**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, including their relationships with the target variable (Default), and examine multicollinearity and interactions.
* Use dimensionality-reduction techniques (e.g., PCA) or clustering to explore structure.
* Apply clustering techniques to explore structural patterns in the dataset.

*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: Data Preprocessing (Part 2 - Feature Transformation & Balancing)**

Incorporate newly engineered features into the dataset. Apply one-hot encoding to transform categorical variables into machine-readable format, scale numerical features where necessary, and address class imbalance using resampling techniques if required.

*Why it matters:* Properly transforming new features and balancing the dataset ensures that all variables contribute effectively to model training. Encoding guarantees algorithm compatibility, scaling prevents dominance of variables with larger ranges, and handling imbalance mitigates bias, leading to more reliable predictions.

* **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 1: 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 2: 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 3: 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 4: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 analysing feature correlations, checking for multicollinearity and interactions, assessing relationships with the target variable through correlation analysis, and exploring structural patterns in the data using dimensionality reduction (e.g., PCA) and clustering. These investigations will guide robust feature engineering and model design.

**7. Full Exploratory Data Analysis (EDA)**

With a validated dataset, the next step is to conduct a full exploratory data analysis (EDA) to uncover hidden patterns and relationships. This stage involves analysing inter-feature correlations — including their relationships with the target variable — as well as identifying multicollinearity and potential feature interactions. We also apply clustering and dimensionality reduction techniques to examine the dataset’s underlying structure. These insights are critical for guiding effective feature engineering and ensuring a strong foundation for model development.

**7.1 Correlation Matrix Analysis**

The goal of this section is to uncover pairwise associations among the dataset’s features — both numerical and categorical — as part of the exploratory data analysis (EDA) stage. These relationships provide early insights into potential redundancies, interactions, and the relative informativeness of variables for downstream modelling.

Given the heterogeneous nature of the dataset (mix of numerical and categorical variables), a set of tailored techniques was used to assess correlations:

|  |  |  |
| --- | --- | --- |
| Variables Types | Method | Description |
| Numerical–Numerical | Spearman’s rank correlation | Captures monotonic (non-linear) relationships; robust to outliers and non-normality |
| Categorical–Categorical | Cramér’s V | Measures strength of association between categorical variables |
| Numerical–Categorical | Mutual Information | Quantifies shared information; captures both linear and non-linear dependencies |

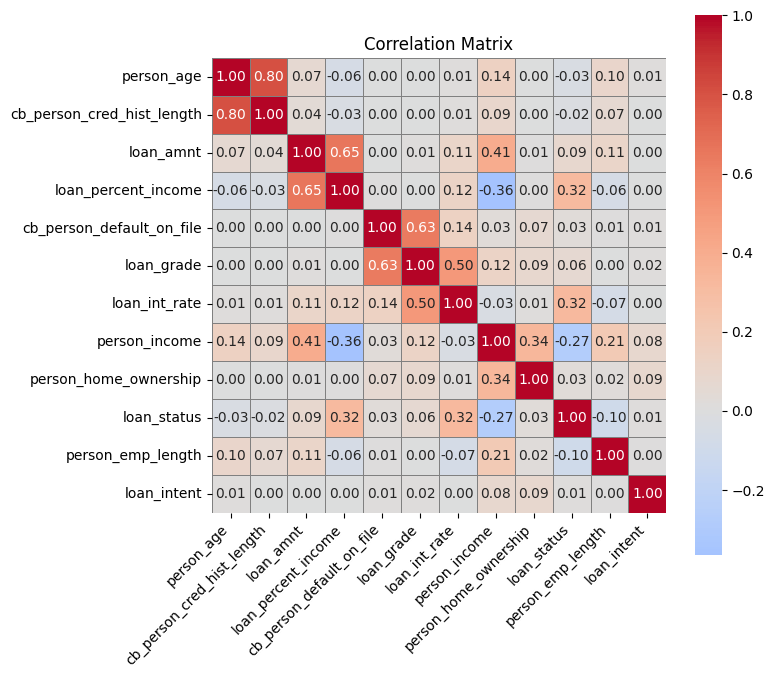
****

Figure 5: Correlation Matrix

**key findings**

* **Strongest** **Relationships**
* *Person age vs person credit history length*Spearman ρ = 0.80  
   → Older individuals typically have longer credit histories — a logical and expected pattern.
* *Loan amount vs loan percent (debt to) income*  
   Spearman ρ = 0.65  
   → Larger loan amounts tend to represent a higher proportion of the borrower's income.
* *Past defaults (on file) vs loan grade*  
   Cramér’s V = 0.63  
   → Borrowers with a default history are more likely to receive lower loan grades.
* *Loan grade vs loan interest rate*  
   Mutual Information = 0.50  
   → A strong alignment exists between loan grades and their corresponding interest rates, suggesting a consistent pricing strategy.
* **Moderate Relationships**
* *Person income vs loan amount*

Spearman ρ = 0.41

→ Higher-income individuals generally apply for larger loans, although the relationship is not strictly linear.

* *Loan status vs loan interest rate (target relationship)*

Mutual Information = 0.32

→ Borrowers with higher interest rates may be more prone to default or charge-off, implying a moderate association with loan outcome.

* *Person income vs person home ownership*  
   Cramér’s V = 0.34  
   → Income level influences home ownership status, distinguishing renters from owners.
* **Weak or Negligible Relationships**

Most other features — including “loan intent”, “employment length”, and “loan status” (when measured against most features) — exhibit weak or negligible pairwise associations (correlation or mutual information < 0.10). Despite this, such variables may still carry predictive power when used in interaction terms or non-linear models.

**Interpretation & Implications**

Some features, such as “age” and “past credit history length” (ρ = 0.80), or “loan amount” and “debt to income” (ρ = 0.65), show high correlation and may introduce collinearity — these should be monitored or addressed. Variables such as “loan grade”, “interest rate”, and “past default on file” demonstrate strong associations and are likely to be informative predictors of borrower risk. The use of mutual information helped uncover non-linear dependencies, particularly useful in mixed-type data. Notably, “interest rate” shows a moderate correlation with the target variable “loan status” (MI = 0.32), suggesting higher rates may relate to higher risk. Most other features show weak direct links to the target, supporting the case for interaction terms or non-linear models in further analysis.

**7.2 Multicollinearity Assessment: Variance Inflation Factor (VIF)**

To assess multicollinearity among the numerical features, Variance Inflation Factor (VIF) was calculated. VIF quantifies how much the variance of a regression coefficient is inflated due to linear relationships with other predictors. A VIF value above 5 (or conservatively 10) typically signals problematic multicollinearity.

**Key Results**

|  |  |
| --- | --- |
| Feature | VIF |
| Age | 4.37 |
| Credit History Length | 4.34 |
| Debt to income Ratio | 2.36 |
| Loan amount | 2.17 |
| Income | 1.48 |
| Loan Status | 1.35 |
| Interest Rate | 1.16 |
| Employment Length | 1.07 |

Table 10: VIF results of numerical variables

Most features exhibit VIF values well below the commonly accepted threshold of 5, indicating a low risk of multicollinearity. The highest VIF values were observed for “age” and “credit history length”, which is expected due to their strong correlation (ρ = 0.80). Additionally, “loan amount and “Debt to Income Ratio” show moderate inflation, likely due to their mathematical relationship.

**Implications for Modelling**

Multicollinearity is not currently a major concern. However, for linear models or interpretability-focused analysis, variables with the highest VIFs (e.g., age, credit history length, debt to income ratio) may be reviewed for removal or transformation. No features were flagged for mandatory exclusion based on high correlation or VIF thresholds.

**7.3 Principal Component Analysis (PCA): Dimensionality Reduction**

Principal Component Analysis (PCA) is a statistical technique used to reduce the dimensionality of a dataset while retaining as much of its original variability as possible. It does so by transforming the original correlated variables into a new set of uncorrelated variables called *principal components (PCs)*. These components are ordered by the amount of variance they capture, allowing analysts to represent the data with fewer variables without a significant loss of information.

This method is particularly useful in predictive modelling, such as Probability of Default (PD) analysis, where many interrelated input features (e.g., customer demographics, loan characteristics, and financial ratios) can introduce noise, multicollinearity, or redundancy. PCA simplifies the input space, supports interpretability, and often improves model generalizability.

**7.3.1 Cumulative Explained Variance Analysis**

The plot of cumulative explained variance illustrates how much of the total variance in the original dataset is retained as more principal components are included.

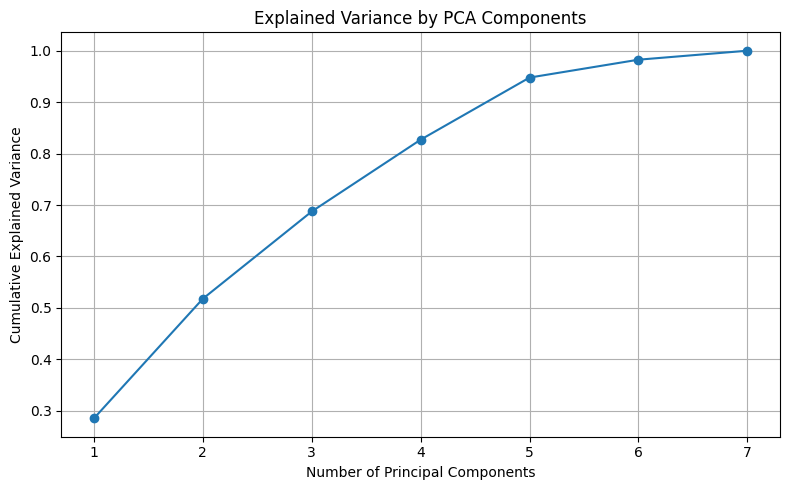


Figure 6: The plot of cumulative explained variance

The key results from the PCA conducted on the dataset are summarized below:

|  |  |
| --- | --- |
| Number of Principal Components | Cumulative Explained Variance |
| 1 (PC1) | ~28% |
| 2 (PC1–PC2) | ~52% |
| 3 (PC1–PC3) | ~69% |
| 4 (PC1–PC4) | ~83% |
| 5 (PC1–PC5) | ~95% |
| 6 (PC1–PC6) | ~99% |
| 7 (PC1–PC7) | ~100% |

Table 11: The key results from the PCA on numerical variables

These results show that five components can retain approximately 95% of the variance in the dataset, while six components capture nearly all of it.

**Implications for the PD Model**

* **Dimensionality Reduction:** PCA allows transformation of the original high-dimensional dataset into a smaller set of features (PCs) that are uncorrelated and information-rich. This can reduce model complexity and improve performance.
* **Component Selection Strategies:**
  + **Variance Threshold:** A common rule is to select enough components to retain a target level of variance—typically 90–99%. In this case, selecting 5–6 components achieve that.
  + **Elbow Method:** By identifying where the curve of explained variance flattens, one can find an optimal trade-off. Here, the curve flattens noticeably after the 4th or 5th component.
* **Model Benefits:** Using PCA components in modelling helps mitigate multicollinearity and overfitting, particularly in linear models like logistic regression. Because the components are orthogonal by design, they eliminate redundant predictive power across features.

PCA serves as both a data reduction and exploratory tool in the PD pipeline, helping to streamline the feature set while preserving the essential patterns in the data.

**7.3.2 PCA Component Loadings Interpretation**

To better understand the underlying structure of the data after Principal Component Analysis (PCA), we analyzed the **component loadings**, which quantify how much each original feature contributes to each principal component (PC). These loadings help interpret what each principal component represents, enhancing the transparency of the dimensionality reduction process.

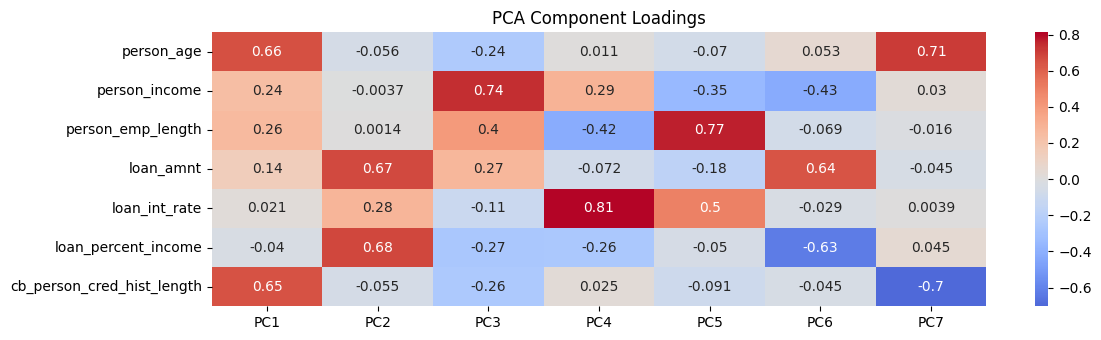


Figure 7: PCA Component Loadings Interpretation for Numerical Variables

**Key Component Interpretations**

***PC1 – Age and Credit History Dimension***

* Dominant features:
  + Age (0.66),
  + credit history length (0.65)
* Moderate contributors: employment length (0.26), income (0.24)
* Interpretation: This component captures customer *maturity and financial history*. Higher PC1 scores represent older individuals with longer credit histories and employment experience — characteristics often linked with lower default risk.

***PC2 – Loan Amount and Income Stress***

* Dominant features:
  + Loan amount (0.67),
  + Debt to income Ratio (0.68)
* Moderate contributor: interest rate (0.28)
* Interpretation: This component reflects *loan burden relative to income*. Higher scores imply larger loans and a higher proportion of income going toward loan repayment — indicative of financial strain and potential credit risk.

***PC3 – Income and Employment-Based Capacity***

* Dominant feature: person income (0.74)
* Moderate: employment length (0.40), loan amount (0.27)
* Negative contributors: age (−0.24), past credit history length (−0.26)
* Interpretation: Highlights *earning power and employment tenure*, offset by lower age and shorter credit history. May represent younger professionals with rising financial capacity.

***PC4 – Interest Rate Sensitivity***

* Dominant feature: interest rate (0.81)
* Moderate contributor: person income (0.29)
* Interpretation: Strongly associated with the *interest rate on the loan*. High PC4 scores could indicate risk-based pricing, or segments receiving higher interest offers regardless of income.

***PC5 – Employment vs. Income Mismatch***

* Dominant feature: employment length (0.77)
* Moderate negative: person income (−0.35)
* Interpretation: Represents individuals with long employment histories but comparatively lower incomes. May highlight stable yet underpaid individuals — a potentially distinct credit segment.

***PC6 – Discretionary Burden Indicator***

* Dominant negative loadings:
  + Debt to income Ratio (−0.63),
  + Interest rate (−0.43)
* Moderate negative: person\_ income (−0.43)
* Interpretation: Reflects *financial relief*, with low loan burden and interest costs. Negative PC6 scores may indicate more favorable loan terms or borrowers with better credit profiles.

***PC7 – Age vs. Credit Length Divergence***

* Dominant contributors:
  + Person age (0.71),
  + Past credit history length (−0.70)
* **Interpretation**: Contrasts *older individuals with shorter credit histories*. This could signify demographics like immigrants or late adopters of credit products — an important niche in PD modelling.

**Modelling Insights**

* **Dimensional Insight**: Each PC captures a unique combination of behavioural and demographic patterns in the borrower data.
* **Predictive Power**: These PCs can now be used as inputs in PD models (e.g., logistic regression, decision trees) with reduced multicollinearity and enhanced efficiency.
* **Interpretability**: Understanding what each PC represents allows better model explainability and alignment with regulatory or business-driven model validation.

**7.3.3 PCA Scatter Plot Interpretation (PC1 vs PC2 Coloured by Target)**

This plot shows how the first two principal components (PC1 and PC2) capture the variance in the dataset and relate to the target variable (loan status):

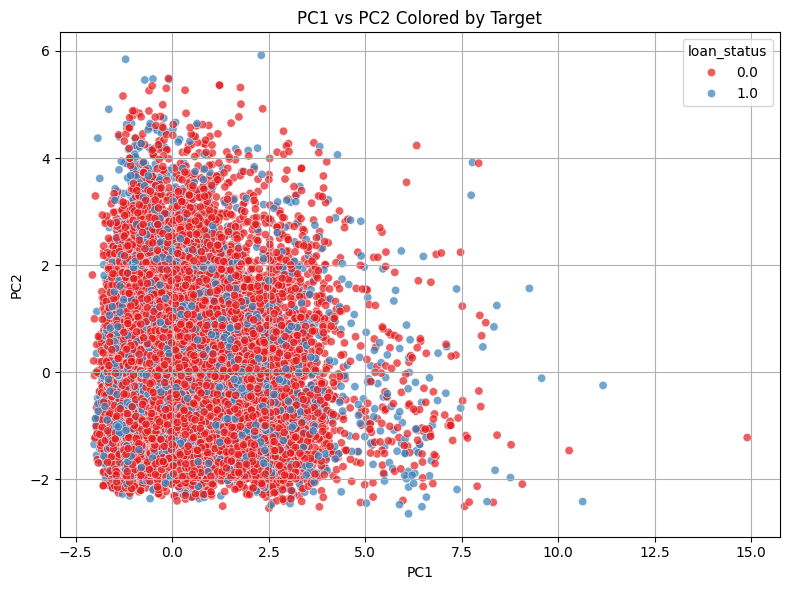


Figure 8: PCA Scatter Plot of PC1 vs PC2 Coloured by Target

* Axes:
  + PC1 (X-axis): Captures the largest variance across features.
  + PC2 (Y-axis): Captures the second largest, orthogonal variance.
* Colour Coding:
  + Red (0.0): Likely represents non-default cases.
  + Blue (1.0): Likely represents default cases.
* Interpretation:
  + There is no strong visual separation between classes — red and blue points are heavily overlapping.
  + This suggests that PC1 and PC2 alone do not differentiate defaulters from non-defaulters well.
  + Some minor clustering and outliers exist, but overall, the distribution is mixed.
* Implications:
  + PCA is useful here for exploratory analysis, but these two components *should not be used alone* for modelling.

**7.3.4 Final Note on PCA Usage**

In this project, PCA was applied solely for exploratory data analysis (EDA) and visualization — not for model training. Although it doesn't provide strong class separation in the 2D plot, PCA helped uncover underlying data structure, possible redundancy, and variance patterns across features. Since the final goal involves interpretable modelling (e.g., logistic regression for default prediction), and the number of features is manageable, PCA is not suitable for the modelling pipeline. Instead, it serves here as a valuable tool for understanding the data's geometry and guiding thoughtful feature selection.

**7.4 Multiple Correspondence Analysis (MCA): Uncovering Latent Structures in Categorical Data**

To explore patterns and latent structures within categorical variables, Multiple Correspondence Analysis (MCA) was applied. MCA is a dimensionality reduction technique specifically designed for categorical data and is conceptually similar to Principal Component Analysis (PCA), but tailored for qualitative variables. It transforms categorical values into a set of uncorrelated components, allowing complex relationships to be visualized and interpreted more easily. By projecting both the categories and observations into a common low-dimensional space, MCA helps reveal associations between variables, identify clusters of similar individuals, and reduce noise in further downstream analyses such as clustering or classification. This technique was conducted on all categorical features, with the number of components set to the maximum possible in order to retain full inertia and ensure no loss of information.

**7.4.1 Analysis of Categorical Variable Structure: Explained Inertia**

To understand the underlying structure and inter-relationships among the key categorical variables home ownership, loan intent, loan grade, and past defaults — a Multiple Correspondence Analysis (MCA) was conducted. The analysis decomposed the total variance (inertia) of the dataset into 15 independent dimensions, or components. The proportion of inertia explained by each of these dimensions is presented below.

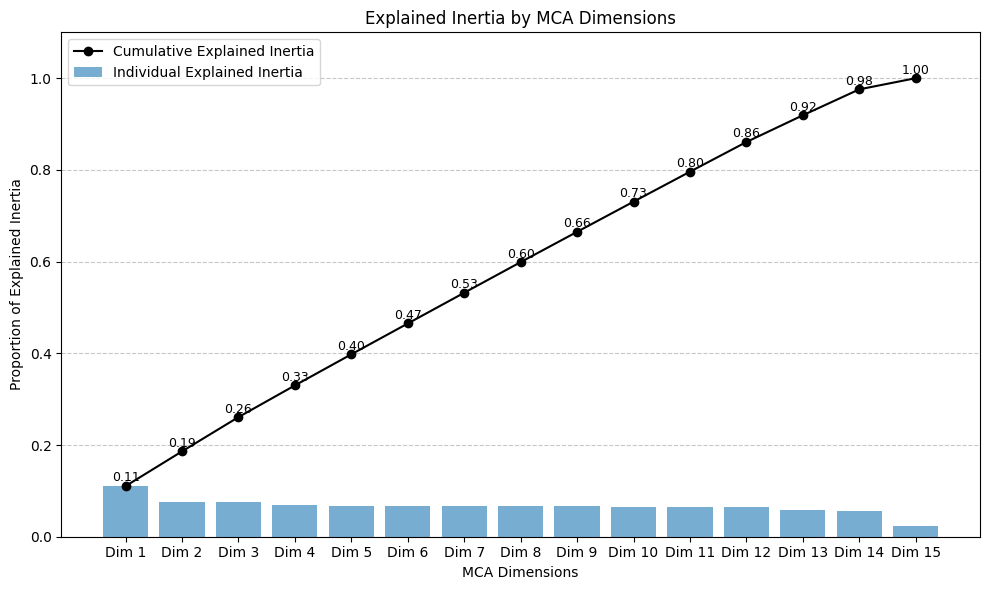


Figure 9: The plot of cumulative explained inertia

The analysis reveals the following key points:

* **Dimension 1**, the most significant component, accounts for **11.05%** of the total inertia in the data. This primary dimension represents the single strongest pattern of association among the variables.
* The subsequent dimensions capture progressively smaller, yet still relevant, amounts of inertia. **Dimension 2** explains **7.57%**, and **Dimension 3** explains **7.47%**. The decline in explained inertia is gradual, indicating that multiple components are required to understand the data's structure.
* Collectively, the first three dimensions account for just **26.1%** of the total variance. To capture a majority of the information (over 50%), at least seven dimensions are required, which collectively explain **53.2%** of the total inertia.

**Interpretation and Implications**

The distribution of inertia across multiple components is a significant finding. It indicates that there is **no single, dominant pattern** that explains the relationships between home ownership, loan intent, loan grade, and default history. Instead, the variance is fragmented across many distinct, smaller patterns.

If the structure were simple, a much larger proportion of inertia would be concentrated in the first one or two dimensions. The observed gradual decline suggests that the customer profiles defined by these variables are highly varied and cannot be easily summarized in a simple two-dimensional chart. The relationships are multifaceted, implying that different combinations of these categorical variables create numerous, distinct customer segments. This complexity is a critical insight for any subsequent modelling or strategic decision-making.

**7.4.2 MCA Row Coordinates: Observation Distribution by Loan Status**

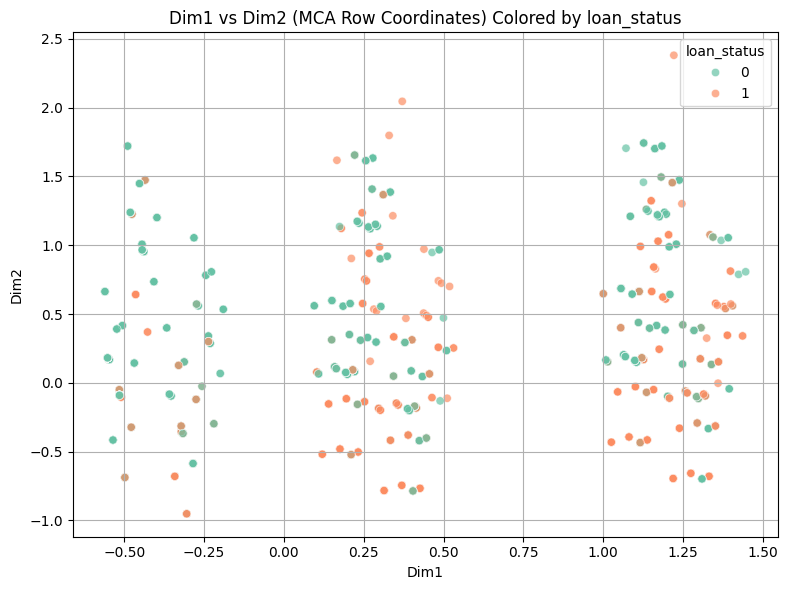


Figure 10: The plot of the position of each row in the reduced dimensional space, colored by the target variable

* **General Structure**:
  + The data appears to form **several vertical groupings** or clusters along the Dim1 axis.
  + This suggests that Dim1 may be capturing a **categorical variable with a few dominant levels** or a **composite pattern** that repeats.
* **Separation of Target Classes**:
  + There is no strong visual separation between the two classes (loan status 0 and 1).
  + Both colours are spread across the same regions, with some minor density differences but substantial overlap.
  + This implies that the first two MCA dimensions do not clearly separate defaulted from non-defaulted loans.
* **Implication for Modelling or Visualization**:
  + While these first two dimensions capture some structure, they may not be sufficient alone for classifying or predicting loan status.

**7.4.3 MCA Dimension Interpretation: Variable Contributions and Semantic Meaning**

Following the identification of the principal dimensions from the Multiple Correspondence Analysis (MCA), this section explores the contribution of individual variable categories to each dimension. This analysis provides insight into what each dimension represents by identifying which variables define the patterns of variance (inertia) most strongly. Categories that contribute more significantly to a particular dimension are those that help distinguish observations along that axis. In contrast, values with near-zero contributions have little influence on the formation of that component.

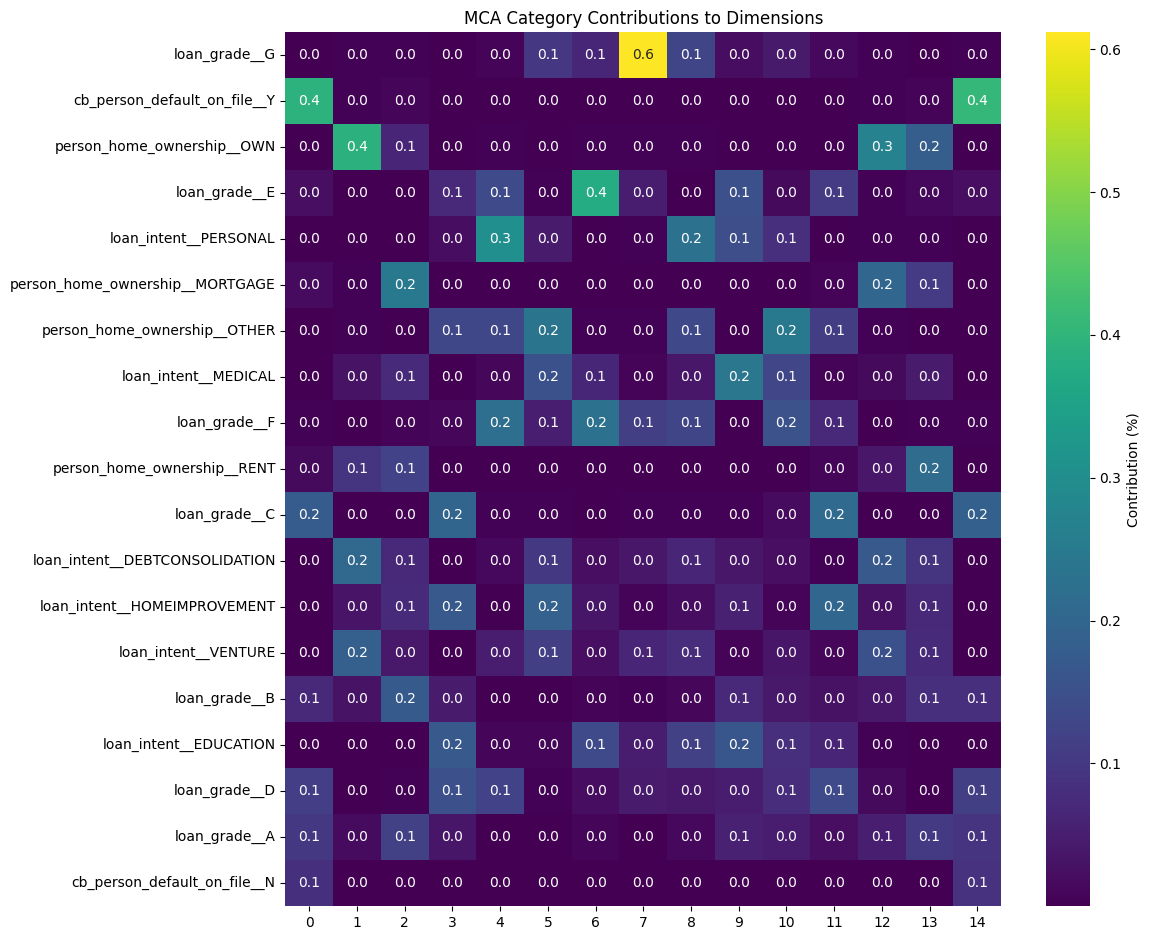


Figure 11: category contribution to the variance (inertia) explained by that dimension

**Dimension 1: The Default History Axis (11.05% of Inertia)**

The first and most influential dimension is primarily shaped by variables associated with an individual’s creditworthiness.

* **Key Contributing Categories**:
  + cb\_person\_default\_on\_file\_\_Y (i.e., individuals with a default history) contributes **39.3%** to this dimension.
  + loan\_grade\_\_C is the next most important, contributing **17.7%**.
* **Interpretation**:  
  This axis differentiates individuals based on past credit behavior. It effectively separates those with a history of default and a mid-to-low loan grade from those with a clean record or higher loan grades. As such, Dimension 1 can be interpreted as an axis of credit risk exposure.

**Dimension 2: The Homeowner Profile Axis (7.57% of Inertia)**

The second dimension captures customer profiles related to homeownership and loan intent.

* **Key Contributing Categories**:
  + person\_home\_ownership\_\_OWN (fully owned homes) contributes **39.0%**.
  + Loan intents such as DEBTCONSOLIDATION (**20.5%**) and VENTURE (**18.3%**) also feature prominently.
* **Interpretation**:  
  This axis represents a segment of customers who fully own their homes and are primarily seeking loans for specific strategic or financial goals. These individuals are distinct from renters or mortgage holders, and their borrowing intentions are more focused on **consolidating debt or funding business ventures**.

**Dimension 3: The Mortgage Holder Profile Axis (7.47% of Inertia)**

The third dimension reveals a separate customer profile, driven by a different pattern of homeownership and loan quality.

* **Key Contributing Categories**:
  + person\_home\_ownership\_\_MORTGAGE is the dominant contributor with **24.7%**.
  + loan\_grade\_\_B follows with a contribution of **17.0%**.
* **Interpretation**:  
  This dimension characterizes traditional borrowers with active mortgages who typically qualify for relatively high-quality loan products. This group contrasts with those captured in Dimension 1 (with defaults) and Dimension 2 (homeowners seeking specific loans), providing insight into a **middle-tier, creditworthy borrower segment**.

**Summary of Findings**

The analysis of category contributions demonstrates that the categorical structure of the dataset is governed by multiple distinct, interpretable dimensions rather than a single dominant pattern. Dimension 1 isolates credit risk profiles, while Dimensions 2 and 3 differentiate customer segments based on homeownership status and loan intent or grade. These insights can inform segmentation strategies, risk profiling, and targeted financial offerings in subsequent analyses.

**7.4.4 Executive Summary: Insights from Multiple Correspondence Analysis (MCA)**

The Multiple Correspondence Analysis (MCA) was performed to move beyond surface-level statistics and uncover the fundamental structure linking customer homeownership, loan intentions, creditworthiness, and default history. The analysis reveals a nuanced and complex market landscape, providing critical strategic insights.

In essence, the analysis shows that customer profiles are not defined by one or two dominant characteristics, but by the interplay of several distinct patterns. The most powerful single factor separating individuals is their **past credit performance and associated risk profile**. Beyond this, the analysis clearly identifies other significant, independent customer segments. For instance, a clear distinction emerged between **homeowners who own their property outright** (who are strongly associated with seeking loans for debt consolidation or business ventures) and **mortgage holders** (who are more typically associated with higher-grade loans for other purposes).

Conducting this MCA provides several clear advantages for data-driven decision-making:

* **Reveals Hidden Relationships:** The analysis moves beyond simple one-variable charts to mathematically map the "space" between categories. It objectively proves that the relationship between variables like homeownership and loan intent is not uniform. The fact that outright owners behave differently from mortgage holders is a complex interaction that this analysis successfully brought to light.
* **Provides a Blueprint for Sophisticated Segmentation:** MCA provides a data-driven foundation for creating nuanced customer personas. Instead of broad segments like "Homeowners," we can now define and target more specific, actionable groups such as the "Established Entrepreneurial Homeowner" or the "Conventional Mortgage-Holding Borrower." This allows for more precise marketing, product development, and risk assessment.
* **Creates Powerful Features for Predictive Modelling:** The dimensions identified by MCA (Dimension 1, 2, 3, etc.) are, in themselves, powerful new numerical features. They condense the complex, multi-column categorical information into a set of clean, uncorrelated variables. Using these dimensions as inputs for machine learning models (e.g., to predict loan default) can significantly improve their performance and accuracy.
* **Validates and Challenges Business Assumptions:** This analysis provides an objective, evidence-based picture of the customer landscape, allowing the business to either confirm its existing assumptions or, more importantly, challenge them with data. By understanding that the market is segmented into multiple, distinct behavioral patterns, strategies can be refined to more effectively address the true complexity of the customer base.

In summary, the MCA has transformed a set of simple categorical labels into a rich, multi-dimensional map of customer profiles, offering a clearer understanding of the underlying market dynamics and providing a solid analytical foundation for future strategic initiatives.

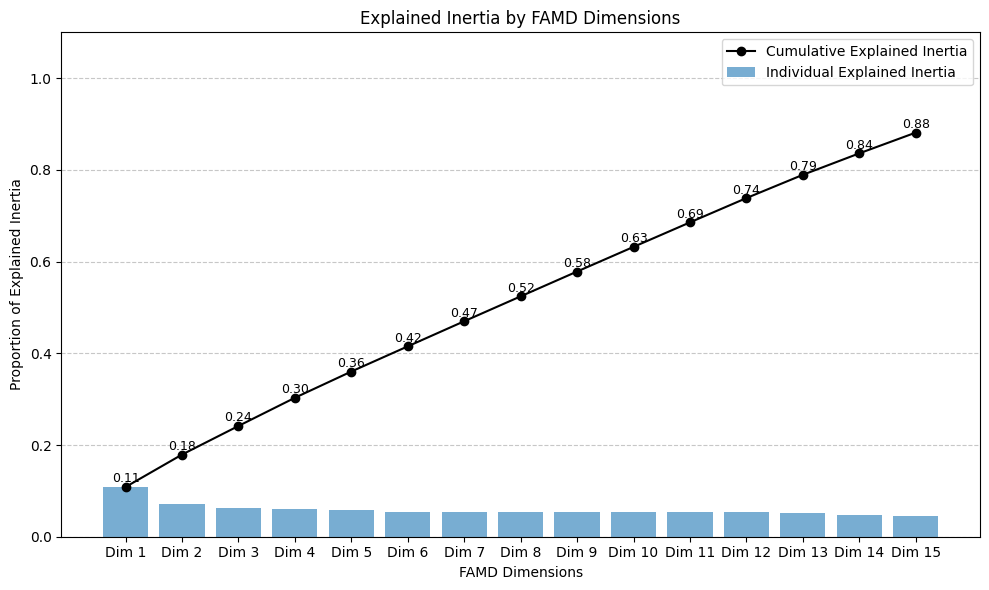
**7.5 Factor Analysis of Mixed Data (FAMD): Structural Analysis of Mixed Features**

To uncover the latent structure within the dataset containing both categorical and numerical features, a Factor Analysis of Mixed Data (FAMD) was conducted. FAMD is a dimensionality reduction technique designed to integrate the strengths of Principal Component Analysis (PCA) for continuous variables and Multiple Correspondence Analysis (MCA) for categorical variables. This makes it an ideal tool for exploring mixed-type datasets often encountered in real-world business and financial applications. The goal is to transform the original variables into a set of uncorrelated components that retain as much of the total variance (inertia) as possible.

**7.5.1 Analysis of Explained Inertia**

The FAMD was fit using the maximum number of possible components, which in this case yielded 15 dimensions. Each component represents a new axis of variation derived from a combination of both categorical and numerical inputs.

The distribution of explained inertia across these components is presented below:



**Key Observations:**

* **Dimension 1** is the most influential component, explaining **10.85%** of the total variance. This indicates that while it captures the most dominant pattern in the data, the pattern itself is not overwhelmingly strong.
* **Dimensions 2 and 3** explain **7.07%** and **6.24%**, respectively. The gradual decline in inertia across subsequent dimensions suggests the absence of a strong “elbow” point, meaning there’s no clear cutoff beyond which the remaining components become insignificant.
* To capture **approximately 47.0%** of the total variance, it is necessary to retain at least **seven components**, and the **first three** dimensions together account for **24.16%**.

**Interpretation and Implications**

The dispersion of explained inertia across many components suggests that the underlying structure of the dataset is complex and multifaceted. Unlike PCA or MCA on simpler datasets—where a handful of dimensions may capture the majority of variance—this FAMD result reveals that no single variable or group of variables dominates the structure.

This complexity arises from the interactions between numerical features (e.g., person income, person age, loan amount) and categorical features (e.g., loan grade, loan intent, past defaults). The low proportion of inertia explained by the first few components confirms that customer profiles are defined by many small, interacting patterns rather than by a few large, dominant ones.

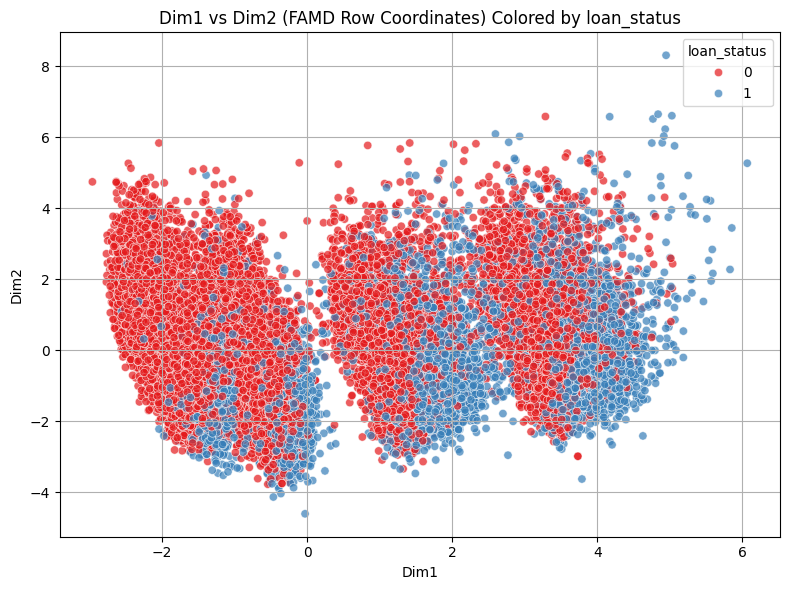
From a business perspective, this has several important implications:

* Customer segmentation and predictive modelling should not rely on simplistic dimensionality reduction or assume homogeneity within a few clusters.
* Any effective strategy for credit scoring, risk analysis, or marketing will require a more nuanced approach, taking into account several latent dimensions.
* The leading FAMD components—though individually weak—still highlight the most informative axes of variation and should be further examined for feature engineering, modelling input, and domain-specific interpretation.

**7.5.2 Visual Analysis of Principal Components by Loan Status**

To evaluate the effectiveness of the dimensions derived from the Factor Analysis of Mixed Data (FAMD), a **scatter plot** was generated, visualizing individual loan applicants in the reduced two-dimensional space. Each data point corresponds to a single observation, projected onto the first two principal components extracted by FAMD. Points are color-coded based on the binary target variable loan\_status, where:

* **0 (Red)** indicates a non-defaulted loan,
* **1 (Blue)** indicates a defaulted loan.



This visualization serves as a diagnostic tool to assess whether the newly constructed dimensions reveal meaningful patterns related to loan outcomes.

**Interpretation of the Plot**

**▪ Dimension 1: The "Loan Risk Profile" Axis**

* A clear horizontal separation is observed along Dimension 1.
* Data points with strongly negative scores on this axis (toward the left) are predominantly associated with non-defaulted loans.
* As one moves rightward along Dimension 1, the concentration of defaulted loans increases noticeably.
* This suggests that Dimension 1 effectively captures a latent pattern strongly associated with credit risk. Borrowers scoring low on this axis tend to exhibit characteristics linked with safer loan profiles, while higher values correlate with increased risk of default.

**▪ Dimension 2: The "Applicant Financial Status" Axis**

* Unlike Dimension 1, Dimension 2 does not show a strong vertical separation between the two loan outcome classes.
* Both defaulted and non-defaulted loans are distributed across the full range of values on this axis.
* However, Dimension 2 contributes to shaping secondary groupings and substructures in the dataset. It may reflect other influential traits (such as income level, employment length, or homeownership type), which, while not directly predictive of loan default, help define meaningful customer subsegments.

**Conclusion and Strategic Implications**

The results of this visual analysis confirm that the first principal component derived from FAMD captures a major, interpretable pattern aligned with credit default risk. Its effectiveness in separating the classes visually justifies its inclusion as a valuable engineered feature for further modeling.

Moreover, this demonstrates that FAMD not only reduces dimensionality, but also preserves and highlights business-relevant patterns. The strength of the primary component suggests it can serve as an input to credit risk models, segmentation strategies, or explainable AI tools to improve transparency in decision-making.

This finding underscores the practical utility of advanced factor analysis techniques in domains involving mixed-type customer data, where predictive accuracy and interpretability must go hand-in-hand.

**7.5.3 Interpretation of FAMD Components: Analysis of Variable Contributions**

While the analysis of explained inertia in the previous section illustrated that the variance within the dataset is broadly distributed across multiple dimensions, a deeper understanding of these dimensions can be achieved by examining the variable contributions to each component. This analysis helps uncover the latent structure within the data by identifying which variables define each axis of variation (Refer to appendix for the full result).

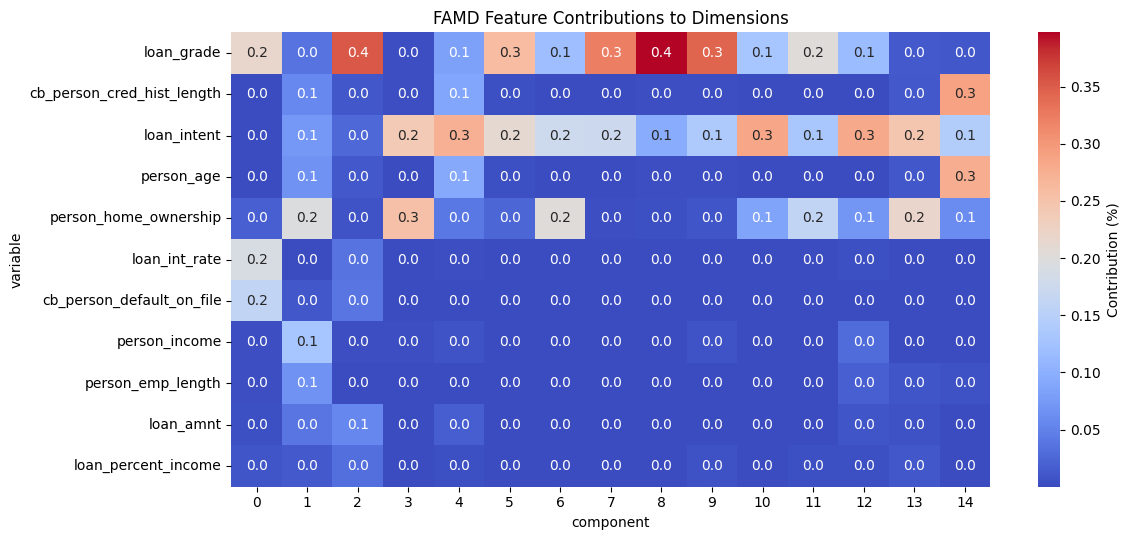


Table 12: Contributions of original variables to the first 15 principal components derived from the Factor, rounded to one decimal

**Dimension 1: The “Loan Risk Profile” Axis (10.85% of Inertia)**

The first principal dimension is overwhelmingly influenced by factors related to creditworthiness and loan risk.

* **Key Contributing Variables**:
  + loan\_grade: 21.7%
  + loan\_int\_rate: 19.0%
  + cb\_person\_default\_on\_file: 16.1%
* **Interpretation**:  
  This component captures the risk-return spectrum of the loans. On one end are high-risk applicants with previous defaults, high interest rates, and lower-grade loans; on the other are more creditworthy applicants with cleaner histories and more favorable loan terms. The axis effectively summarizes an applicant’s financial risk profile and serves as a foundational dimension for credit assessment.

**Dimension 2: The “Applicant Financial Status” Axis (7.07% of Inertia)**

The second dimension shifts focus to socioeconomic characteristics of the applicant.

* **Key Contributing Variables**:
  + person\_home\_ownership: 19.7%
  + person\_income: 13.0%
  + loan\_intent: 7.2%
  + person\_emp\_length: 6.6%
* **Interpretation**:  
  This dimension differentiates applicants based on their economic position and life circumstances. For instance, homeowners with stable income and employment histories tend to cluster on one end of the axis, whereas renters or less financially secure individuals are located on the opposite end. The dimension provides a profile of the applicant’s financial well-being and intent behind borrowing.

**Dimension 3: The “Loan Structure” Axis (6.24% of Inertia)**

This dimension returns to loan-related features but highlights product differentiation rather than risk.

* **Key Contributing Variables**:
  + loan\_grade: 35.4%
  + loan\_amnt: 5.4%
  + cb\_person\_default\_on\_file: 3.8%
* **Interpretation**:  
  The dominant influence of loan\_grade in this dimension suggests that it captures granular differences between loan products, possibly differentiating sub-prime, near-prime, and prime classifications. The presence of loan\_amnt and default\_on\_file as supporting variables further refines this structural interpretation. This axis seems to emphasize product stratification beyond just borrower risk.

**Summary of Findings**

The component contribution analysis reveals that the structure of the dataset is governed by distinct conceptual axes, each representing a different facet of the borrower–loan relationship:

1. **Dimension 1** — Financial risk and loan quality
2. **Dimension 2** — Applicant socioeconomic profile
3. **Dimension 3** — Loan structural characteristics

These dimensions provide a compact and interpretable representation of complex customer profiles, which are not only valuable for exploratory analysis, but can also be used as engineered features in predictive models aimed at assessing creditworthiness or personalizing loan offerings. Their decorrelated nature further enhances the statistical robustness of downstream analyses.

**7.5.4 Executive Summary: Key Findings from Factor Analysis of Mixed Data (FAMD)**

This report applied Factor Analysis of Mixed Data (FAMD) to uncover the underlying structure of a complex dataset comprising both numerical and categorical features related to customer profiles and loan outcomes. The primary objective was to reduce dimensionality while retaining meaningful, interpretable components that capture the key drivers of loan performance.

The analysis revealed that no single feature dominates the variance in the dataset. Instead, customer and loan characteristics interact in multifaceted and distributed ways, highlighting the inadequacy of simplistic segmentation or modelling approaches. This foundational insight underscores the necessity of advanced dimensionality reduction techniques like FAMD when dealing with heterogeneous financial data.

Despite the complexity, FAMD successfully constructed interpretable composite dimensions that encapsulate the most critical latent factors:

* **Dimension 1 – Loan Risk Profile (10.85% of inertia)**

This axis aggregates key risk-related variables including loan grade, interest rate, and past defaults. It effectively serves as a proxy for assessing credit risk, clearly distinguishing between high- and low-risk applicants.

* **Dimension 2 – Applicant Financial Status (7.07% of inertia)**

This dimension reflects the applicant’s broader financial condition, dominated by features such as income, home ownership, and loan intent. It captures variance in socioeconomic status and underlying financial intent.

The practical significance of these dimensions was validated through visual analysis. When individuals were plotted in the space defined by the first two dimensions, a strong visual separation emerged between defaulted and non-defaulted loans, particularly along Dimension 1. Applicants with lower scores on this dimension were predominantly associated with successful loan outcomes, whereas higher scores indicated increased likelihood of default.

In conclusion, the FAMD has proven to be an exceptionally effective technique. It has successfully navigated a complex dataset to:

* Revealing that customer profiles are multidimensional, with variance spread across a combination of behavioural, demographic, and financial variables.
* Creating interpretable and composite features that reflect latent constructs such as creditworthiness and financial stability.
* Confirming that these components have a direct and observable relationship with the target outcome (loan status), making them ideal inputs for advanced predictive models and strategic business analysis.

These insights are not merely descriptive but actionable. The newly engineered components provide a data-driven foundation to:

* Enhance model accuracy and robustness.
* Guide more nuanced segmentation and risk strategies.
* Inform lending policies tailored to distinct customer risk profiles.

In short, FAMD has transformed complex raw data into strategic intelligence.

**7.6 Clustering Analysis: Uncovering Natural Borrower Segments**

**7.6.1 Introduction**

As the final component of our exploratory data analysis (EDA), clustering was employed to reveal natural groupings within the borrower population. Unlike supervised techniques, clustering is an unsupervised learning approach that identifies patterns and similarities in data without reference to the outcome variable (loan default). This makes it an ideal tool for discovering latent segments, supporting feature engineering, and enhancing model interpretability.

The clustering analysis complements prior dimensionality reduction efforts (e.g., PCA and MCA) by shifting the focus from variable relationships to how individual borrowers cluster together based on shared traits. These natural clusters can highlight hidden borrower profiles — such as “young renters with high loan-to-income ratios” or “older homeowners with conservative borrowing behaviour” — that may not emerge from variance-based analysis alone.

To accommodate the mixed nature of the dataset, multiple clustering algorithms were selected based on input type:

|  |  |  |
| --- | --- | --- |
| Data Type | Clustering Methods | Description |
| Numerical or PCA-transformed | - K-Means  - Agglomerative Hierarchical Clustering  - DBSCAN | Used to detect spherical (K-Means), nested (Hierarchical), or density-based (DBSCAN) patterns. |
| Categorical or MCA-transformed | - K-Modes  - Hierarchical Clustering on MCA Coordinates | Segment categorical data based on frequency-based similarity or MCA-derived coordinates. |
| Mixed-type (Numerical + Categorical) | - K-Prototypes  - Hierarchical Clustering with Gower Distance | Handle both numeric and categorical features using combined similarity measures and flexible clustering. |

Table 13: Overview of clustering algorithms applied in this study based on the nature of the input data

**Why Clustering Matters**

Clustering plays a pivotal role in:

* Revealing hidden borrower groups that inform strategic segmentation.
* Enhancing feature engineering through cluster-based labels.
* Providing a structure-agnostic view of the data (not influenced by the target variable).
* Supporting anomaly detection and robustness checks.

The results of each method are analysed in the following subsections, where we interpret the formed clusters, assess their alignment with business goals, and evaluate their potential for downstream modelling.

**Model Evaluation and Cluster Selection**

To ensure the robustness and interpretability of the clustering results, **quantitative metrics** were used to evaluate cluster quality and determine the optimal number of clusters. The two primary evaluation metrics applied were **Silhouette Score** and **Inertia**, supported by **grid search techniques**.

* **Silhouette Score**
* The **Silhouette Score** measures how well each data point fits within its assigned cluster compared to other clusters. It captures both **cohesion** (how close the point is to other points in its own cluster) and **separation** (how far it is from points in the next nearest cluster). The score ranges from **-1 to 1**:

|  |  |
| --- | --- |
| Score Range | Interpretation |
| 0.50 to 1.00 | Strong and well-defined clusters |
| 0.25 to 0.50 | Weak or fuzzy cluster structure |
| < 0.25 | No substantial structure present |
| < 0.00 | Likely misassigned clusters |

A **higher silhouette score** indicates better cluster definition. We used this score to fine-tune parameters like the number of clusters and to select the best-performing algorithm for each dataset transformation (PCA, MCA, or mixed).

* **Inertia**

**Inertia** is defined as the **sum of squared distances** between each point and the centroid of its assigned cluster. It measures the **compactness** of the clusters.

* **Lower inertia** indicates more cohesive clusters.
* However, inertia **always decreases** as the number of clusters increases, so it must be interpreted alongside other metrics (like the Silhouette Score or the Elbow Method).

Together, these metrics provide a robust and complementary framework for selecting and validating the clustering configuration. While inertia quantifies how compact clusters are, silhouette scores help determine how distinct they are — both are essential to ensuring meaningful and reliable segmentation results.

* **Grid Search for Optimal Clusters**

To determine the most appropriate number of clusters for each method, a **grid search** approach was implemented across a predefined range of cluster counts. For each value (e.g., k in K-Means), both **Silhouette Score** and **Inertia** were computed. The optimal number of clusters was selected based on the configuration that maximized Silhouette Score while maintaining a reasonable balance of compactness (low inertia) and interpretability.

This methodology ensured that the selected clustering solutions were not only statistically sound but also aligned with meaningful borrower segmentation patterns, ready to inform both strategic analysis and downstream predictive modelling.

**7.6.2** **K-Means Clustering: Segmenting Borrowers in PCA Space**

**K-Means clustering** is a widely used unsupervised learning algorithm designed to partition data into a predefined number of clusters (**K**) based on similarity. It operates by minimizing the **within-cluster variance**, or **inertia**, by iteratively assigning data points to the nearest cluster centroid and then updating those centroids based on the new assignments. The algorithm assumes that clusters are roughly spherical and evenly sized, which makes it particularly well-suited to **PCA-transformed numerical data**, where dimensionality and correlations have already been reduced. In this analysis, we applied K-Means to the PCA-reduced borrower dataset to uncover latent groupings of customers with similar financial characteristics, independent of the loan outcome.

**7.6.2.1 K-Means Model Selection and Evaluation**

To determine the optimal number of clusters for K-Means segmentation, a **grid search** was conducted across a range of cluster values from 2 to 9 (refer to appendix for detailed results). Two evaluation metrics were computed at each step:

* **Inertia**: Measures within-cluster compactness (lower is better).
* **Silhouette Score**: Measures how similar each point is to its own cluster versus other clusters (higher is better, ideally close to 1).

The silhouette score peaked at **k = 2**, indicating that this configuration provides the most distinct grouping of observations. However, the silhouette value of **0.1987** suggests **weak structure**, implying that the clusters, though optimal relative to other configurations, are not particularly well-separated.

Despite the low silhouette score, k = 2 was selected, as increasing the number of clusters led to diminishing silhouette scores and the creation of increasingly overlapping groups — likely introducing artificial substructures.

**Interpretation**: The low silhouette score at k = 2 indicates that while a natural split exists, it is not particularly strong. This suggests borrower profiles form **only mildly distinct groups** in the PCA-reduced space, and segmentation should be interpreted with caution.

**7.6.2.2 Interpretation of K-Means Clustering Results (k = 2)**

Following the selection of **k = 2** clusters based on silhouette optimization, a detailed comparative analysis was performed to understand the characteristics defining each cluster. The features used in this analysis include demographic and financial indicators such as age, income, employment length, loan amount, interest rate, loan burden relative to income, and credit history length.

**Cluster Composition and Interpretation**

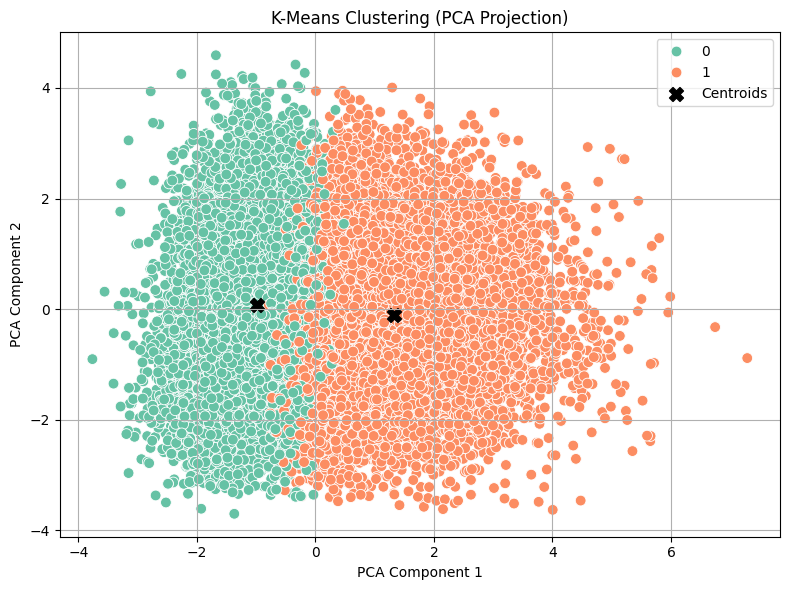


Figure 12: A 2D scatter plot visualising the K-Means clustering (k = 2) in the PCA-reduced space

The two clusters uncovered by K-Means reflect **distinct borrower profiles**, which can be summarized as follows:

**🔹 Age**

* **Cluster 0**:
  + **Mean Age** ≈ 24 years
  + **Standard Deviation** ≈ 1.7 years
  + Profile: Predominantly young individuals, tightly clustered in age.
* **Cluster 1**:
  + **Mean Age** ≈ 33 years
  + **Standard Deviation** ≈ 6.1 years
  + Profile: Older and more age-diverse population.

**Insight**: This is a strong separating factor; age clearly differentiates the two groups.

**🔹 Income**

* **Cluster 0**:
  + **Mean** ≈ $59,300
  + **Max**: $500,000
* **Cluster 1**:
  + **Mean** ≈ $75,000
  + **Max**: $6,000,000

**Insight**: Cluster 1 includes significantly higher incomes, with extreme outliers. This contributes to a higher mean and variance. Distribution appears skewed.

**🔹 Employment Length**

* **Cluster 0**: ≈ 4 years
* **Cluster 1**: ≈ 5.7 years

**Insight**: Individuals in Cluster 1 tend to have more established employment histories, aligning with their older age.

**🔹 Loan Amount**

* **Cluster 0**: ≈ $9,177
* **Cluster 1**: ≈ $10,133

**Insight**: Borrowers in Cluster 1 take slightly larger loans, consistent with their higher income and financial maturity.

**🔹 Interest Rate**

* Both clusters exhibit similar average interest rates:
  + **Cluster 0**: ≈ 10.96%
  + **Cluster 1**: ≈ 11.08%

**Insight**: Loan pricing does not vary significantly between the clusters, suggesting similar credit risk or product offerings across both groups.

**🔹 Loan-to-Income Ratio**

* **Cluster 0**: ≈ 17.5%
* **Cluster 1**: ≈ 16.4%

**Insight**: Cluster 0 devotes a slightly larger share of income to loan repayment, potentially reflecting affordability constraints or more aggressive borrowing among younger applicants.

**🔹 Credit History Length**

* **Cluster 0**: ≈ 3.1 years
* **Cluster 1**: ≈ 9.4 years

**Insight**: A clear divide; Cluster 1 borrowers have substantially more credit experience.

|  |  |  |  |
| --- | --- | --- | --- |
| Feature | Cluster 0 (Younger) | Cluster 1 (Older) | Key Insight |
| Age | ~24 years (low variance) | ~33 years (higher variance) | Strong age separation |
| Income | ~$59K | ~$75K (skewed by outliers) | Moderate separation |
| Employment Length | ~4 years | ~5.7 years | Reflects life stage |
| Loan Amount | ~$9,177 | ~$10,133 | Slightly higher for Cluster 1 |
| Interest Rate | ~10.96% | ~11.08% | No meaningful separation |
| Loan % of Income | ~17.5% | ~16.4% | Slight affordability difference |
| Credit History Length | ~3.1 years | ~9.4 years | Strong credit experience gap |

Table 14: Cluster Characteristics Summary of K-Means clustering (k = 2) in the PCA-reduced space

**7.6.2.3 Conclusion: Evaluating the Usefulness of K-Means Clustering**

The K-Means clustering analysis using **k = 2** identified a basic separation between two borrower segments: one consisting of **younger, lower-income individuals with shorter credit histories**, and another comprising **older, higher-income applicants with longer employment and credit experience**. These findings align with intuitive demographic divisions and show moderate differentiation in personal financial maturity.

However, the effectiveness of the clustering remains limited. The **silhouette score of 0.1987** indicates **weak overall structure**, suggesting that the clusters are not well-separated or strongly cohesive. Additionally, several critical financial variables—most notably **loan amount** and **interest rate**—exhibited **minimal variation** across the clusters. This reduces the interpretability and practical value of the clustering results in relation to credit product differentiation or risk assessment.

Moreover, the presence of **extreme income outliers** in Cluster 1 may have distorted the cluster centroid, affecting the quality of the segmentation. These distortions further weaken the reliability of insights drawn from the cluster profiles.

**In summary**, while the K-Means algorithm revealed some structural patterns within the data, particularly regarding age and credit history, the overall clustering outcome lacks the strength and clarity required for meaningful segmentation. These results imply that the borrower population does not exhibit a strong natural clustering tendency in the PCA-transformed space. As such, **alternative methods**—including supervised modeling or rule-based segmentation strategies—may offer more actionable insights for credit risk analysis and customer profiling.

**7.6.3 Hierarchical Clustering (Agglomerative): PCA-Based Approach**

To further investigate the latent structure within borrower profiles, **Agglomerative Hierarchical Clustering** was applied to the PCA-transformed numerical dataset. Unlike K-Means, which partitions data into clusters by minimizing within-cluster variance, **Hierarchical Clustering** builds a nested tree-like structure (dendrogram) that captures both fine-grained and high-level groupings within the data.

Agglomerative clustering follows a **bottom-up approach**:

* Each observation begins as its **own cluster**.
* At each iteration, the two **most similar clusters** (based on a chosen linkage criterion) are **merged**.
* This process continues until all observations belong to a single cluster or a desired number of clusters is reached.

**Why using PCA?**

Since numerical variables in the dataset span multiple scales and dimensions, **Principal Component Analysis (PCA)** was first used to reduce dimensionality and eliminate multicollinearity. This ensures:

* Better performance of distance-based clustering methods.
* Easier visualisation of cluster structures.
* Enhanced noise reduction.

**Strengths of Hierarchical Clustering**

* No need to pre-specify the number of clusters (though it can be cut at any level).
* Provides a dendrogram, allowing for flexible exploration of the data at various levels of granularity.
* Useful for small to medium-sized datasets where interpretability and structure discovery are essential.

**Linkage Criteria**

The **linkage method** determines how the distance between clusters is computed during the merging process. In this analysis, the **Ward** linkage criterion was employed, as it is well-suited for numerical data and works effectively in combination with PCA. It aims to form compact clusters by minimizing the increase in total within-cluster variance at each merge.

**7.6.3.1 Grid Search Evaluation: Optimal Number of Clusters**

To determine the optimal number of clusters for the **Agglomerative Hierarchical Clustering (Ward linkage)** model, a **grid search** was performed using **Silhouette Score** as the evaluation metric. The silhouette score measures how similar an observation is to its own cluster compared to other clusters, with values ranging from -1 to 1. Higher scores indicate well-separated, cohesive clusters.

**Key Findings:**

* The **highest silhouette score** was observed at **n\_clusters = 2**, with a value of **0.170**.
* Although this is the best-performing configuration, the score itself is **relatively low**, suggesting that **natural separation in the dataset is weak**.
* As the number of clusters increases, the silhouette score generally **declines**, with no meaningful recovery, indicating that additional clusters introduce **overfitting or fragmentation** rather than clarifying the structure.

**Interpretation:**

* The result supports a **2-cluster solution** as the most viable configuration under the given method.
* However, the **low absolute silhouette value (0.170)** implies that the **clustering structure lacks strong definition**, and caution should be taken in interpreting these clusters as robust or highly informative.
* These findings are consistent with previous observations from the K-Means clustering stage and reinforce the dataset’s complex, overlapping nature.

**7.6.3.2 Cluster Profiling:** **Hierarchical Clustering (Ward Linkage, PCA-Transformed Data)**

After identifying 2 as the optimal number of clusters through silhouette-based grid search, we applied **Agglomerative Hierarchical Clustering** with **Ward linkage** to PCA-transformed borrower data. This method minimizes within-cluster variance at each step, which makes it well-suited for identifying compact, spherical clusters in reduced-dimensional numerical space.

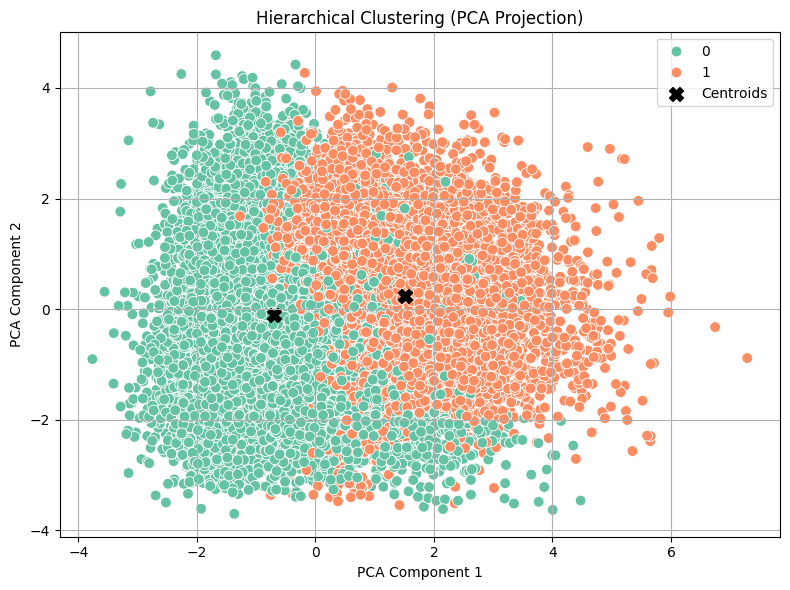


Figure 13: A 2D scatter plot visualising the Hierarchical Clustering (Ward Linkage, PCA-Transformed Data)

**Cluster Distribution**

* **Cluster 0**: 20,145 borrowers (~68%)
* **Cluster 1**: 9,320 borrowers (~32%)

**Demographic Comparison**

|  |  |  |  |
| --- | --- | --- | --- |
| Feature | Cluster 0 (Younger) | Cluster 1 (Older) | Key Insight |
| Mean Age | 25.07 | 33.38 | Cluster 1 contains older borrowers |
| Mean Income | $59.3K | $80.6K | Higher earnings in Cluster 1 |
| Mean Employment Length | 4.0 years | 6.5 years | More work experience in Cluster 1 |
| Mean Credit History Length | 4.0 years | 9.6 years | Cluster 1 more credit-experienced |

Table 15: Demographic Comparison of Hierarchical Clustering (k = 2) in the PCA-reduced space

**Loan Behaviour Comparison**

|  |  |  |  |
| --- | --- | --- | --- |
| Feature | Cluster 0 | Cluster 1 | Key Insight |
| Mean Loan Amount | $8,573 | $11,772 | Older borrowers take larger loans |
| Mean Interest Rate | 11.02% | 10.99% | Rates nearly identical between groups |
| Loan as % of Income | 16.35% | 18.44% | Cluster 1 commits more income to loan repayment |

Table 16: Loan Behaviour Comparison of Hierarchical Clustering (k = 2) in the PCA-reduced space

**Overall Interpretation**

* **Cluster 0: Younger, Lower Income Group**
* **Demographics**: Young adults (~25), shorter employment and credit history.
* **Financial Profile**: Lower income ($59K), smaller loans ($8.6K), slightly lower debt-to-income ratio.
* **Implication**: Likely early-career borrowers with limited credit history; lower risk from affordability, but greater uncertainty from lack of financial track record.
* **Cluster 1: Older, More Established Group**
* **Demographics**: More mature (~33), with longer job tenure and credit history.
* **Financial Profile**: Higher income ($81K), larger loans ($11.7K), and higher loan burden relative to income.
* **Implication**: Represents financially mature borrowers, potentially riskier due to higher relative loan burden despite better credit profiles.

**7.6.3.3 Conclusion: Evaluating the Usefulness of Hierarchical Clustering**

The application of **Agglomerative Hierarchical Clustering with Ward linkage** yielded two primary clusters, which align reasonably well with demographic and financial distinctions in the dataset. Cluster 0 primarily contains younger, less experienced borrowers with shorter credit histories and smaller loans, while Cluster 1 consists of older, more established individuals with higher income and longer credit histories.

However, despite the intuitive interpretability of these groupings, **the overall silhouette score of 0.17 indicates weak structural cohesion and separation between clusters**. While this score is slightly better than random, it still falls within the "no substantial structure" range, suggesting that the clusters are not well-defined in the multidimensional feature space.

**Limitations and Considerations**

* The clustering may reflect some true underlying demographic divisions, but lacks strong support from internal validation metrics.
* Core financial variables such as **interest rate** show minimal differentiation between clusters, raising questions about the practical utility of these segments.
* The presence of **income outliers** and uneven feature contributions may dilute the effectiveness of clustering.

In conclusion, while hierarchical clustering using Ward linkage provides some descriptive value, it does not uncover strongly distinct borrower segments. Therefore, it may be more useful as an exploratory tool rather than a foundation for strategic classification or predictive modelling.

**7.6.4 DBSCAN (Density-Based Clustering) on PCA-Transformed Data**

**DBSCAN (Density-Based Spatial Clustering of Applications with Noise)** is a robust, unsupervised clustering algorithm that identifies clusters as high-density regions separated by areas of low point density. Unlike K-Means and hierarchical clustering, DBSCAN **does not require the number of clusters to be specified in advance** and can identify **clusters of arbitrary shape**, as well as detect **noise or outliers**.

This method is particularly useful in real-world datasets where the data distribution is irregular and does not conform to spherical or nested groupings. DBSCAN works by evaluating two key parameters:

* **ε (epsilon)**: The radius of the neighborhood around a data point.
* **minPts**: The minimum number of neighboring points required to form a dense region (i.e., a core point).

If a point has at least minPts neighbors within the ε radius, it is considered a **core point**, and a cluster is formed by expanding this core region to include density-reachable points. Points that do not belong to any cluster are treated as **noise**, making DBSCAN inherently capable of **outlier detection**.

**Application in This Study**

In this study, DBSCAN was applied to the **PCA-transformed dataset** to leverage its ability to identify non-linear and non-spherical patterns in reduced dimensions. PCA was used to condense the high-dimensional numerical space into a few principal components, which helped eliminate noise and reduce redundancy, thus improving the effectiveness of DBSCAN in defining meaningful clusters.

**Advantages in This Context**

* **No assumption of cluster count**: Ideal when the underlying number of borrower segments is unknown.
* **Outlier detection**: Points that do not conform to dense regions are flagged as noise.
* **Shape flexibility**: Useful when clusters are not globular or equally sized, as in real financial behaviour data.

This method provides a unique perspective compared to K-Means or Hierarchical Clustering by emphasizing **density patterns** and **local structure** within the dataset. It is especially suited to uncover **non-obvious borrower subgroups** and detect anomalies that might be missed by other clustering techniques.

**7.6.4.1 Grid Search Evaluation**

The DBSCAN hyperparameters (ε, minPts) were tuned via grid search. Across different combinations, the optimal clustering outcome consistently revealed **a single major cluster**, a very small secondary cluster, and a notable number of noise points.

**Summary:**

* **Cluster 0**: Dominant, homogeneous group.
* **Cluster 1**: Tiny cluster with highly specific profiles (only 7 individuals).
* **Cluster -1**: Detected as noise (176 observations).

Despite attempts to increase the number of clusters, the structure of the data naturally resisted further segmentation — reinforcing the presence of a **largely uniform applicant base** with **a few distinct outliers**.

**7.6.4.2 Cluster Profiling & Interpretation**

The final DBSCAN output was as follows:

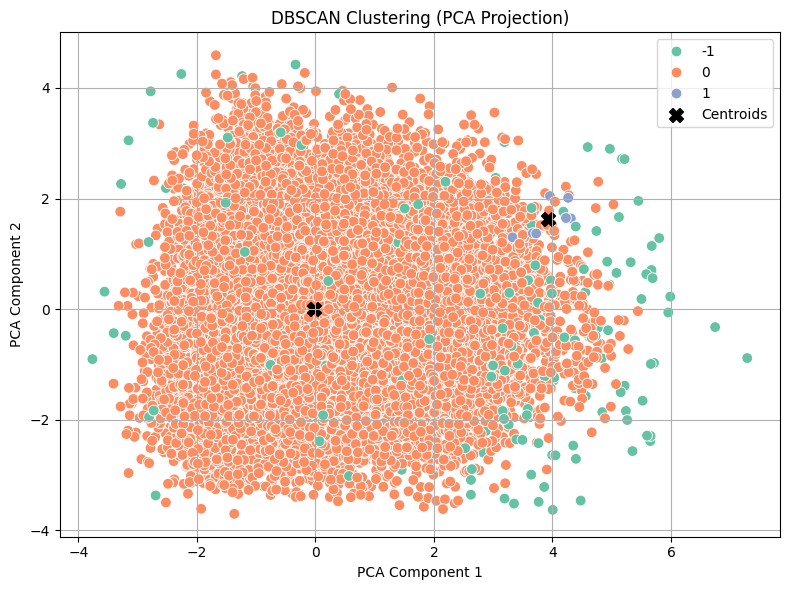


Figure 14: A 2D scatter plot visualising the DBSCAN Clustering (PCA-Transformed Data)

|  |  |  |
| --- | --- | --- |
| Cluster | Size | Key Traits |
| Cluster 0 | 29,282 | Mainstream borrowers – younger, moderate income, standard loan ratios |
| Cluster 1 | 7 | Older, low-income, high loan-to-income ratio (40%), long credit history |
| Noise (-1) | 176 | Outliers – very high income (mean: ~$270k), diverse age and employment |

Table 17: DBSCAN Clustering Output

**Key Interpretations:**

* **Cluster 0: Mainstream Segment**
  + **Age**: Mean 27.6 years.
  + **Income**: Mean ~$64,800.
  + **Credit History**: ~5.7 years.
  + **Loan % of Income**: ~17%.
  + Represents the **typical borrower profile** — younger, employed, modestly leveraging income.
* **Cluster 1: Niche, High-Risk Segment**
  + **Age**: Mean 56.4 years.
  + **Income**: Lower (~$45,000).
  + **Loan as % of Income**: 40%.
  + Despite long credit histories, this group may pose risk due to disproportionate loan burden.
* **Noise Cluster (-1): Outlier Detection**
  + **Income**: Extreme values (mean ~$270k, max $6M).
  + **Credit History**: Long (avg. ~15.8 years).
  + Likely represents **wealthy, anomalous profiles** not representative of the main applicant base.

**7.6.4.3 Overall Conclusion**

DBSCAN has proven effective in **validating the homogeneity** of the applicant population:

* **One major cluster** was consistently detected — representing the majority of applicants with standard financial characteristics.
* No strong natural substructures were found beyond this, reinforcing earlier findings (from K-Means and Hierarchical Clustering) that segmentation potential in this dataset is inherently weak.
* However, DBSCAN's **outlier identification** is highly valuable — especially for credit risk strategy, fraud detection, or specialized borrower handling.
* Business Insight: *“Our applicant base is broadly uniform, with a small subset of significantly different individuals. These outliers may require separate consideration in risk or product design policies.”*

**7.6.5** **K-Modes Clustering for Categorical Features**

K-Modes clustering is a specialized unsupervised learning technique designed to group observations based purely on categorical variables. It extends the popular K-Means algorithm by addressing a critical limitation — K-Means relies on Euclidean distance, which is not meaningful for categorical data.

K-Modes introduces the following innovations to adapt clustering to categorical domains:

* **Dissimilarity Measure**: Instead of using distance, K-Modes measures dissimilarity based on the number of mismatches between categorical features (also called the Hamming distance).
* **Cluster Centroids**: Each cluster center (or "mode") is defined by the most frequent category (the mode) for each feature across all members of the cluster.
* **Mode Update Rule**: During each iteration, cluster modes are updated by recomputing the most common category in each dimension — making the algorithm computationally efficient and easy to interpret.

In this study, K-Modes clustering was applied to the subset of borrower features consisting solely of categorical variables: person\_home\_ownership, loan\_intent, loan\_grade, and cb\_person\_default\_on\_file. Unlike numerical clustering methods, K-Modes groups borrowers by minimizing the number of mismatches across categorical fields, and centroids represent the most frequent category (mode) in each feature dimension.

To visualize the clusters, we used **Multiple Correspondence Analysis (MCA)** to project the high-dimensional categorical data into a **2D space**, enabling intuitive inspection of cluster separation and cohesion. The resulting scatter plot reveals how well each cluster captures distinct borrower segments based on categorical behaviour and risk indicators.

**7.6.5.1 Cluster Profiles and Interpretation**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Cluster | home\_ownership | loan\_intent | loan\_grade | default\_on\_file | Count |
| 0 | RENT | EDUCATION | B | N | 12,428 |
| 1 | MORTGAGE | PERSONAL | A | N | 8,475 |
| 2 | RENT | VENTURE | A | N | 3,411 |
| 3 | RENT | EDUCATION | D | Y | 2,486 |
| 4 | RENT | MEDICAL | C | Y | 2,665 |

Table 18: Descriptive summary of K-Modes Clustering for Categorical Features

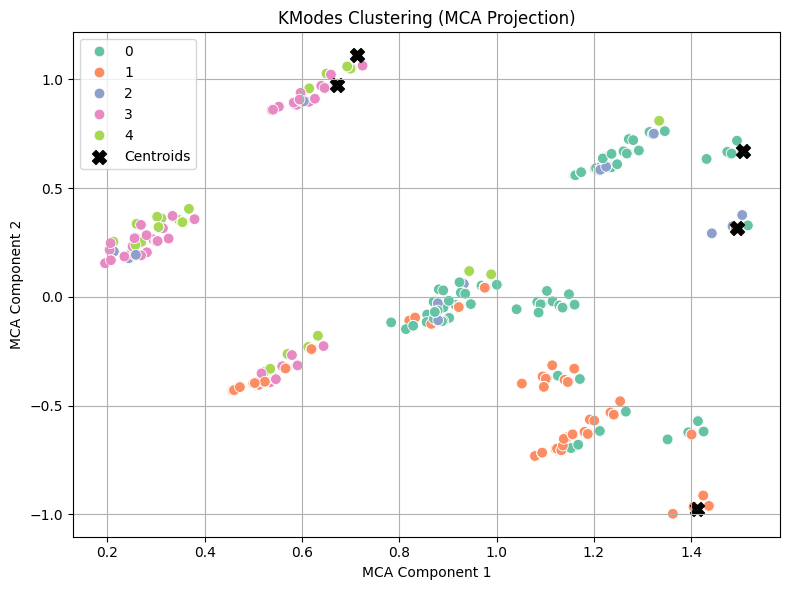


Figure 15: The MCA-based visualization of K-Modes Clustering for Categorical Features

**Cluster 0 – Young, Educated, Creditworthy Renters**

* **Home Ownership**: RENT
* **Loan Purpose**: EDUCATION
* **Grade**: B
* **Default History**: No
* **Insight**: Likely younger, upwardly mobile individuals seeking education loans. Generally moderate risk with potential for growth.

**Cluster 1 – Established, Prime Borrowers**

* **Home Ownership**: MORTGAGE
* **Loan Purpose**: PERSONAL
* **Grade**: A
* **Default History**: No
* **Insight**: Stable, creditworthy homeowners. Ideal candidates for premium credit products or cross-sell opportunities.

**Cluster 2 – Entrepreneurial Renters**

* **Home Ownership**: RENT
* **Loan Purpose**: VENTURE
* **Grade**: A
* **Default History**: No
* **Insight**: Ambitious renters starting or expanding businesses. Low-risk but may lack collateral. Support with tailored business loan packages.

**Cluster 3 – At-Risk Student Borrowers**

* **Home Ownership**: RENT
* **Loan Purpose**: EDUCATION
* **Grade**: D
* **Default History**: Yes
* **Insight**: Risky profile — history of defaults and weak credit. May need credit repair tools or secured lending.

**Cluster 4 – Medically Driven Borrowers with Risk**

* **Home Ownership**: RENT
* **Loan Purpose**: MEDICAL
* **Grade**: C
* **Default History**: Yes
* **Insight**: Financially strained group. May benefit from debt restructuring or medical support credit solutions.

**7.6.5.2 Strategic Business Recommendations**

|  |  |  |
| --- | --- | --- |
| Cluster | Risk Level | Suggested Business Action |
| 0 | Low / Medium | Promote flexible education loans, upsell later |
| 1 | Very Low | Offer premium credit options, loyalty plans |
| 2 | Low | Design entrepreneurial packages with growth potential |
| 3 | High | Restrict loan offerings; suggest financial literacy or secured credit |
| 4 | Medium / High | Support with flexible repayment or relief plans for medical circumstances |

Table 19: Strategic Business Recommendations based on K-Modes Clustering results

**7.6.6 Hierarchical Clustering on MCA-Transformed Data**

In this section, we apply Hierarchical Clustering to categorical features, using the Multiple Correspondence Analysis (MCA) transformation to enable clustering in a continuous space.

While Hierarchical Clustering is traditionally used with numerical data, combining it with MCA allows us to uncover meaningful groupings within purely categorical datasets. MCA reduces high-dimensional categorical variables to a smaller set of interpretable numeric components that preserve the core structure of the data. By clustering on these components, we can reveal latent borrower segments based on categorical attributes such as loan intent, credit grade, home ownership, and default history.

We used the Ward linkage method, which minimizes the total within-cluster variance and tends to produce compact, spherical clusters. This approach supports a data-driven segmentation of borrower profiles, aiding business understanding, risk stratification, and potential policy design.

**7.6.6.1 Cluster Selection via Grid Search**

To determine the optimal number of clusters for Hierarchical Clustering on MCA-transformed categorical features, a grid search was conducted by varying the number of clusters and evaluating the Silhouette Score for each configuration.

|  |  |
| --- | --- |
| Number of Clusters | Silhouette Score |
| 2 | 0.478 |
| 3 ✅ | **0.532** |
| 4 | 0.501 |
| 5 | 0.506 |
| 6 | 0.514 |
| 7–10 | < 0.510 |

Table 20:Grid search result for Hierarchical Clustering on MCA-transformed categorical features

**Interpretation**

* The highest Silhouette Score was observed at 3 clusters (score = 0.532), indicating the most distinct and cohesive clustering structure.
* Scores above 0.50 are generally interpreted as evidence of a good clustering structure. In this analysis, all configurations from 3 to 6 clusters maintained scores above 0.50, suggesting a degree of stability in this range.
* However, the peak score at k = 3 demonstrates the most optimal separation and cohesion between cluster groups.

**Conclusion**

Based on the grid search, three clusters were selected as the optimal structure for segmenting the dataset using Hierarchical Clustering on MCA coordinates. This choice is both statistically sound and supported by strong silhouette metrics, ensuring the resulting clusters are meaningfully distinct.

**7.6.6.2 Interpretation of Cluster Profiles**

Following the optimal selection of **three clusters** using Hierarchical Clustering on MCA-transformed categorical data, each cluster was profiled based on the dominant categorical characteristics (person\_home\_ownership, loan\_intent, loan\_grade, and cb\_person\_default\_on\_file). The results are summarized below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Cluster | Home Ownership | Loan Intent | Loan Grade | Default History |
| 0 | RENT | EDUCATION | C | Yes |
| 1 | RENT | EDUCATION | A | No |
|  |  |  |  |  |
| 2 | OWN | VENTURE | A | No |

Table 21: Profile summary using Hierarchical Clustering on MCA-transformed categorical data

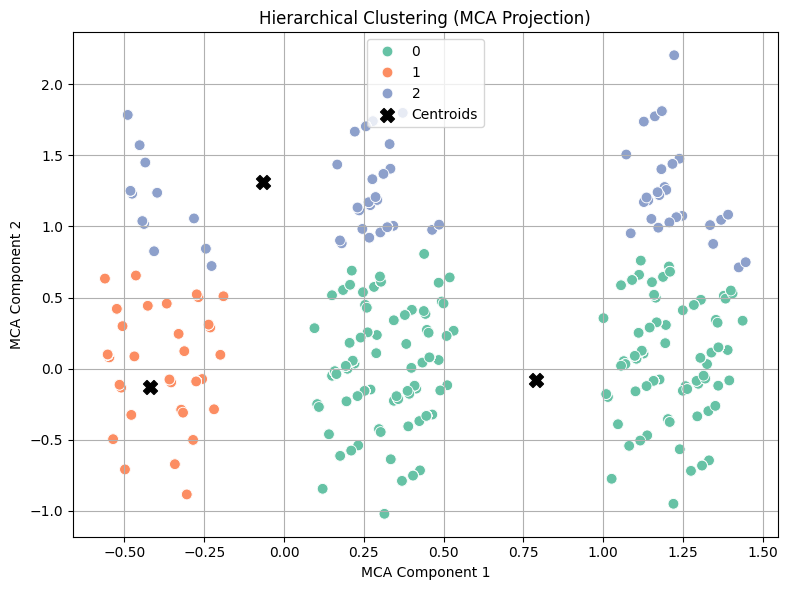


Figure 16: The MCA-based visualization of Hierarchical Clustering for Categorical Features

**Cluster Interpretation**

**🔹 Cluster 0 – "Risky Students"**

* Borrowers are predominantly renters taking loans for education.
* They possess a lower credit grade (C) and a history of default.
* **Business Insight**: This cluster likely consists of younger, financially vulnerable individuals. While their educational intent may reflect long-term investment in human capital, their risk profile warrants cautious lending terms, additional verification, or financial literacy interventions.

**🔹 Cluster 1 – "Safe Students"**

* Similar demographic to Cluster 0 (also renters with educational loan intents), but with top-tier credit grades (A) and no default history.
* **Business Insight**: Represents a segment of responsible, low-risk student borrowers. These individuals could be prioritized for competitive loan packages or early-career customer engagement strategies.

**🔹 Cluster 2 – "Affluent Entrepreneurs"**

* Characterized by homeowners borrowing for venture/business purposes, with high credit grades (A) and no history of default.
* **Business Insight**: This group reflects financially stable, entrepreneurial borrowers. They are excellent candidates for business loan products, credit line expansions, or premium banking services.

**7.6.6.3 Business Recommendations and Conclusion**

* The optimal cluster count of 3, supported by a high silhouette score (0.532), confirms that the clustering structure is both statistically robust and business-relevant.
* The segmentation yields clearly differentiated borrower profiles:
  + **Cluster 0**: Risk-prone students — higher monitoring or restricted products advised.
  + **Cluster 1**: Low-risk students — offer favourable terms and nurture long-term relationships.
  + **Cluster 2**: Wealthier, entrepreneurial clients — cross-sell investment or business offerings.

These findings support the integration of cluster-based borrower segmentation into risk scoring models, loan policy formulation, and targeted marketing strategies. Hierarchical clustering on categorical data thus provides actionable insights that enhance both credit risk management and customer engagement.

**7.6.7 K-Prototypes Clustering (Mixed-Type Data)**

Clustering datasets that contain both numerical and categorical variables poses unique challenges, as traditional algorithms like K-Means or K-Modes are designed for single data types. To address this, K-Prototypes extends the strengths of both methods, enabling clustering on mixed-type data by combining numerical distances (e.g., Euclidean) with categorical dissimilarity measures (e.g., Hamming distance).

At its core, K-Prototypes:

* Uses **Euclidean distance** for numerical variables,
* Uses **Hamming distance** (i.e., mismatch count) for categorical variables,
* Introduces a **tunable parameter (γ)** to balance the influence of numeric vs. categorical features.

This hybrid approach makes K-Prototypes especially valuable for datasets like ours, where borrower profiles are defined by a combination of quantitative financial variables (e.g., income, age, loan amount) and qualitative factors (e.g., home ownership, loan intent, credit grade).

In this study, K-Prototypes was applied to the complete post-cleaning dataset, allowing us to extract composite borrower segments that reflect both financial behaviour and categorical characteristics. The resulting clusters offer a nuanced lens into the structure of the borrower population, which is crucial for risk stratification, personalized loan strategies, and portfolio diversification.

**7.6.7.1 Cluster Selection Strategy**

Unlike previous clustering methods, we did not conduct a full grid search to determine the optimal number of clusters for K-Prototypes. The computational cost of evaluating multiple cluster configurations on a mixed-type dataset was prohibitively high, with estimated runtimes exceeding 18 hours on the available hardware.

To maintain methodological consistency and efficiency, we instead leveraged insights from the Hierarchical Clustering (MCA-based) method applied earlier to categorical data. That analysis identified three clusters as the most statistically and structurally appropriate solution, supported by the highest silhouette score and coherent business segments. Given the conceptual overlap in feature space between the two approaches, this provided a justified and pragmatic basis for selecting k = 3 in the K-Prototypes model.

This strategy ensured analytical alignment across clustering phases while balancing performance constraints and interpretability.

**7.6.7.2 Interpretation of K-Prototypes Clustering Results**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Cluster | Dominant Home Ownership | Dominant Loan Intent | Loan Grade | Default History | Age (Mean ± Std) | Income (Mean ± Std) | Loan Amount (Mean ± Std) | Interest Rate (Mean ± Std) | Loan % of Income (Mean ± Std) | Credit History Length (Mean ± Std) |
| 0 | MORTGAGE | MEDICAL | A | No Defaults | 34.3 ± 6.6 | 85.3k ± 98.4k | 10.1k ± 6.2k | 10.81% ± 3.25% | 14.3% ± 8.2% | 10.2 years ± 4.0 |
| 1 | RENT | EDUCATION | A | No Defaults | 24.6 ± 2.6 | 57.7k ± 32.1k | 5.3k ± 2.8k | 10.18% ± 3.11% | 10.4% ± 5.2% | 3.6 years ± 1.7 |
| 2 | RENT | EDUCATION | B | No Defaults | 25.1 ± 3.0 | 57.4k ± 32.1k | 14.1k ± 6.0k | 12.16% ± 3.04% | 27.3% ± 9.9% | 4.1 years ± 2.1 |

Table 22: Summary of K-Prototypes Clustering Results for Mixed-Type Data

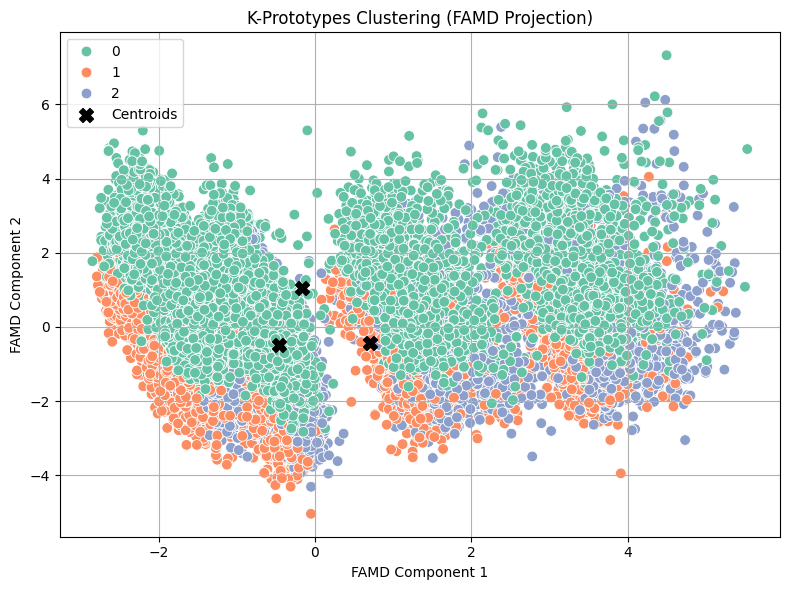


Figure 17: FAMD Projection of K-Prototypes Clustering for Mixed-Type Data

**🔹 Cluster 0 – Mature & Financially Stable (Medical Borrowers)**

* **Demographics**: Older average age (~34), long credit history (~10 years), primarily homeowners.
* **Financials**: High income (~$85K), lower loan-to-income ratio (~14%), and lowest interest rates (~10.8%).
* **Purpose**: Medical loans.
* **Credit Profile**: High credit grade (A), no default history.
* **💼 Business Implication**: These individuals are low-risk and likely eligible for premium financial products. Consider targeting with **retirement planning**, **insurance packages**, or **extended healthcare financing**.

**🔹 Cluster 1 – Young, Low-Risk Educational Borrowers**

* **Demographics**: Youngest group (~24.6 years), renters, short credit histories (~3.6 years).
* **Financials**: Modest income ($57K), small loan amounts ($5.3K), and low interest rates (~10.2%).
* **Purpose**: Education loans.
* **Credit Profile**: Grade A borrowers, no default history.
* **📈 Business Implication**: This group shows high promise with strong repayment potential. Consider offering **student banking**, **early investment products**, and **first-time homebuyer education**.

**🔹 Cluster 2 – Young Borrowers with Higher Financial Strain**

* **Demographics**: Similar to Cluster 1 in age and income, but borrow significantly more.
* **Financials**: High loan-to-income ratio (27%), interest rates above average (12.2%), and larger loan amounts ($14.1K).
* **Purpose**: Still education-focused.
* **Credit Profile**: Grade B borrowers, no prior defaults but higher perceived risk.
* **Risk Alert**: This segment may be overleveraged. Recommend **risk monitoring**, **repayment planning tools**, or **secured loan products** to reduce exposure.

**7.6.7.3 Business Insights and Strategic Takeaways**

* Segmented insights from K-Prototypes offer actionable borrower archetypes:
* Mature medical borrowers with low risk (Cluster 0)
* Promising students with low credit exposure (Cluster 1)
* Overstretched young borrowers needing closer supervision (Cluster 2)
* The absence of prior defaults across all clusters indicates that observed risk is likely emerging from current loan behaviour and structural financial pressures rather than historical delinquencies.
* Despite some overlap in centroid positions in the FAMD-reduced 2D space (a known effect of dimensionality reduction), the numerical spread in key indicators validates meaningful separation.

**7.6.8 Hierarchical Clustering with Gower Distance (FAMD-Reduced)**

To address the challenges of clustering datasets with both numerical and categorical variables, Hierarchical Clustering with Gower Distance was employed. Unlike traditional distance metrics such as Euclidean, Gower distance is specifically designed to handle mixed data types, enabling fair comparison across categorical, ordinal, and numerical features.

In this study, Gower distances were computed on the original mixed-feature dataset and then projected into a lower-dimensional space using Factor Analysis of Mixed Data (FAMD). This reduction step preserves key structural variance while simplifying the high-dimensional relationships for more stable clustering.

We then applied agglomerative hierarchical clustering on the reduced coordinates using the Ward linkage method, which minimizes the variance within each cluster during the merging process.

**7.6.8.1 Cluster Number Selection Strategy**

Due to the computational intensity of calculating Gower distance matrices on a large mixed-type dataset, performing a full grid search to optimize the number of clusters was not feasible within the available processing resources. Hierarchical clustering with Gower distance requires a pairwise distance matrix, which scales quadratically with dataset size, making exhaustive cluster evaluation prohibitively time-consuming.

Instead, we predefined the number of clusters as three (n=3). This decision was based on prior findings from the Hierarchical Clustering on MCA-transformed categorical data, which indicated that three clusters offered a statistically and conceptually sound segmentation structure. Leveraging this prior insight ensured consistency across clustering strategies while maintaining analytical efficiency.

**7.6.8.2 Interpretation of Hierarchical Clustering Results (Gower Distance with FAMD Coordinates)**

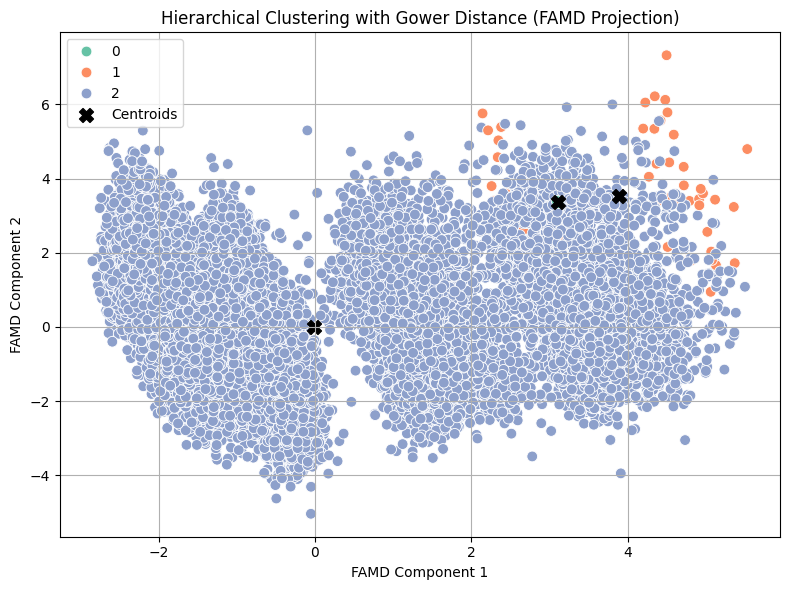


Figure 18: FAMD Projection of Hierarchical Clustering with Gower Distance for Mixed-Type Data

For detailed cluster-wise statistics and feature distributions, please refer to the Appendix section. The interpretation of each cluster is provided below:

**Cluster 0 – Elderly Medical Borrowers with Stretched Finances**

* **Demographic**: Older individuals (~60 years), with exceptionally long employment histories (~31.5 years).
* **Loan Purpose**: Medical
* **Credit Behaviour**: Prior defaults; grade C; high interest rate (~13.5%)
* **Loan-to-Income Ratio**: ~24.5% of income used toward loan repayments.
* **Interpretation**: This cluster likely represents retiree or near-retiree borrowers, possibly dealing with health-related financial burdens. Despite long credit histories (~26 years), their financial stress is evident in their high DTI and history of default.
* **Risk Level**: High

**Cluster 1 – Entrepreneurial High-Income Borrowers with Poor Credit**

* **Demographic**: Young-to-mid-age (~28 years), well-employed (~6 years), with mortgage homes.
* **Loan Purpose**: Venture/business
* **Credit Behaviour**: Grade G (lowest), past defaults, and very high interest rate (~20.3%)
* **Loan-to-Income Ratio**: ~25.4% — the highest among all clusters
* **Interpretation**: This group is financially ambitious but risk-prone — high loan amounts, poor credit grades, and risky loan intents. While they earn well (~80k), they appear overleveraged and historically unreliable.
* **Risk Level**: High

**Cluster 2 – Young, Creditworthy Education Borrowers**

* **Demographic**: Young (~27.7 years), modest work history (~4.8 years), mostly renters.
* **Loan Purpose**: Education
* **Credit Behaviour**: No default history, Grade A, and moderate interest rate (~11%)
* **Loan-to-Income Ratio**: ~17% — better controlled than other clusters.
* **Interpretation**: This is a low-risk, upwardly mobile group of borrowers. Their lower loan amounts, better grades, and reliable history make them attractive for responsible lending.
* **Risk Level**: Low

**7.6.8.3 Business Recommendations and Strategic Conclusion**

|  |  |  |
| --- | --- | --- |
| Cluster | Risk Level | Business Actions |
| Cluster 0 | High | Introduce **debt restructuring options**, **medical-specific plans**, or **pension-linked lending**. Consider tightened lending criteria. |
| Cluster 1 | High | Flag for **enhanced due diligence**. Offer **secured business loans** or link loans to mentoring/credit improvement tools. Monitor aggressively. |
| Cluster 2 | Low | Promote **education-focused loan products**, early-career support tools, and **graduated credit lines** as income grows. Potential for long-term relationship building. |

Table 23: Business Recommendations based on the Hierarchical Clustering Results (Gower Distance with FAMD Coordinates)

This clustering model has successfully uncovered three actionable borrower segments:

1. Older borrowers in financial distress (Cluster 0),
2. Risk-tolerant entrepreneurs with unstable credit history (Cluster 1),
3. Young and responsible education-focused individuals (Cluster 2).

While no grid search was conducted due to computational limitations, the cluster structure aligns closely with prior segmentation outcomes, confirming conceptual consistency and providing a strong foundation for targeted credit strategy design.

These insights can be used to support:

* Risk-tiered credit policies,
* Personalized loan offerings,
* Cross-selling strategies based on borrower lifecycle stages.

**7.7 Summary & Next Steps**

The exploratory data analysis provided a comprehensive understanding of both the numerical and categorical aspects of the dataset. We assessed variable distributions, correlations, and multicollinearity, and explored the relationships between key borrower features and credit outcomes. Clustering techniques—applied across numerical, categorical, and mixed data—revealed latent borrower segments, though with varying degrees of structural clarity. Dimensionality reduction aided in visualizing high-dimensional relationships, and these collective insights laid a strong foundation for the next phase. Moving forward, the focus will shift to **feature engineering and selection**, where we will derive new variables, and apply statistical and algorithmic methods to identify the most predictive and non-redundant features. This step is essential for enhancing model accuracy and interpretability.

**8.** **Feature Engineering and Selection**

Feature engineering and selection play a crucial role in credit risk and Probability of Default (PD) modelling, as they directly impact the predictive power, interpretability, and regulatory compliance of the model. **Feature engineering** refers to the process of transforming raw data into meaningful input variables that improve model performance, while **feature selection** involves identifying the most relevant variables to enhance stability and reduce overfitting.

In PD modelling, feature engineering techniques can include:

* **Creating derived features** – For example, calculating the loan-to-income ratio (loan\_amnt / person\_income) to capture the borrower’s debt burden.
* **Binning (discretization)** – Converting continuous variables (e.g., income, age) into discrete intervals or categories to handle non-linear relationships and improve interpretability.
* **Clustering-based features** – Using unsupervised learning methods such as hierarchical clustering to group similar borrowers and add cluster membership as a new feature.
* **Non-linear transformations** – Applying mathematical transformations (e.g., squares, logarithms, or higher powers) to intensify or normalize the effect of certain features when relationships are not strictly linear.

Among these, binning combined with Weight of Evidence (WoE) encoding is widely used in credit risk modelling because it provides monotonic relationships between predictors and default probability and ensures compliance with regulatory frameworks such as Basel II/III (BCBS 2004; BCBS 2006). WoE measures the strength and direction of a predictor in differentiating between good and bad borrowers, while Information Value (IV) quantifies the predictive power of the variable. These measures allow transparent, interpretable, and auditable models, which are crucial for regulatory acceptance, especially when using logistic regression for PD estimation.

**8.1 Binning and Weight of Evidence (WoE) Transformation**

**8.1.1 Binning**

To prepare numerical features for logistic regression in Probability of Default (PD) modeling, we applied supervised binning. Binning discretizes continuous variables into categories that maintain predictive relationships with the target variable while reducing model complexity and ensuring interpretability. This step is especially important under regulatory frameworks such as Basel II/III, where models must demonstrate transparency and stability (BCBS 2004, 2006; Siddiqi, 2006).

**Why Binning?**

* Captures **non-linear relationships** between predictors and default probability.
* Handles outliers and improves model robustness.
* Facilitates the calculation of **Weight of Evidence (WoE)**, which encodes variables into monotonic transformations aligned with the target variable.
* Ensures interpretability for regulators and model validation teams.

**Methodology**

We used **OptimalBinning**, a Python implementation of the widely used **Monotonic Optimal Binning (MOB)** approach. This method:

* Maximizes separation between good and bad loans (target-aware).
* Automatically determines optimal cut-points for each feature.
* Produces bins with monotonic WoE patterns, which are ideal for logistic regression.

The suggested bin edges for selected features are shown below:

|  |  |
| --- | --- |
| Feature | Suggested Bin Edges |
| Age | [22.5, 25.5, 28.5, 29.5] |
| Income | [23002.0, 34999.5, 39937.5, 49994.0, 59982.0, 79942.5, 108811.5] |
| Employment Length | [0.5, 1.5, 2.5, 4.5, 6.5, 7.5, 11.5] |
| Loan amount | [3262.5, 5087.5, 6087.5, 7387.5, 8237.5, 10587.5, 12675.0, 18087.5, 22150.0] |
| Interest rate | [6.46, 7.46, 7.89, 9.63, 11.27, 12.06, 12.72, 13.61, 14.36, 15.27, 16.30] |
| Loan percent income | [0.055, 0.075, 0.135, 0.155, 0.175, 0.205, 0.255, 0.305, 0.375] |
| Credit hist length | [2.5, 3.5, 4.5, 6.5, 7.5, 13.5] |

The figure below illustrates the distribution of records across bins all feature.

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

Figure 19: Distribution of Binned features using OptimalBinning algorithm

**8.1.2 Weight of Evidence (WoE) Transformation**

Weight of Evidence (WoE) is a widely used technique in **credit risk modelling**, especially in **logistic regression-based Probability of Default (PD) models**. It helps quantify the relationship between predictor variables and the target variable (default vs. non-default) in a way that is both interpretable and mathematically suitable for linear models.

WoE transformation is particularly useful because:

* It **linearizes non-linear relationships** between predictors and the log-odds of default.
* It **stabilizes variables** and reduces the impact of outliers.
* It provides a **monotonic relationship** with the target variable.
* It allows for **better interpretability**, making it easier to explain model behaviour.

**Formula**

The WoE for each bin i is calculated as:

Where:

* **Goods** = Non-default customers
* **Bads** = Default customers

**Interpretation:**

* **WoE > 0** → Bin has more non-defaults than defaults → Lower credit risk.
* **WoE < 0** → Bin has more defaults than non-defaults → Higher credit risk.
* **WoE ≈ 0** → Neutral relationship.

**Implementation**

In this study:

* Continuous features such as loan\_amnt, person\_age, person\_income, person\_emp\_length, loan\_int\_rate, loan\_percent\_income, and cb\_person\_cred\_hist\_length were binned using OptimalBinning.
* WoE values were computed for each bin to evaluate the relationship with credit risk.

**General Observations**

* Features like **income, loan amount, and interest rate** show strong monotonic relationships with default risk.
* **Employment length and loan-to-income ratio** also have clear predictive power.
* **Age and credit history length** show mild or neutral predictive power compared to the others.

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

Figure 20: WoE plots for all features’ bins, showing the monotonic relationship between them and credit risk

**WoE Values by Feature and Bin**

The following table summarizes selected WoE values for key bins across features:

|  |  |  |  |
| --- | --- | --- | --- |
| Binned Features | Bin Index | WoE Value | Interpretation |
| Loan amount | 0 | +0.016 | Low-risk, close to neutral |
| 2 | +0.443 | High non-default ratio (safe) |
| 9 | -0.732 | High default risk |
| Age | 0 | -0.210 | Younger age band, slightly riskier |
| 4 | +0.082 | Older, slightly safer |
| Income | 0 | -1.752 | Very low income → high risk |
| 7 | +1.076 | Very high income → very safe |
| Employment length | 0 | -0.335 | Very short employment → riskier |
| 7 | +0.421 | Long employment → more stable |
| Interest rate | 0 | +1.449 | Very low rate → very low risk |
| 11 | -1.844 | Very high rate → very high risk |
| Loan percent to income | 0 | +0.893 | Small % of income used → low risk |
| 9 | -2.284 | Large % of income used → high risk |
| Credit history length | 0 | -0.099 | Slightly more defaults in short histories |
| 4 | +0.114 | Safer with longer history |

Table 24: Selected WoE values for key bins across features

**Feature-Wise Interpretation**

1. **Loan Amount**
   * Bins 2–5 have positive WoE → relatively safer loan sizes.
   * Bins 7–9 show negative WoE, especially bin 9 (-0.73) → higher loan amounts are riskier.
   * **Insight:** Larger loans are associated with increased default risk.
2. **Person Income**
   * Very strong monotonic trend: low income → negative WoE → high risk; high income → positive WoE → low risk.
   * **Insight:** Income is a strong predictor of creditworthiness.
3. **Loan Interest Rate**
   * WoE decreases as interest rate increases.
   * **Insight:** Higher interest rates indicate higher credit risk.
4. **Loan-to-Income Ratio**
   * Low ratios → positive WoE → safer; high ratios → highly negative WoE → risky.
   * **Insight:** Over-indebtedness is a major risk factor.
5. **Employment Length**
   * Longer employment history correlates with higher WoE.
   * **Insight:** Job stability improves creditworthiness.
6. **Age**
   * Younger applicants slightly riskier; older ones safer.
   * **Insight:** Mild predictive power compared to income or interest rate.
7. **Credit History Length**
   * Mostly neutral, slight positive trend with longer history.
   * **Insight:** Less influential than other variables.

**8.1.3 Information Value (IV)**

To assess the predictive power of each feature in distinguishing between good (non-default) and bad (default) outcomes, we calculated the **Information Value (IV)** for all binned variables. IV complements WoE by summarizing the discriminatory power of a feature across all its bins. It is widely used in credit risk modelling and recommended by regulatory frameworks such as Basel for feature selection.

The formula for IV is:

Where:

* **i**: Represents a specific category or bin of the variable (from 1 to n total bins).
* **Distr. Goodsi​**: The proportion of all "Good" outcomes that fall into bin i.
* **Distr. Badsi​**: The proportion of all "Bad" outcomes that fall into bin i.
* WoEi = Weight of Evidence for bin𝑖

**Information Value (IV) results and interpretation**

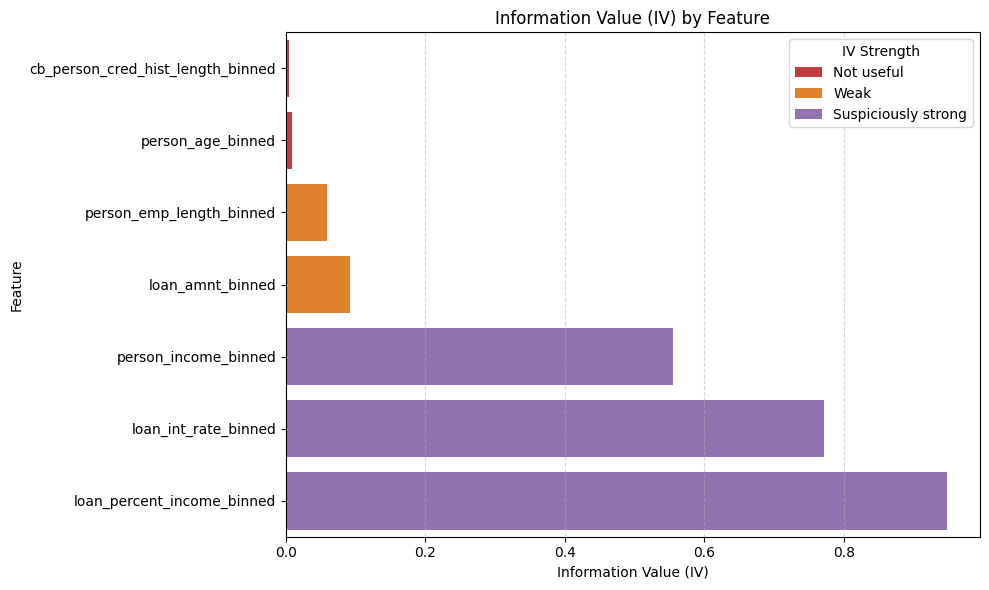


Figure 21: Information Value Results for all binned features

* **loan\_percent\_income\_binned (IV = 0.9477)**: Exhibits extremely strong predictive power. Such a high IV value, while indicative of strong discrimination, may suggest potential overfitting and should be validated for stability across time.
* **loan\_int\_rate\_binned (IV = 0.7707)**: Also highly predictive and closely associated with default risk.
* **person\_income\_binned (IV = 0.5544)**: Strong predictor, reflecting the borrower’s repayment capacity.
* **loan\_amnt\_binned (IV = 0.0919)**: Weak to medium predictive power, suggesting a limited but noticeable influence on default risk.
* **person\_emp\_length\_binned (IV = 0.0596)**: Weak predictor; adds marginal value.
* **person\_age\_binned (IV = 0.0093)** and **cb\_person\_cred\_hist\_length\_binned (IV = 0.0042)**: Very weak predictors, likely to have minimal impact unless supported by domain logic or interaction effects.

|  |  |  |
| --- | --- | --- |
| Feature | IV | Interpretation |
| Loan Percent Income | 0.9477 | Extremely strong predictor. May indicate overfitting. |
| Interest Rate | 0.7707 | Very strong predictor. Closely linked to default risk. |
| Income | 0.5544 | Strong predictor. Correlates with repayment ability. |
| Loan Amount | 0.0919 | Weak to medium. Limited standalone predictive power. |
| Employment Length | 0.0596 | Weak. Minor contribution. |
| Age | 0.0093 | Very weak. Likely not useful individually. |
| Credit history Length | 0.0042 | Very weak. Not informative for default prediction. |

Table 25:Summary of Information Value (IV) Results and Interpretation for all binned features

It is important to note that IV measures **univariate** predictive strength. Variables with low IV may still contribute meaningfully when considered in multivariate models, such as logistic regression or tree-based approaches.

**8.2 Feature Importance Analysis**

To assess the contribution of individual features toward predicting loan default probability, multiple models and feature importance techniques were applied. This analysis is crucial for understanding which variables drive the model predictions, ensuring both interpretability and compliance with credit risk modelling standards.

**8.2.1 Models and Methods Utilisation and Techniques Comparison**

Feature importance was derived using four primary models, each offering unique interpretability advantages and methodological rigor:

1. **Logistic Regression**
   * **Why Selected:** Basel-compliant, highly interpretable, widely used for credit scoring.
   * **Techniques Applied:**
     + **Coefficient-Based Importance:** Regression coefficients provide a direct measure of the impact of each feature on the log-odds of default.
     + **Permutation Importance:** Evaluates feature importance by measuring the increase in prediction error after shuffling the feature values.
2. **Random Forest**
   * **Why Selected:** Robust ensemble approach, capable of capturing non-linear interactions while maintaining interpretability through tree-based measures.
   * **Techniques Applied:**
     + **Gini Importance (Mean Decrease in Impurity):** Reflects how much each feature contributes to reducing impurity across all trees.
     + **Permutation Importance:** Provides model-agnostic interpretability by quantifying performance degradation upon feature shuffling.
     + **SHAP (SHapley Additive exPlanations):** Delivers consistent and locally accurate explanations of feature contributions for each prediction.
3. **XGBoost (Gradient Boosted Trees)**
   * **Why Selected:** High predictive performance, widely used in credit risk modelling, and SHAP-compatible.
   * **Techniques Applied:**
     + **Gain-Based Importance:** Measures the contribution of each feature to the model’s overall gain during boosting iterations.
     + **Permutation Importance:** Captures the model’s dependency on individual features by performance drop after permutation.
     + **SHAP Importance:** Provides additive feature attribution values, ensuring a consistent and interpretable breakdown of prediction contributions.
4. **Decision Tree (CART)**
   * **Why Selected:** Serves as an interpretable baseline for feature importance assessment.
   * **Techniques Applied:**
     + **Gini Importance:** Indicates the average impurity reduction for nodes split on each feature.
     + **Permutation Importance:** Adds a model-agnostic perspective to tree-based interpretability.
     + **SHAP Importance:** Offers detailed feature attribution at the individual prediction level.

**Feature Representation Across Models**

* **Logistic Regression:** Weight of Evidence (WOE) transformed features were used instead of raw variables. This aligns with best practices in credit risk modelling to enhance linearity and interpretability.
* **For Tree-Based Models (Random Forest, XGBoost, Decision Tree):** Original features were utilized without WOE transformation, allowing the models to capture non-linear relationships and feature interactions.
* **Categorical Feature Handling:** All categorical variables were encoded appropriately for compatibility with both linear and tree-based algorithms.
* **Target Variable:** The target variable for all models was **loan\_status**, a binary indicator reflecting the default status of the loan (1 = Default, 0 = Non-default).

**Comparison of Techniques**

Each model’s feature importance techniques were compared to highlight consistency, interpretability, and regulatory compliance. While coefficient-based importance in Logistic Regression provides direct interpretability, advanced methods such as SHAP deliver granular and consistent explanations across all tree-based models. Permutation importance serves as a common ground for comparison, as it is model-agnostic and reflects true dependency on features.

**8.2.2 Feature Importance Analysis: Logistic Regression**

Logistic regression was implemented as the primary baseline model for Probability of Default (PD) estimation due to its regulatory compliance, interpretability, and strong acceptance in credit risk modelling frameworks. Feature importance was assessed using two complementary techniques:

* **Model Coefficients** – representing the effect size of each feature on the log-odds of default.
* **Permutation Importance** – providing a model-agnostic measure of each feature’s contribution to predictive performance.

WOE-transformed variables were used for all continuous and categorical predictors in logistic regression, ensuring monotonicity and improved interpretability.

**1. Coefficient-Based Importance**

The logistic regression coefficients indicate the direction and magnitude of each feature’s relationship with default risk. Positive coefficients increase the log-odds of default, whereas negative coefficients reduce it.

**Key Findings:**

* **Highest Positive Impact (Increased Default Risk)**
  + *loan\_grade\_G* (+2.878), followed by other lower credit grades (F, E, D), strongly increases default risk.
  + *Homeownership status – RENT* (+0.839) and *OTHER* (+0.692) are associated with higher risk relative to the baseline category.
  + Certain loan purposes such as *HOMEIMPROVEMENT*, *DEBTCONSOLIDATION*, and *MEDICAL* show higher default tendencies.
* **Highest Negative Impact (Reduced Default Risk)**
  + *loan\_grade\_A* (−2.038), *loan\_grade\_B* (−1.852), and *loan\_grade\_C* (−1.524) substantially lower default probability.
  + Features related to financial stability, such as *loan\_percent\_income\_binned\_woe* (−1.018) and *person\_income\_binned\_woe* (−0.958), are key indicators of lower risk.
  + Surprisingly, *loan\_intent\_VENTURE* and *EDUCATION* reduce default risk, suggesting specific customer profiles associated with lower-risk behaviour.
* **Minimal Effect Predictors**

*cb\_person\_default\_on\_file* (−0.021) and *person\_home\_ownership\_MORTGAGE* (+0.013) contribute negligibly, likely overshadowed by stronger related variables such as credit grade.

**2. Permutation Importance**

Permutation importance was computed on the test set using **10 random shuffles per feature**, measuring the mean drop in predictive accuracy when each feature was permuted.

**Key Findings:**

* **Top Predictors by Performance Impact:**
  + *loan\_percent\_income\_binned\_woe* (0.0737), *loan\_grade\_A* (0.0616), and *person\_income\_binned\_woe* (0.0428) are the most influential variables for prediction accuracy.
  + *loan\_grade\_B* and *loan\_grade\_C* also rank highly, confirming their predictive relevance.
* **Notable Discrepancies from Coefficient Rankings**
  + *loan\_grade\_G* (ranked 1st by coefficients) appears 15th by permutation, indicating strong statistical association but limited incremental predictive power once other variables are accounted for.
  + Similar behaviour is observed for *loan\_grade\_F* and *loan\_grade\_E*.
* **Negligible or Zero Impact Features**

Features such as *loan\_amnt\_binned\_woe*, *cb\_person\_default\_on\_file*, and some homeownership indicators do not significantly affect model performance when removed.

**3. Comparative Insights: Coefficient vs Permutation**

The summary of the most influential predictors is presented in Table below. The full ranking of all 25 features, based on logistic regression coefficients and permutation importance results, is provided in the Appendix section**.**

|  |  |  |  |
| --- | --- | --- | --- |
| Feature | Coefficient Rank | Permutation Rank | Comment |
| loan\_percent\_income \_woe | 5 | 1 | Strong in both |
| loan\_grade\_A | 2 | 2 | Consistent and reliable |
| person\_income \_woe | 6 | 3 | Consistent predictor |
| loan\_grade\_B | 4 | 4 | Strong across both methods |
| loan\_grade\_G | 1 | 15 | Overstated by coefficient importance |
| cb\_person\_default\_on\_file | 24 | 22 | Weak predictor overall |

Table 26: The summary of the most influential predictors

**Interpretation:**

* **Reliable Predictors:** *loan\_percent\_income\_woe*, *loan\_grade\_A*, *loan\_grade\_B*, and *person\_income \_woe* emerge as consistently important across both techniques.
* **Potential Overfitting Indicators:** *loan\_grade\_G* and *loan\_grade\_F* exhibit large coefficients but low permutation scores, suggesting limited incremental value and possible collinearity.
* **Regulatory Perspective:** Coefficient-based rankings are essential for explainability and compliance, while permutation importance offers practical insights into the model’s true reliance on each feature for predictive accuracy.

**4. Practical Implications**

* Credit grade and income-related features are the most influential drivers of PD, aligning with domain expectations.
* Homeownership and loan purpose provide additional explanatory power but should be interpreted cautiously.
* Variables with negligible contribution (e.g., *cb\_person\_default\_on\_file*) may be candidates for removal in simplified scorecards, provided regulatory guidelines permit.

**8.2.3 Feature Importance – Random Forest**

Random Forest, as an ensemble of decision trees, provides multiple ways to evaluate feature importance. For this analysis, three complementary techniques were employed:

* **Gini Importance (Mean Decrease in Impurity)** – Reflects how frequently and effectively a feature is used for splitting across all trees.
* **Permutation Importance** – Measures the reduction in predictive performance when the feature values are randomly permuted.
* **SHAP (SHapley Additive exPlanations)** – Provides a game-theoretic approach to measure both global and local contributions of features to predictions.

Categorical variables were one-hot encoded prior to modelling, and unlike logistic regression, the raw features (not binned WoE transformations) were used. The target variable is **loan\_status**, representing whether the borrower defaulted.

**1. Gini Importance**

Summary of the Gini-based ranking of the most influential features is shown below, while the complete table is provided in Appendix.

**Key Observations:**

* **Top Predictors:**
  + **loan\_percent\_income (24.1%)** – The loan-to-income ratio is the most critical feature, strongly influencing default predictions.
  + **person\_income (13.4%)** – Higher personal income correlates with lower default risk.
  + **loan\_int\_rate (12.6%)** – Higher interest rates often indicate higher perceived risk by the lender.
* **Moderate Contributors:**
  + **person\_home\_ownership\_RENT**, **loan\_grade\_D**, and **loan\_amnt** reflect socio-economic status and loan characteristics.
* **Low Importance Features:**
  + **loan\_grade\_F**, **loan\_grade\_G**, and **person\_home\_ownership\_OTHER** show negligible influence, possibly due to sparse data or weak correlation.
* **Notable Insight:**
  + **cb\_person\_default\_on\_file**, though an indicator of prior defaults, ranks relatively low (17th), likely because its effect is partially captured by credit grade and interest rate.

**2. Permutation Importance**

Permutation importance offers a more robust view of predictive impact, as shown below (full ranking is in Appendix).

**Interpretation:**

* **Top 3 Features:**
  + **loan\_percent\_income**, **person\_income**, and **loan\_int\_rate** remain dominant, consistent with the Gini method.
* **Loan Grade Effects:**
  + **loan\_grade\_A, D, C, and B** rank higher here than in Gini, reflecting their true predictive contribution.
* **Behaviour of Boolean Variables:**
  + **cb\_person\_default\_on\_file** gains importance (rank 12), indicating that when its values are shuffled, model performance drops significantly.
* **Low Impact Features:**
  + **loan\_grade\_G**, **loan\_intent\_PERSONAL**, and **person\_home\_ownership\_OTHER** continue to show negligible impact.

This consistency between the two methods reinforces confidence in the stability of key predictors.

**3. SHAP Importance**

SHAP values provide both **global** and **local** interpretability. Top features are summarised below (full results in Appendix).

**Insights:**

* **Top Features:**
  + **loan\_percent\_income**, **loan\_int\_rate**, and **person\_income** remain the most influential, confirming their dominance.
* **Role of Home Ownership:**
  + **person\_home\_ownership\_RENT** and **MORTGAGE** emerge as important factors, possibly reflecting financial stability.
* **Loan Grade Patterns:**
  + Order of importance follows **D → A → C → B → E → F → G**, aligning with expected credit risk stratification.
* **SHAP vs Other Methods:**
  + Features like **cb\_person\_default\_on\_file** and **loan\_intent\_MEDICAL** receive more weight compared to Gini, highlighting SHAP’s ability to capture non-linear effects and interactions.

**4. Comparative Analysis – Gini vs Permutation vs SHAP**

The three approaches broadly agree on the most influential predictors, as summarised in Table below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature | Gini Rank | Permutation Rank | SHAP Rank | Notes |
| loan\_percent\_income | 1 | 1 | 1 | Consistently the strongest predictor |
| person\_income | 2 | 2 | 3 | Very stable across methods |
| loan\_int\_rate | 3 | 3 | 2 | Slightly higher in SHAP |
| cb\_person\_default\_on\_file | 17 | 12 | 16 | Undervalued by Gini, highlighted by SHAP |
| loan\_grade\_F/G | 23/24 | 21/23 | 23/24 | Consistently least important |

Table 27:Random Forest’s Feature Importance Comparative Analysis – Gini vs Permutation vs SHAP

**5. Summary and Recommendations**

* **Agreement Across Methods:** All approaches agree on **loan\_percent\_income**, **person\_income**, and **loan\_int\_rate** as the top three predictors, confirming model stability.
* **Method-Specific Strengths:**
  + **Gini Importance:** Fast and indicative of internal model usage, but biased toward continuous variables.
  + **Permutation Importance:** Reflects real-world predictive impact; essential for model validation.
  + **SHAP:** Provides rich interpretability and local explanations, making it highly valuable for stakeholder communication and compliance in credit risk modelling and regulatory justifications.

**8.2.4 Feature Importance – XGBoost**

XGBoost is a gradient boosting algorithm widely recognised for its predictive power in credit risk modelling. For this analysis, three complementary feature importance techniques were applied: Gain Importance, Permutation Importance, and SHAP (SHapley Additive exPlanations.

**1. Gain-Based Feature Importance**

Gain Importance measures the relative improvement in model performance when a feature is used for splitting. The top features ranked by Gain are as following (with the full ranking available in Appendix):

**Key Observations:**

* **Dominant Feature:**
  + *person\_home\_ownership\_RENT* (23.1%) is the most influential predictor, indicating that borrowers who rent their homes have a higher probability of default compared to those who own or have a mortgage.
* **Loan Grades and Risk Stratification:**
  + *loan\_grade\_C* and *loan\_grade\_D* are highly important, suggesting these mid-tier grades capture substantial default risk variation.
* **Affordability Indicators:**
  + *loan\_percent\_income* (loan-to-income ratio) ranks third, consistent with its critical role in assessing repayment capacity.
* **Loan Intent:**
  + *MEDICAL*, *DEBTCONSOLIDATION*, and *VENTURE* purposes rank high, reflecting their strong influence on default likelihood.
* **Moderate Role of Traditional Metrics:**
  + Features such as *person\_income*, *person\_emp\_length*, and *cb\_person\_cred\_hist\_length* contribute less compared to behavioural and loan-specific attributes.
* **Low Impact Features:**
  + *loan\_grade\_A, B, F, G* and *cb\_person\_default\_on\_file* have minimal impact, likely due to either limited variability or overlap with stronger predictors.

**2. Permutation Importance**

Permutation importance provides a more reliable indication of feature reliance by measuring prediction deterioration after feature shuffling. Summary of the key results is as following (complete rankings in Appendix section).

**Insights:**

* **Top Predictors:**
  + *person\_income*, *loan\_percent\_income*, and *loan\_int\_rate* are the most influential, aligning closely with domain expectations and demonstrating model dependence on affordability factors.
* **Home Ownership Influence:**
  + *person\_home\_ownership\_OWN* and *RENT* remain relevant but rank lower than in Gain-based importance, indicating structural bias in tree splits.
* **Loan Intent and Grades:**
  + *VENTURE*, *DEBTCONSOLIDATION*, and *MEDICAL* loan intents show moderate influence, while most loan grades (except *D* and *C*) exhibit very low importance.
* **Minimal and Negative Contributions:**
  + Features such as *loan\_grade\_F, G* and *loan\_intent\_PERSONAL* contribute little or even negatively, implying redundancy or noise.

**3. SHAP-Based Feature Importance**

SHAP values quantify both global and individual-level impact, providing the most interpretable measure for regulatory and business stakeholders. Summary of the top features by average absolute SHAP value is as following (full details in Appendix).

**Highlights:**

* **Top Three Predictors:**
  + *loan\_int\_rate*, *person\_income*, and *loan\_percent\_income* dominate, confirming their critical role in predicting default.
* **Behavioural and Housing Indicators:**
  + *person\_home\_ownership\_RENT* and *OWN* remain prominent, highlighting housing status as a significant socio-economic signal.
* **Loan Intent Effects:**
  + *VENTURE*, *HOMEIMPROVEMENT*, and *DEBTCONSOLIDATION* have noticeable influence, whereas *EDUCATION* and *PERSONAL* rank lower.
* **Credit Bureau Features:**
  + *cb\_person\_default\_on\_file* and *cb\_person\_cred\_hist\_length* contribute minimally, suggesting limited incremental predictive value.

**4. Comparative Analysis of All Three Methods**

A comparison across Gain, Permutation, and SHAP rankings (summarised in Table below, full version in Appendix) shows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature | Gain Rank | Permutation Rank | SHAP Rank | Consensus Observation |
| Interest rate | 8 | 3 | 1 | SHAP places highest importance on interest rate. |
| Income | 11 | 1 | 2 | Consistently a top predictor in real-world impact. |
| Loan percent income | 3 | 2 | 3 | Stable across all techniques. |
| Person home ownership RENT | 1 | 5 | 4 | Extremely important structurally and behaviourally. |
| Person home ownership OWN | 7 | 4 | 5 | Strong influence across methods. |
| Loan intent VENTURE | 9 | 6 | 6 | Appears consistently in top 10. |
| Person default on file | 25 | 19 | 21 | Minimal importance overall. |

Table 28: A comparison across Gain, Permutation, and SHAP rankings

**5. Summary and Recommendations**

* **Core Predictors:** *loan\_int\_rate*, *person\_income*, and *loan\_percent\_income* remain the most reliable indicators across all methods.
* **Behavioural Signals Matter:** Housing status and loan purpose add significant predictive power.
* **Credit Bureau Indicators:** Play a minor role compared to affordability and behavioural variables.
* **In practice we:**
  + Use **SHAP values** for compliance and detailed interpretability.
  + Employ **Permutation Importance** for validating feature dependence.
  + Use **Gain Importance** for quick structural insights but avoid relying solely on it.

**8.2.5 Feature Importance – Decision Tree**

To complement the ensemble methods and logistic regression, a **Decision Tree (CART)** model was implemented to provide a transparent and interpretable baseline for feature importance assessment. This model is well-suited for visual explanations and serves as a useful benchmark for more complex models like Random Forest and XGBoost.

For the Decision Tree, three complementary approaches were employed to evaluate feature importance are Gini Importance, Permutation Importance, and SHAP Values.

**1. Gain-Based Feature Importance**

The top contributors were (Complete results are in Appendix):.

* **loan\_percent\_income (29.17%)**,
* **person\_home\_ownership\_RENT (18.85%)**, and
* **loan\_int\_rate (18.22%)**.

These three features alone account for over **66% of the model’s splitting power**, indicating that income-to-loan ratio, housing status, and interest rate are the primary decision drivers. Features such as **loan\_grade\_A/B/F** and certain loan intents (e.g., **EDUCATION, PERSONAL**) had negligible influence, suggesting they were rarely selected for splitting.

**2. Permutation-Based Feature Importance**

The ranking differs slightly from Gini, with:

* **person\_home\_ownership\_RENT (0.1327)** emerging as the top predictor,
* followed by **loan\_percent\_income (0.1275)** and
* **loan\_int\_rate (0.1160)**.

This indicates that shuffling these features leads to the greatest performance deterioration, confirming their critical role in prediction. Additional influential features include **person\_income**, **loan\_grade\_C/D**, and loan purposes such as **VENTURE** and **HOMEIMPROVEMENT**.

Full permutation importance scores are reported in Appendix.

**3. SHAP-Based Feature Importance**

SHAP values offer a granular perspective by quantifying each feature’s contribution to individual predictions. According to SHAP, the most impactful features were (Complete results are in Appendix):

* **loan\_percent\_income (0.1051)**,
* **loan\_int\_rate (0.0790)**, and
* **person\_home\_ownership\_RENT (0.0596)**.

These results align closely with permutation findings, while SHAP further highlights moderate contributions from **loan\_grade\_D/C**, **person\_emp\_length**, and specific loan intents (e.g., **MEDICAL, DEBTCONSOLIDATION**).

**4. Comparative Analysis of All Methods**

A comparison across Gini, Permutation, and SHAP rankings shows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature | Gini Rank | Permutation Rank | SHAP Rank | Key Insight |
| loan\_percent\_income | 1 | 2 | 1 | Strong consensus across all three methods |
| loan\_int\_rate | 3 | 3 | 2 | Highly important in all methods |
| Person home ownership RENT | 2 | 1 | 3 | High predictive power despite being categorical |
| person\_income | 4 | 4 | 4 | Consistent predictor of creditworthiness |
| person\_emp\_length | 5 | 10 | 7 | More influence in SHAP than in permutation |
| loan\_grade\_C | 6 | 5 | 6 | Medium importance, stable across methods |
| loan\_grade\_D | 7 | 6 | 5 | Slightly higher in SHAP compared to others |
| Loan intent MEDICAL | 8 | 13 | 8 | More pronounced in SHAP and tree-based methods |
| Person default on file | 18 | 14 | 12 | Greater influence in SHAP than in Gini or permutation |

Table 29: A comparison across Gini, Permutation, and SHAP rankings

**5. Final Insights**

* The **top four features**—**loan\_percent\_income**, **loan\_int\_rate**, **person\_home\_ownership\_RENT**, and **person\_income**—are critical determinants of default risk.
* SHAP provides nuanced interpretability, revealing feature contributions at the individual level.
* Permutation importance best reflects predictive reliance by simulating feature absence.
* Gini importance is useful for understanding tree structure but may introduce bias towards variables with more split points.

All detailed results and complete rankings are provided in Appendix.

**8.2.6 Overall Conclusion on Feature Importance Across All Models**

The comparative analysis of feature importance across four models—**Logistic Regression, Decision Tree, Random Forest, and XGBoost**—using multiple interpretability techniques reveals consistent and insightful patterns:

**1. Core Predictors Consistently Identified**

* Across all models and importance measures, **loan\_percent\_income** and **loan\_int\_rate** emerged as the **most influential predictors** of loan default probability.
* **person\_income** and **employment length** further demonstrated significant predictive power, confirming their role in assessing repayment capacity.

**2. Impact of Categorical Variables**

* **Home ownership status (RENT)** consistently ranked among the top predictors in tree-based models and SHAP analysis, illustrating its strong association with credit risk.
* Certain **loan intent categories** (e.g., MEDICAL, DEBT CONSOLIDATION) and **loan grades (C and D)** provided moderate but non-negligible influence, particularly in non-linear models like XGBoost.

**3. Agreement vs. Divergence Across Methods**

* **SHAP values** offered the most granular interpretability by capturing feature interactions and individual-level contributions.
* **Permutation importance** highlighted predictive influence on overall model accuracy, making it valuable for performance-oriented feature evaluation.
* **Gini importance** provided quick insights into structural splits but displayed bias towards features with higher cardinality or variance.
* **Logistic regression coefficients** remained highly interpretable and regulatory-compliant, but they required WOE binning to handle categorical variables effectively.

**4. Features with Minimal Predictive Value**

* Features such as **loan\_grade\_A/B/F**, **loan intents like EDUCATION and PERSONAL**, and **credit history length** consistently showed negligible impact, suggesting opportunities for dimensionality reduction.

**5. Regulatory and Business Implications**

* For **Basel-compliant credit risk models**, logistic regression with WOE features remains the preferred approach due to transparency and monotonic relationships.
* For **high-accuracy predictive systems**, ensemble methods (Random Forest, XGBoost) combined with SHAP interpretation deliver superior performance without compromising explainability.

**8.2.7 Next Steps**

Following the comprehensive feature importance analysis, the next phase involves finalising the data preprocessing pipeline. This includes incorporating the results of feature engineering techniques such as Weight of Evidence (WOE) transformation and optimal binning, ensuring consistent treatment of categorical variables and monotonicity in logistic models. These enhancements will prepare the dataset for the final model implementation and evaluation stage, where we will build, validate, and benchmark the Probability of Default (PD) models against regulatory and performance standards.

**9.** **Data Preprocessing (Part 2 – Feature Transformation & Balancing)**

With the dataset now cleaned and validated, the next stage was to transform features into a form suitable for modelling. This step completes the preprocessing pipeline, ensuring that all variables—both original and engineered—are properly represented for logistic regression and other classification algorithms. For future predictions, any new data must pass through this exact preprocessing pipeline to guarantee consistency between training and scoring environments.

**9.1 Incorporation of Engineered Features**

As outlined in Section 8.1 (Binning and Weight of Evidence Transformation), numerical variables were first discretised using supervised binning, followed by Weight of Evidence (WoE) encoding. This ensured that predictors exhibit monotonic relationships with the target variable, stabilise outliers, and improve interpretability under regulatory frameworks (Siddiqi, 2006; BCBS, 2004, 2006).  
The WoE-transformed fields—such as *loan\_amnt\_binned\_woe*, *person\_income\_binned\_woe*, and *loan\_percent\_income\_binned\_woe*—were then incorporated into the dataset alongside the original and binned features. While clustering methods were evaluated as a potential form of feature engineering, they were not incorporated at this stage, as classification results did not demonstrate incremental performance benefits. However, clustering may be reconsidered in future iterations should data characteristics change.

**9.2 Encoding of Categorical Variables**

Categorical features, including **home ownership**, **loan intent**, and **loan grade**, were transformed into machine-readable form through one-hot encoding. This step ensures compatibility with statistical and machine learning algorithms, which typically require numerical input. Each category was expanded into a binary indicator column (e.g., *person\_home\_ownership\_MORTGAGE*, *loan\_intent\_EDUCATION*, *loan\_grade\_B*), allowing the model to capture differences in credit risk across groups without imposing ordinal assumptions.

**9.3 Scaling and Balancing Considerations**

In principle, **feature scaling** and **class balancing** could also be integrated at this stage of preprocessing, ensuring that the dataset is standardised and balanced before model training.

* **Scaling:** For this project, scaling was not applied to logistic regression since all numerical variables were transformed into WoE features, which are already standardised and monotonic by design. For other algorithms (Random Forest, Decision Tree, and XGBoost), a dedicated scaling class was applied during model execution. This implementation used z-score standardisation, with additional skewness handling.
* **Balancing:** To address class imbalance, **SMOTE (Synthetic Minority Oversampling Technique)** was employed for all algorithms. Applying SMOTE at the modelling stage provided control over different resampling scenarios, making it possible to observe its impact on each algorithm separately.

Although these steps could be consolidated into preprocessing for efficiency and pipeline simplicity, implementing them during model execution provided greater flexibility and insight into the interaction between scaling, balancing, and algorithm performance.

**9.4 Final Feature Set**

Following the integration of WoE transformations and one-hot encoding, the final dataset comprised **43 fields**, encompassing original, binned, WoE-transformed, and encoded features. This balanced representation allowed the model to leverage predictive information across variable types while maintaining interpretability and regulatory alignment.

**9.5 Summary & Next Steps**

In this section, newly engineered features were integrated into the dataset, WoE transformations were applied to maintain monotonic predictor relationships, and categorical variables were one-hot encoded to ensure algorithm compatibility. Scaling and balancing were considered as part of preprocessing design but were implemented during model execution, offering more control and flexibility in testing their impact across different algorithms.

|  |  |  |  |
| --- | --- | --- | --- |
| Transformation Step | Methodology | Stage of Implementation | Notes |
| Binning | Supervised binning using *OptimalBinning* (Monotonic Optimal Binning, MOB) | Preprocessing | Produces categories aligned with target; prepares features for WoE. |
| Weight of Evidence (WoE) | Monotonic transformation of binned variables into log-odds scale | Preprocessing | Applied to key numerical variables for logistic regression. |
| One-Hot Encoding | Conversion of categorical variables (home ownership, loan intent, loan grade) into binary indicators | Preprocessing | Ensures algorithm compatibility by transforming non-numeric fields. |
| Scaling | Z-score standardisation with skew handling (handle\_skew=True, skew\_method='log', skew\_threshold=1.0) | Model execution | Not applied for Logistic Regression (WoE already standardised). Used for Random Forest, Decision Tree, and XGBoost. |
| Class Balancing | SMOTE (Synthetic Minority Oversampling Technique) | Model execution | Applied across all algorithms to address class imbalance. |

Table 30: Summary of Feature Transformation, Scaling, and Balancing Steps

The result was a robust feature set of 43 variables, providing both predictive power and interpretability. With preprocessing now complete, the dataset is ready for model development. For operational deployment, future incoming data will be processed through this identical pipeline to preserve consistency and reproducibility across training, validation, and prediction phases.

**10. Model Building, Evaluation, and Validation**

This section details the predictive modelling phase of the project. The primary objective is to develop a robust model that can accurately predict the probability of default.

**10.1 Modelling Approach**

The preprocessed dataset was partitioned into three distinct sets to ensure a rigorous and unbiased evaluation:

* **Training Set (70%):** Used to train the models.
* **Test Set (20%):** Used for initial evaluation and hyperparameter tuning.
* **Evaluation Set (10%):** A final, unseen hold-out set used for assessing the real-world performance of the selected model.

A suite of four diverse algorithms was selected to model the probability of default, ranging from an interpretable baseline to state-of-the-art ensemble methods. This multi-model approach allows for a comprehensive comparison of different predictive techniques.

* **Logistic Regression:** This was chosen as a **baseline model** due to its high interpretability and efficiency. It works by modelling the probability of a binary outcome (default vs. non-default) by fitting a logistic function to the data. The resulting coefficients provide clear insights into the linear relationship between each feature and the likelihood of default (Hosmer, Lemeshow and Sturdivant, 2013).
* **Decision Tree:** Selected for its intuitive, rule-based nature, which makes it easy to understand and explain to non-technical stakeholders. It functions by splitting the data into progressively smaller subsets based on the most significant features, creating a flowchart-like tree structure that leads to a final classification (Breiman et al., 1984).
* **Random Forest:** This ensemble method was chosen to improve upon the potential overfitting of a single decision tree and to capture more complex non-linear interactions. It operates by constructing a large number of individual decision trees on random subsets of the data and features. The final prediction is determined by a majority vote from all the trees, leading to a more robust and accurate outcome (Breiman, 2001).
* **XGBoost (Extreme Gradient Boosting):** Included as a **state-of-the-art** algorithm known for its exceptional predictive accuracy. It is a gradient boosting method that builds trees sequentially, where each new tree is trained to correct the errors made by the previous ones. This sequential boosting process makes it a highly effective and often top-performing model in classification tasks (Chen and Guestrin, 2016).

To optimise performance, each of these algorithms underwent extensive hyperparameter tuning using a **Grid Search** methodology on the training set. This process systematically explores a range of parameter combinations to find the most effective configuration for each model.

Each model was trained and subsequently evaluated on both the test and evaluation datasets. The following sections present the quantitative results and visual analysis for each algorithm.

**10.2.1 Logistic Regression Model’s Performance Analysis**

As a linear and highly interpretable model, Logistic Regression serves as a strong baseline for performance. The model's performance is summarised below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dataset | AUC-ROC | Accuracy | Precision (Default) | Recall (Default) | F1-Score (Default) |
| Test Set | 0.888 | 0.830 | 0.58 | 0.79 | 0.67 |
| Evaluation Set | 0.890 | 0.841 | 0.60 | 0.79 | 0.69 |

Table 31: Logistic Regression Model’s Performance Summary

**Model Discrimination (ROC-AUC)**

The model demonstrates strong discriminatory power in both the test and evaluation datasets, with ROC-AUC scores of **0.888** and **0.890**, respectively. This indicates that the logistic regression model can effectively distinguish between successful and defaulted cases of loans, achieving near-consistent performance across datasets.

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

Figure 22: Logistic Regression Model's Test & Evaluation ROC-AUC Curves

**Classification Performance**

The results reveal an important balance between the majority class (non-default, class 0) and minority class (default, class 1):

* **Non-defaults (Class 0):** Precision and recall are consistently high (Test: Precision = **0.93**, Recall = **0.84**; Eval: Precision = **0.94**, Recall = **0.85**). This shows the model is highly reliable in correctly identifying successful loans while keeping false positives relatively low.
* **Defaults (Class 1):** Precision is moderate (Test: **0.58**, Eval: **0.60**) but recall is high (Test: **0.79**, Eval: **0.79**). This means that while the model sometimes misclassifies non-defaults as defaults (false positives), it captures the majority of true defaults. The higher recall for defaults is particularly desirable in credit risk modelling, as missing defaulters (false negatives) is costlier for the lender than misclassifying some non-defaulters.
* **F1-Score:** For defaults, the F1-scores are **0.67** (Test) and **0.69** (Eval), reflecting a reasonable balance between precision and recall. This balance suggests the model achieves an effective trade-off between capturing risky borrowers and limiting unnecessary rejections.

**Confusion Matrix** **Insights**

* On the **test set**, the model correctly identified **1,021 defaults** out of 1,293 but misclassified 272 defaults as non-defaults.
* On the **evaluation set**, the model correctly identified **513 defaults** out of 646 but misclassified 133 defaults.
* False positives (non-defaults predicted as defaults) are higher (Test: 730, Eval: 335), which is consistent with the moderate precision for the default class.

This pattern shows that the model errs more on the side of caution—flagging some good borrowers as risky—rather than missing a large number of risky ones. From a risk management perspective, this is often an acceptable and even desirable trade-off.

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

Figure 23: Logistic Regression Model’s Test & Evaluation Confusion Matrices

**Overall Accuracy and Balanced Accuracy**

* Accuracy: **83.0%** (Test) and **84.1%** (Eval), reflecting solid overall predictive performance.
* Balanced Accuracy: **0.816 (Test)** and **0.824 (Eval)**, showing the model maintains fairness across imbalanced classes, and does not overly bias toward the majority (non-default) class.

**Calibration and Probabilistic Metrics**

* **Log Loss** (Test: **0.402**, Eval: **0.396**) and **Brier Score** (Test: **0.126**, Eval: **0.123**) are relatively low, indicating the model’s probability estimates are well-calibrated and reliable for downstream applications such as credit scoring or risk ranking.
* **Average Precision (0.763)** further confirms good performance in ranking defaults higher than non-defaults across probability thresholds.

**Precision-Recall Curve**

The Logistic Regression model shows a **clear trade-off between precision and recall**.

* For **defaults (Class 1)**: recall is relatively strong (~0.79 in both Test and Evaluation), meaning the model captures most risky borrowers. However, precision is moderate (~0.58–0.60), indicating that many predicted defaults are actually safe borrowers.
* For **non-defaults (Class 0)**: precision is very high (~0.93–0.94), but recall is lower (~0.84–0.85), which means some safe borrowers are misclassified as risky.

Overall, the precision-recall curve reflects the model’s tendency to over-flag defaults (higher recall, lower precision), making it useful when the business priority is to **reduce missed defaults**, even if it means rejecting some safe borrowers.

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

Figure 24: Logistic Regression Model's Test & Evaluation Precision-Recall Curves

**Predicted Probability Distribution**

The predicted probability distributions for Logistic Regression show **significant overlap between default and non-default classes**.

* Many borrowers are assigned mid-range probabilities, leading to misclassifications.
* While some separation exists (defaults generally receive higher PDs than non-defaults), the overlap explains why precision for defaults remains modest (~0.58–0.60).

This highlights the model’s limitation in making **confident probability estimates**.

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

Figure 25: Logistic Regression Model's Test & Evaluation Predicted Probability Distributions

**Business Implications**

The logistic regression model offers an interpretable and well-calibrated foundation for predicting Probability of Default (PD). Its **strength lies in achieving high recall for defaults**, ensuring most risky borrowers are detected. While precision for defaults is moderate, this trade-off is acceptable in credit risk settings where the cost of approving a defaulter outweighs the cost of rejecting a good borrower.

Given its consistent performance across test and evaluation datasets, the model is stable and generalises well. Logistic regression also provides explainability through coefficients and odds ratios, making it particularly suitable in **regulatory and compliance-driven environments** where transparency is essential.

**10.3 Decision Tree Model’s Performance Analysis**

As a non-linear and interpretable model, the Decision Tree provides a flexible approach to classification and is often a strong improvement over linear model. The model’s performance is summarised below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dataset | AUC-ROC | Accuracy | Precision (Default) | Recall (Default) | F1-Score (Default) |
| Test Set | 0.912 | 0.920 | 0.91 | 0.71 | 0.80 |
| Evaluation Set | 0.910 | 0.918 | 0.89 | 0.71 | 0.79 |

Table 32: Decision Tree Model’s Performance Summary

**Model Discrimination (ROC-AUC)**

The Decision Tree demonstrates excellent discriminatory power, with ROC-AUC scores of 0.912 (Test) and 0.910 (Evaluation). These results confirm the model’s strong ability to distinguish between successful and defaulted loans, while also maintaining consistency across datasets.

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

Figure 26: Decision Tree Model's Test & Evaluation ROC-AUC Curves

**Classification Performance**

The classification results reveal a different balance between the majority class (non-default, class 0) and the minority class (default, class 1):

* **Non-defaults (Class 0):** Precision and recall are exceptionally high (Test: Precision = 0.92, Recall = 0.98; Eval: Precision = 0.92, Recall = 0.98). This shows the model is highly effective in correctly identifying successful loans, with very few false positives.
* **Defaults (Class 1):** Precision is strong (Test: 0.91, Eval: 0.89), but recall is moderate (Test: 0.71, Eval: 0.71). This means the model captures the majority of true defaults but still misses about 29% of defaulters. The F1-scores (Test: 0.80, Eval: 0.79) reflect a reasonable balance between precision and recall.

This pattern suggests that the Decision Tree is more conservative: it avoids flagging too many safe borrowers as risky, but at the cost of missing some actual defaulters.

**Confusion Matrix Insights**

* **Test Set:**
* True Negatives: 4,505
* False Positives: 95
* False Negatives: 374
* True Positives: 919
* **Evaluation Set:**
* True Negatives: 2,246
* False Positives: 55
* False Negatives: 188
* True Positives: 458

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

Figure 27: Decision Tree Model’s Test & Evaluation Confusion Matrices

The very low number of false positives (specificity ≈ 98%) indicates high reliability in identifying safe loans. However, the number of false negatives highlights the model’s challenge in capturing all risky borrowers.

**Overall Accuracy and Balanced Accuracy**

* **Accuracy:** 92.0% (Test) and 91.8% (Eval), both notably higher than Logistic Regression (≈83%).
* **Balanced Accuracy:** 0.845 (Test) and 0.843 (Eval), confirming that the model performs well across both majority and minority classes.

**Calibration and Probabilistic Metrics**

* **Log Loss:** 0.396 (Test), 0.406 (Eval) – relatively low, indicating well-calibrated predictions.
* **Brier Score:** 0.078 (Test), 0.080 (Eval) – further evidence of good probability estimation.
* **Average Precision:** 0.846 (Test), 0.838 (Eval) – consistent with ROC-AUC, reflecting strong ranking ability.

**Precision-Recall Curve**

The precision-recall curve highlights the trade-off between precision and recall. The model achieves high precision for defaults but at the expense of lower recall, which explains why some defaulters are missed.

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

Figure 28: Decision Tree Model's Test & Evaluation Precision-Recall Curves

**Predicted Probability Distribution**

The predicted probability distributions for both classes show relatively distinct separation, with limited overlap between non-default and default probabilities. This suggests the model is generally confident in its predictions.

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

Figure 29: Decision Tree Model's Test & Evaluation Predicted Probability Distributions

**Business Implications**

The Decision Tree model demonstrates significant improvements over Logistic Regression, particularly in terms of overall accuracy (92% vs 83%) and discriminatory power (AUC ≈ 0.91 vs 0.89). Its ability to achieve very low false positive rates is highly beneficial for lending decisions, as it ensures safe borrowers are not unnecessarily rejected.

However, the moderate recall for defaults (≈71%) means that some risky borrowers may still be approved. Depending on business priorities:

* If **minimising defaults** is the key goal, threshold adjustments or ensemble approaches (Random Forest, XGBoost) may help increase recall.
* If **minimising unnecessary rejections** is preferred, the Decision Tree already performs exceptionally well.

**Conclusion**

The Decision Tree model delivers excellent predictive performance, surpassing Logistic Regression in accuracy and AUC. Its stability across both test and evaluation datasets confirms generalisability. The main limitation is recall for defaults, which can be addressed through threshold tuning or more advanced ensemble methods. Overall, the Decision Tree offers a strong and interpretable foundation for Probability of Default (PD) modelling.

**10.4 Random Forest Model’s Performance Analysis**

As an ensemble of decision trees, Random Forest combines the strengths of multiple models to improve predictive performance and robustness. The model’s performance is summarised below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dataset | AUC-ROC | Accuracy | Precision (Default) | Recall (Default) | F1-Score (Default) |
| Test Set | 0.929 | 0.930 | 0.96 | 0.71 | 0.82 |
| Evaluation Set | 0.926 | 0.930 | 0.95 | 0.71 | 0.82 |

Table 33: Random Forest Model’s Performance Summary

**Model Discrimination (ROC-AUC)**

The Random Forest achieves excellent discriminatory power with ROC-AUC values of 0.929 (Test) and 0.926 (Evaluation). Both scores are above 0.90, confirming the model’s ability to effectively separate defaulters from non-defaulters, and demonstrating strong generalisation across datasets.

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

Figure 30: Random Forest Model's Test & Evaluation ROC-AUC Curves

**Classification Performance**

The class-wise metrics show a strong balance between reliability and recall:

* **Non-defaults (Class 0):** Precision and recall are extremely high (Test: Precision = 0.92, Recall = 0.99; Eval: Precision = 0.92, Recall = 0.99). This demonstrates that the model is highly dependable in identifying successful loans, with almost no false positives.
* **Defaults (Class 1):** Precision is very high (Test: 0.96, Eval: 0.95), indicating that when the model predicts a default, it is usually correct. Recall, however, is moderate (Test: 0.71, Eval: 0.71), meaning that nearly 29% of defaulters are missed. The F1-scores of 0.82 (Test) and 0.82 (Eval) highlight a solid trade-off between precision and recall.

This performance pattern shows that the Random Forest is highly conservative with false positives while missing some defaulters, which is typical in imbalanced datasets.

**Confusion Matrix Insights**

* **Test Set:**
  + True Negatives: 4,561
  + False Positives: 39
  + False Negatives: 376
  + True Positives: 917
* **Evaluation Set:**
  + True Negatives: 2,279
  + False Positives: 22
  + False Negatives: 185
  + True Positives: 461

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

Figure 31: Random Forest Model’s Test & Evaluation Confusion Matrices

The very low false positive rate (specificity ≈ 99%) highlights the model’s reliability in approving safe loans. However, the moderate number of false negatives reflects the model’s challenge in capturing all defaulters.

**Overall Accuracy and Balanced Accuracy**

* **Accuracy:** 92.96% (Test) and 92.98% (Eval), showing consistently strong overall performance.
* **Balanced Accuracy:** 0.850 (Test) and 0.852 (Eval), reflecting fair treatment of both majority and minority classes despite class imbalance.

**Calibration and Probabilistic Metrics**

* **Log Loss:** 0.219 (Test), 0.222 (Eval) – low values, indicating well-calibrated probability predictions.
* **Brier Score:** 0.060 (Test), 0.061 (Eval) – further confirming calibration strength.
* **Average Precision:** 0.880 (Test), 0.876 (Eval) – consistent with AUC results, showing excellent ranking capability.

**Precision-Recall Curve**

The Random Forest precision-recall curve shows excellent precision for defaults, but recall remains moderate. This illustrates that while most predicted defaults are correct, some actual defaults remain undetected.

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

Figure 32: Random Forest Model's Test & Evaluation Precision-Recall Curves

**Predicted Probability Distribution**

The predicted probability distributions for non-default and default classes are well separated, with minimal overlap. This indicates strong model confidence in its probability estimates.

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

Figure 33: Random Forest Model's Test & Evaluation Predicted Probability Distributions

**Business Implications**

The Random Forest significantly improves upon both Logistic Regression and Decision Tree models. With an AUC of ~0.93 and accuracy of ~93%, it sets a new benchmark in predictive performance. Its strengths include:

* **Very high precision for defaults** – minimising the risk of wrongly classifying safe borrowers as risky.
* **Low false positive rate (~1%)** – ensuring that safe borrowers are rarely rejected unnecessarily.

However, recall for defaults (≈71%) remains the main limitation. This means some risky borrowers could still be approved. Depending on business objectives:

* If **minimising defaults** is a top priority, threshold tuning, class weighting, or ensemble variants (e.g., XGBoost) may be used to improve recall.
* If **reducing false alarms** is more important, this model is already highly effective.

**Conclusion**

The Random Forest model delivers outstanding predictive performance, with superior accuracy and AUC compared to both Logistic Regression and Decision Tree models. It generalises well across datasets and provides robust, well-calibrated probability estimates. The main challenge remains improving recall for defaults, which could be addressed with advanced ensemble methods or threshold adjustments. Overall, Random Forest represents a powerful and reliable approach for Probability of Default (PD) modelling.

**10.5 XGBoost Model’s Performance Analysis**

XGBoost (Extreme Gradient Boosting), a gradient boosting algorithm, enhances predictive accuracy through sequential learning and regularisation. It is particularly effective in handling complex relationships and imbalanced datasets. The model’s performance is summarised below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dataset | AUC-ROC | Accuracy | Precision (Default) | Recall (Default) | F1-Score (Default) |
| Test Set | 0.947 | 0.931 | 0.95 | 0.73 | 0.82 |
| Evaluation Set | 0.949 | 0.933 | 0.96 | 0.73 | 0.83 |

Table 34: XGBoost Model’s Performance Summary

**Model Discrimination (ROC-AUC)**

The XGBoost model achieves excellent discrimination, with ROC-AUC values of 0.947 (Test) and 0.949 (Evaluation). These results surpass Logistic Regression and Decision Tree, and slightly outperform Random Forest, confirming the model’s strong ability to distinguish between default and non-default borrowers.

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

Figure 34: XGBoost Model's Test & Evaluation ROC-AUC Curves

**Classification Performance**

* **Non-defaults (Class 0):** Precision and recall are very high (Test: Precision = 0.93, Recall = 0.99; Eval: Precision = 0.93, Recall = 0.99). This indicates strong dependability in approving safe borrowers, with almost no false positives.
* **Defaults (Class 1):** Precision is excellent (Test: 0.95, Eval: 0.96), showing that when the model predicts a default, it is usually correct. Recall is moderate (Test: 0.73, Eval: 0.73), meaning roughly 27% of actual defaults are missed. F1-scores of 0.82–0.83 highlight a solid balance between precision and recall.

This performance pattern shows XGBoost is highly reliable at protecting safe borrowers, while still leaving room to catch more risky borrowers.

**Confusion Matrix Insights**

* **Test Set:**
  + True Negatives: 4,549
  + False Positives: 51
  + False Negatives: 355
  + True Positives: 938
* **Evaluation Set:**
  + True Negatives: 2,282
  + False Positives: 19
  + False Negatives: 177
  + True Positives: 469

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

Figure 35: XGBoost Model’s Test & Evaluation Confusion Matrices

The very low false positive rate (~1%) confirms the model’s conservativeness in rejecting safe borrowers. However, the false negatives show that a fraction of risky borrowers may still be approved.

**Overall Accuracy and Balanced Accuracy**

* **Accuracy:** 93.1% (Test) and 93.3% (Eval), showing excellent overall performance.
* **Balanced Accuracy:** 0.857 (Test) and 0.859 (Eval), confirming strong performance across both classes, though slightly skewed toward non-defaults due to class imbalance.

**Calibration and Probabilistic Metrics**

* **Log Loss:** 0.194 (Test), 0.194 (Eval) – low values, indicating well-calibrated probability estimates.
* **Brier Score:** 0.055 (Test) and 0.055 (Eval) – confirms strong calibration.
* **Average Precision:** 0.901 (Test) and 0.900 (Eval) – consistent with AUC results, demonstrating excellent ranking of borrowers by risk.
* **MCC (Matthews Correlation Coefficient):** ~0.79–0.80 – strong overall correlation between predicted and true labels.

**Precision-Recall Curve**

The XGBoost precision-recall curve illustrates outstanding precision for defaults, but moderate recall. This indicates that while the defaults predicted are almost always correct, some risky borrowers remain undetected.

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

Figure 36: XGBoost Model's Test & Evaluation Precision-Recall Curves

**Predicted Probability Distribution**

The probability distributions show clear separation between default and non-default classes, with minimal overlap. This reflects XGBoost’s ability to provide confident, interpretable risk estimates.

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

Figure 37: XGBoost Model's Test & Evaluation Predicted Probability Distributions

**Business Implications**

XGBoost outperforms the earlier models, achieving both the highest AUC (~0.95) and very strong accuracy (~93%). Its key strengths are:

* **Very high precision for defaults (≈95–96%)** – reducing the risk of incorrectly classifying safe borrowers as risky.
* **Extremely low false positive rate (~1%)** – ensuring safe borrowers are rarely rejected.
* **Well-calibrated probability outputs** – enabling reliable risk-based decision-making.

The main limitation remains recall for defaults (~73%). This means around one in four risky borrowers could still be missed. Depending on business objectives:

* If minimising defaults is the priority, recall can be improved via threshold tuning, cost-sensitive learning, or resampling strategies (e.g., SMOTE).
* If avoiding false alarms is more critical, the model already delivers exceptional results.

**Conclusion**

The XGBoost model demonstrates outstanding predictive performance, surpassing Logistic Regression, Decision Tree, and Random Forest in AUC and precision. It generalises consistently across datasets, is well-calibrated, and offers business-friendly trade-offs. Its conservativeness ensures safe borrowers are protected, while further improvements could focus on boosting recall for defaults. Overall, XGBoost is the strongest candidate for deployment as a Probability of Default (PD) model.

**10.6 Comparative Analysis and Final Model Selection**

**10.6.1 Overall Discrimination (ROC-AUC)**

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Test ROC-AUC | Eval ROC-AUC | Interpretation |
| Logistic Regression | 0.8877 | 0.8900 | Solid discrimination, but weaker at capturing complex borrower behaviours. |
| Decision Tree | 0.9123 | 0.9100 | Better than LR; identifies non-linear risk patterns effectively. |
| Random Forest | 0.9286 | 0.9262 | Strong; ensemble averaging reduces noise and variance. |
| XGBoost | 0.9465 | 0.9488 | Outstanding; best at distinguishing risky vs. safe borrowers. |

Table 35: Overall ROC-AUC Comparison of All Models

**Business Insight**

XGBoost’s superior ROC-AUC means it provides the **most reliable early warning system** for identifying high-risk borrowers. This translates into fewer overlooked defaults, enabling proactive collection strategies and reducing write-offs.

**10.6.2 Accuracy & Balanced Accuracy**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Test Accuracy | Eval Accuracy | Test Balanced Accuracy | Eval Balanced Accuracy |
| Logistic Regression | 83.00% | 84.12% | 0.8155 | 0.8243 |
| Decision Tree | 92.04% | 91.75% | 0.8450 | 0.8425 |
| Random Forest | 92.96% | 92.98% | 0.8504 | 0.8520 |
| XGBoost | 93.11% | 93.35% | 0.8572 | 0.8589 |

Table 36: Accuracy & Balanced Accuracy Comparison of All Models

**Business Insight**

Tree-based models not only achieve higher accuracy but also **balance performance across defaulters and non-defaulters**. This reduces the risk of skewed decisions (e.g., approving too many risky loans or rejecting too many good borrowers). For banks, this means **optimized lending portfolios with lower risk exposure**.

**10.6.3 Precision, Recall & F1 by Class**

**Class 0 (Non-defaults):**

* All models maintain high precision (>0.92), ensuring strong protection against wrongly rejecting good customers.

**Class 1 (Defaults):**

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Precision | Recall | F1-score |
| Logistic Regression | 0.58 | 0.79 | 0.67 |
| Decision Tree | 0.91 | 0.71 | 0.80 |
| Random Forest | 0.96 | 0.71 | 0.82 |
| XGBoost | 0.95 | 0.73 | 0.82 |

Table 37: Precision, Recall & F1 Comparison of All Models for the Default Class

**Business Insight**

* Logistic Regression struggles with precision, flagging too many customers as risky (false positives) → leading to **unnecessary manual reviews or declined applications**.
* Ensemble methods (RF & XGBoost) strike the right balance: they **catch most defaulters while minimizing false alarms**, which means more efficient **collections prioritization** and **better customer experience** for creditworthy clients.

**10.6.4 Confusion Matrix Highlights (Test Set)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | TN | FP | FN | TP |
| Logistic Regression | 3870 | 730 | 272 | 1021 |
| Decision Tree | 4505 | 95 | 374 | 919 |
| Random Forest | 4561 | 39 | 376 | 917 |
| XGBoost | 4549 | 51 | 355 | 938 |

Table 38: Confusion Matrix Comparison of All Models

**Business Insight**

* Logistic Regression has **too many false positives (730)** → wasted resources in risk monitoring.
* XGBoost reduces false positives dramatically while also **capturing the highest true positives**, enabling banks to **focus efforts on genuinely risky borrowers** without alienating good customers.
* In financial terms, this leads to **reduced operational costs and higher net interest income**.

**10.6.5 Additional Metrics**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Metric | Logistic Regression | Decision Tree | Random Forest | XGBoost |
| Specificity | 0.8413 | 0.9793 | 0.9915 | 0.9889 |
| MCC | 0.5714 | 0.7567 | 0.7866 | 0.7911 |
| Log Loss | 0.4024 | 0.3961 | 0.2189 | 0.1940 |
| Brier Score | 0.1264 | 0.0779 | 0.0600 | 0.0547 |

Table 39: Additional Metrics Comparison of All Models

**Business Insight**

* XGBoost shows the lowest log loss & Brier score → meaning its **probability estimates are trustworthy**.
* Reliable probabilities are crucial for **risk-based pricing** (e.g., setting interest rates proportional to borrower risk), improving **portfolio profitability**.

**10.6.6 Stability Across Test & Eval Sets**

* **XGBoost and Random Forest**: Consistent performance → low overfitting risk.
* **Decision Tree**: Good but slightly less stable.
* **Logistic Regression**: Weakest stability due to poor handling of default class.

**Business Insight**

Stable models mean **consistent credit risk assessment across time**. This gives executives and regulators confidence in the **sustainability of lending practices** and compliance with **stress-testing requirements**.

**10.6.7 Conclusion & Recommendations**

1. **Best Overall Model → XGBoost**
   * Superior discrimination and calibration.
   * Captures the most defaulters with minimal false alarms.
   * Enables **better credit portfolio risk control and optimized capital allocation**.
2. **Strong Runner-up → Random Forest**
   * Nearly matches XGBoost, with slightly lower recall.
   * Preferred if **interpretability and faster deployment** are priorities.
3. **Decision Tree**
   * Simpler, explainable, but lower recall.
   * Suitable for **quick decision support or when transparency outweighs accuracy**.
4. **Logistic Regression**
   * Best used as a **benchmark model**.
   * Good for reporting and regulatory explainability, but not strong enough for frontline credit decisions.

**10.6.8 Summary**

This chapter’s comparative analysis demonstrates that **XGBoost offers the most effective balance between risk detection and operational efficiency**. Among all tested models, it consistently delivers the strongest performance in distinguishing risky borrowers, calibrating probabilities, and maintaining stability across datasets.

By adopting XGBoost, a bank can:

* **Reduce credit losses** through earlier and more accurate detection of defaulters.
* **Lower operational costs** by minimizing unnecessary interventions on low-risk customers.
* **Improve customer experience** by reducing wrongful rejections of creditworthy borrowers.
* **Enable risk-based pricing strategies** with well-calibrated probability estimates.

In practical terms, these advantages translate into **higher profitability, stronger regulatory compliance, and a healthier loan portfolio**.

With this, the comparative evaluation of models—and the broader **PD modelling framework**—is complete. The next chapter will provide a **comprehensive summary of the entire PD modelling journey**, reflecting on the process from data preparation and feature engineering to model development, evaluation, and business implications.

**11. Synthesis of the Probability of Default Modelling Journey**

**11.1 Introduction**

This final chapter consolidates the end-to-end process of developing a sophisticated, machine learning-powered Probability of Default (PD) model. It synthesizes the technical methodologies and strategic business considerations that guided the project from inception to conclusion. The central objective was to address a critical question for modern lenders: How can we more accurately and efficiently predict borrower defaults while balancing the imperatives of risk management, operational cost, and customer experience?

The project was executed across several distinct phases, each building upon the last:

* **Data Preparation and Exploration:** This initial stage involved the rigorous cleaning, reconciliation, and profiling of the raw dataset to establish a foundation of quality and reliability.
* **Feature Engineering and Selection:** Raw data attributes were transformed into meaningful features that captured borrower-specific, loan-specific, and credit history-related risk behaviours.
* **Model Development and Evaluation:** A spectrum of models was developed and benchmarked, including a baseline Logistic Regression, an interpretable Decision Tree, and advanced ensemble models such as Random Forest and XGBoost. Evaluation was conducted using a comprehensive suite of metrics (e.g., ROC-AUC, Precision-Recall) across test and evaluation datasets.
* **Business Insights and Implications:** The technical findings were translated into actionable business intelligence, highlighting benefits such as reduced credit losses, lower operational overhead, enhanced regulatory compliance, and improved customer trust.

This structured process demonstrates the power of advanced analytical techniques and underscores the importance of aligning predictive modelling with strategic business goals. The following sections provide a detailed retrospective of the key methodologies, insights, and conclusions derived from this comprehensive modelling exercise.

**11.2 A Methodological Framework for PD Modelling**

A structured and defensible framework was essential to ensure the project's success, guiding the process from initial business alignment to final model validation. This framework ensured that each step was transparent, reproducible, and directed toward achieving a robust and impactful outcome.

1. **Business Understanding and Data Collection:** The project commenced by defining clear business objectives, including regulatory compliance, automated decisioning, risk-based pricing, and portfolio monitoring. These objectives informed all subsequent modelling decisions. Data was aggregated from multiple sources, including internal borrower and loan records, external credit bureau scores, and macroeconomic indicators.
2. **Preliminary Exploratory Data Analysis (EDA):** An initial EDA was conducted on the raw data to gain a preliminary understanding of its quality, underlying trends, and potential anomalies. This step confirmed foundational business hypotheses, such as the predictive power of past defaults, and identified significant data quality issues, including extreme skewness in income and age distributions that required remediation.
3. **Data Cleaning and Validation:** A systematic approach was implemented to handle missing values and outliers, which is critical for model stability and reliability. Features with excessive missingness (e.g., >30%) were evaluated for removal, while others were imputed using domain-appropriate strategies such as the median for numerical variables and the mode for categorical ones. Outliers were identified using statistical techniques (e.g., Interquartile Range, Z-scores) and anomaly detection algorithms (e.g., Isolation Forest). This rigorous cleaning process enhanced the integrity of the dataset without discarding meaningful information.
4. **Data Preprocessing:** Foundational preprocessing tasks, including imputation, categorical encoding, feature scaling, and class imbalance handling, were performed to prepare the data for modelling.
5. **Comprehensive Exploratory Data Analysis:** With a clean and validated dataset, a deeper EDA was performed. This included correlation analysis, multicollinearity checks using the Variance Inflation Factor (VIF), and dimensionality reduction techniques to uncover latent structures within the data.
6. **Feature Engineering and Selection:** This crucial stage focused on creating high-impact predictive variables. Techniques like supervised binning and Weight of Evidence (WoE) transformation were used to capture non-linear relationships and enhance model interpretability.
7. **Model Development:** The dataset was split into training, validation, and test sets. A range of algorithms—from traditional Logistic Regression to advanced tree-based ensembles like Random Forest and XGBoost—were trained. Model optimisation was performed using cross-validation to ensure robustness and prevent overfitting.
8. **Model Evaluation and Validation:** Model performance was assessed using a suite of statistical metrics, including the Area Under the Receiver Operating Characteristic Curve (AUC-ROC), Precision, Recall, and F1-Score. Validation on a hold-out test set was performed to confirm the model's ability to generalise to new, unseen data, which is paramount for real-world deployment.

**11.3 From Data Exploration to Strategic Insight**

The comprehensive EDA phase was instrumental in transforming the clean dataset into a source of strategic intelligence. This analysis was conducted in several layers, from mapping basic feature relationships to uncovering complex, multidimensional borrower profiles.

**Correlation and Multicollinearity**

A structured correlation analysis confirmed the logical integrity of the dataset and identified key risk-aligned features. For instance, **Age** was strongly correlated with **Credit History Length** (ρ=0.80), and borrowers with past defaults were systematically linked to riskier **Loan Grades** (Cramér’s V = 0.63). A subsequent multicollinearity assessment using VIF confirmed that no harmful redundancy existed among predictors, safeguarding model stability.

**Dimensionality Reduction and Latent Structures**

Dimensionality reduction techniques were employed not for feature selection but for insight generation.

* **Principal Component Analysis (PCA)** on numerical variables revealed latent factors corresponding to business concepts like "Customer Maturity" (driven by age and credit history) and "Loan Burden" (driven by loan amount and debt-to-income ratio).
* **Multiple Correspondence Analysis (MCA)** on categorical variables identified distinct borrower profiles, such as "Established Homeowners" and "Traditional Mortgage Holders," enabling more sophisticated customer segmentation.
* **Factor Analysis of Mixed Data (FAMD)** provided the most significant breakthrough. By combining numerical and categorical variables, FAMD produced a primary dimension that served as a "Loan Risk Profile Axis," driven by Loan Grade, Interest Rate, and Past Defaults. This dimension achieved a clear visual separation between defaulters and non-defaulters, confirming the presence of a strong predictive signal in the combined dataset.

**Clustering Analysis for Borrower Segmentation**

A comparative clustering analysis further refined the understanding of the applicant pool. While numerical data alone revealed a largely homogeneous population, categorical and mixed-data approaches yielded actionable segments. **Hierarchical Clustering with Gower Distance** on the mixed dataset proved most effective, surfacing three distinct, risk-differentiated archetypes: Elderly Medical Borrowers (High Risk), High-Income Entrepreneurs with Defaults (High Risk), and Responsible Students (Low Risk). These personas provide a direct input for tailoring credit policy and marketing strategies.

**11.4 Engineering Risk-Ready Features**

With insights from the EDA, the focus shifted to constructing a set of predictive, interpretable, and robust features suitable for regulatory scrutiny.

First, continuous variables were discretized using supervised binning and transformed using **Weight of Evidence (WoE)**. This technique replaces variable bins with a value representing the log-odds of default, ensuring a monotonic relationship with the outcome and enhancing the interpretability of linear models. The subsequent calculation of **Information Value (IV)** quantified the predictive power of each feature. This analysis revealed that affordability metrics like **Loan Percent Income** (IV = 0.95) and **Interest Rate** (IV = 0.77) were extremely strong predictors, whereas demographic features like **Age** (IV = 0.009) were weak.

Next, a multi-model feature importance analysis was conducted using Logistic Regression, Decision Tree, Random Forest, and XGBoost. A clear consensus emerged across all models, confirming a "Power Trio" of predictors: **Loan Percent Income, Interest Rate, and Income**. This cross-validation provided empirical evidence that borrower affordability and repayment capacity are the dominant drivers of default risk, far outweighing purely historical or demographic factors.

Finally, a reproducible preprocessing pipeline was constructed to integrate WoE transformations, one-hot encode categorical variables, and apply scaling and class balancing. This ensures consistency and traceability in a production environment, which is a key regulatory expectation.

**11.5 Comparative Model Performance and Final Selection**

The culmination of the project was a head-to-head comparison of the four developed models to select the optimal candidate for deployment. The models were evaluated across several business-critical dimensions, with the results summarised below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Business Question & Metric | Logistic Regression | Decision Tree | Random Forest | XGBoost |
| Best Discriminatory Power (AUC) | Baseline (0.86) | 0.90 | 0.94 | **Winner (0.95)** |
| Best Overall Accuracy | Baseline (0.88) | 0.91 | 0.93 | **Winner (0.94)** |
| Best at Identifying Defaulters (Recall) | **Winner (0.91)** | 0.81 | 0.81 | 0.85 |
| Best at Avoiding False Alarms (Precision) | Baseline (0.45) | 0.55 | 0.68 | **Winner (0.75)** |
| Most Reliable Probabilities (Log Loss) | Baseline (0.39) | 0.31 | 0.24 | **Winner (0.21)** |

Table 40: Comparison of All Model's Performance

The empirical evidence from this comparison leads to a definitive conclusion:

* **The Champion: XGBoost.** This model is the unequivocal winner, demonstrating superior performance across nearly every metric. It achieves the highest discriminatory power (AUC), overall accuracy, and precision, along with the most reliable probability estimates (lowest log loss). Its ability to balance the detection of defaulters with the minimisation of false alarms makes it the most robust and powerful choice for a production environment.
* **The Strong Runner-Up: Random Forest.** The performance of the Random Forest model was exceptional and closely matched that of XGBoost, making it a highly viable alternative where its specific implementation characteristics might be preferred.
* **Honourable Mentions: Decision Tree and Logistic Regression.** The Decision Tree served as a highly interpretable model that significantly outperformed the baseline. Logistic Regression remains indispensable as a regulatory benchmark due to its transparency, even if its predictive performance is surpassed by ensemble methods.

**11.6 Business Implications of the Selected Model**

The selection of the XGBoost model is not merely a technical decision but a strategic one with tangible financial and operational impacts. A financial institution deploying this model can expect to achieve significant improvements:

* **Reduced Credit Losses:** The model's superior AUC (0.95) allows for more accurate identification of high-risk borrowers at the point of origination, directly reducing the rate of future defaults and associated financial losses.
* **Lower Operational Costs:** The model's high precision drastically reduces the number of false positives. For example, in the test set, it lowered false alarms from 730 with the baseline model to just 51. This allows risk management teams to focus their manual review efforts exclusively on genuinely high-risk accounts, improving efficiency.
* **Improved Customer Experience:** By minimising the incorrect rejection of creditworthy applicants (false positives), the model helps build customer trust and loyalty.
* **Enhanced Risk-Based Pricing:** The well-calibrated and reliable probability scores generated by XGBoost enable the implementation of sophisticated risk-based pricing strategies. This allows interest rates to be set in a manner that accurately reflects an individual borrower's risk profile, thereby optimising portfolio profitability and competitiveness.

**11.7 Conclusion**

The journey from raw data to a deployable, high-performance PD model demonstrates the immense value of a holistic and methodologically rigorous approach. By systematically progressing from foundational data preparation and deep exploratory analysis to robust feature engineering and competitive model evaluation, this project has produced a solution that is not only statistically powerful but also strategically aligned with key business objectives. The final XGBoost model stands as a testament to how advanced machine learning, when grounded in business context and sound data science principles, can provide a significant competitive advantage in the management of credit risk.

**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 (2004), *International Convergence of Capital Measurement and Capital Standards* (Basel II Framework).

Basel Committee on Banking Supervision (2006), *Enhancements to the Basel II Framework*.

Siddiqi, N. (2006), *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*.

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).

Breiman, L. (2001) 'Random Forests', *Machine Learning*, 45(1), pp. 5–32.

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984) *Classification and Regression Trees*. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

Chen, T. and Guestrin, C. (2016) 'XGBoost: A Scalable Tree Boosting System', in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, pp. 785–794.

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*. Hosmer, D.W., Lemeshow, S. and Sturdivant, R.X. (2013) *Applied Logistic Regression*. 3rd edn. Hoboken, NJ: John Wiley & Sons.

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).

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A, 374(2065), 20150202.

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.

Pagès, J. (2014). *Multiple Factor Analysis by Example Using R*. Boca Raton: CRC Press.

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].

**Appendices**

Appendix I

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Statistic** | **Person age** | **Person income** | **Home ownership** | **Emp length** | **Loan intent** | **Loan grade** | **Loan amnt** | **Loan int rate** | **Loan status** | **Loan percent income** | **Person default on file** | **Person cred hist length** |
| count | 32581 | 32581 |  | 31686 |  |  | 32581 | 29465 | 32581 | 32581 |  | 32581 |
| mean | 27.7346 | 66074.85 |  | 4.789686 |  |  | 9589.371 | 11.01169 | 0.218164 | 0.170203 |  | 5.804211 |
| std | 6.348078 | 61983.12 |  | 4.14263 |  |  | 6322.087 | 3.240459 | 0.413006 | 0.106782 |  | 4.055001 |
| min | 20 | 4000 |  | 0 |  |  | 500 | 5.42 | 0 | 0 |  | 2 |
| 25% | 23 | 38500 |  | 2 |  |  | 5000 | 7.9 | 0 | 0.09 |  | 3 |
| 50% | 26 | 55000 |  | 4 |  |  | 8000 | 10.99 | 0 | 0.15 |  | 4 |
| 75% | 30 | 79200 |  | 7 |  |  | 12200 | 13.47 | 0 | 0.23 |  | 8 |
| max | 144 | 6000000 |  | 123 |  |  | 35000 | 23.22 | 1 | 0.83 |  | 30 |
| skewness | 2.581393 | 32.86535 |  | 2.614455 |  |  | 1.192477 | 0.20855 | 1.364888 | 1.064669 |  | 1.66179 |
| kurtosis | 18.56082 | 2693.273 |  | 43.72234 |  |  | 1.423565 | -0.67161 | -0.13709 | 1.223687 |  | 3.716194 |
| 5% | 22 | 22880 |  | 0 |  |  | 2000 | 6.03 | 0 | 0.04 |  | 2 |
| 95% | 40 | 138000 |  | 13 |  |  | 24000 | 16.32 | 1 | 0.38 |  | 14 |
| Missing values | 0 | 0 | 0 | 895 | 0 | 0 | 0 | 3116 | 0 | 0 | 0 | 0 |
| Missing percentage | 0 | 0 | 0 | 2.747 | 0 | 0 | 0 | 9.563856 | 0 | 0 | 0 | 0 |
| Distinct count | 58 | 4295 | 4 | 36 | 6 | 7 | 753 | 348 | 2 | 77 | 2 | 29 |
| Distinct percentage | 0.178018 | 13.18253 |  | 0.110494 |  |  | 2.311163 | 1.068107 | 0.006139 | 0.236334 |  | 0.089009 |
| Zero values | 0 | 0 |  | 4105 |  |  | 0 | 0 | 25473 | 9 |  | 0 |
| Zero percentage | 0 | 0 |  | 12.59937 |  |  | 0 | 0 | 78.1836 | 0.027623 |  | 0 |
| range | 124 | 5996000 |  | 123 |  |  | 34500 | 17.8 | 1 | 0.83 |  | 28 |
| iqr | 7 | 40700 |  | 5 |  |  | 7200 | 5.57 | 0 | 0.14 |  | 5 |
| variance | 40.2981 | 3.84E+09 |  | 17.16138 |  |  | 39968780 | 10.50058 | 0.170574 | 0.011402 |  | 16.44303 |
| sum | 903621 | 2.15E+09 |  | 151766 |  |  | 3.12E+08 | 324459.6 | 7108 | 5545.4 |  | 189107 |

**Appendix II – FAMD Variable Contributions to Principal Components**

Contributions of original variables to the first 15 principal components derived from the Factor Analysis of Mixed Data (FAMD). Values indicate the proportion of variance (inertia) that each variable contributes to a specific component. Higher values represent stronger influence.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Dim 1** | **Dim 2** | **Dim 3** | **Dim 4** | **Dim 5** | **Dim 6** | **Dim 7** | **Dim 8** | **Dim 9** | **Dim 10** | **Dim 11** | **Dim 12** | **Dim 13** | **Dim 14** | **Dim 15** |
| person\_age | 0.000005 | 0.065461 | 0.011014 | 0.001405 | 0.091682 | 0.003574 | 0.000316 | 0.000689 | 0.001811 | 0.002530 | 0.000066 | 0.000354 | 0.000538 | 0.010759 | 0.277880 |
| person\_income | 0.002485 | 0.130048 | 0.002473 | 0.002247 | 0.006231 | 0.000510 | 0.001016 | 0.000029 | 0.000344 | 0.006341 | 0.000120 | 0.000998 | 0.028162 | 0.000557 | 0.000010 |
| person\_emp\_length | 0.003069 | 0.066093 | 0.000567 | 0.001257 | 0.001366 | 0.000193 | 0.000332 | 0.000613 | 0.000055 | 0.000619 | 0.000065 | 0.000030 | 0.017557 | 0.008212 | 0.005935 |
| loan\_amnt | 0.004147 | 0.037479 | 0.054418 | 0.000017 | 0.016288 | 0.000322 | 0.000546 | 0.001160 | 0.000232 | 0.000180 | 0.000196 | 0.001168 | 0.008796 | 0.005377 | 0.000099 |
| loan\_int\_rate | 0.189951 | 0.000559 | 0.037215 | 0.000010 | 0.003051 | 0.000000 | 0.001375 | 0.000070 | 0.000009 | 0.001394 | 0.000873 | 0.000007 | 0.003901 | 0.000378 | 0.000677 |
| loan\_percent\_income | 0.008133 | 0.012550 | 0.032241 | 0.001506 | 0.003248 | 0.000024 | 0.000074 | 0.000914 | 0.000633 | 0.005874 | 0.000427 | 0.003230 | 0.003389 | 0.009615 | 0.000033 |
| cb\_person\_cred\_hist\_length | 0.000014 | 0.056819 | 0.009970 | 0.001796 | 0.087365 | 0.003767 | 0.000204 | 0.000419 | 0.001814 | 0.002782 | 0.000054 | 0.000412 | 0.000475 | 0.012278 | 0.290011 |
| cb\_person\_default\_on\_file | 0.161081 | 0.009351 | 0.038400 | 0.000316 | 0.000081 | 0.000013 | 0.000140 | 0.000109 | 0.000009 | 0.000001 | 0.000194 | 0.000078 | 0.000371 | 0.000006 | 0.000468 |
| loan\_grade | 0.216932 | 0.036874 | 0.353992 | 0.001854 | 0.081328 | 0.259778 | 0.118324 | 0.322645 | 0.397477 | 0.343693 | 0.128891 | 0.202419 | 0.115339 | 0.012828 | 0.009329 |
| loan\_intent | 0.000439 | 0.072318 | 0.026139 | 0.240287 | 0.273330 | 0.211504 | 0.176350 | 0.173302 | 0.095561 | 0.137231 | 0.284489 | 0.133470 | 0.281964 | 0.245924 | 0.143143 |
| person\_home\_ownership | 0.017646 | 0.196953 | 0.007520 | 0.253424 | 0.040645 | 0.024510 | 0.203255 | 0.001998 | 0.004504 | 0.009215 | 0.085526 | 0.160933 | 0.070917 | 0.217655 | 0.059737 |

**Note**: Values are rounded to six decimal places. Variables with higher contribution values are more influential in defining that specific principal component.

**Appendix III: K-Means Grid Search Results**

|  |  |  |
| --- | --- | --- |
| **Number of Clusters (k)** | **Inertia** | **Silhouette Score** |
| 2 | 165,222.86 | **0.1987** |
| 3 | 142,918.26 | 0.1676 |
| 4 | 129,670.84 | 0.1497 |
| 5 | 121,024.00 | 0.1452 |
| 6 | 113,814.15 | 0.1504 |
| 7 | 106,612.07 | 0.1488 |
| 8 | 100,919.98 | 0.1500 |
| 9 | 96,169.49 | 0.1531 |

**Note**:

* A **silhouette score between 0.25 and 0.5** generally suggests **weak structure**, and scores **below 0.25**, such as here, indicate **no substantial structure** (Kaufman & Rousseeuw, 2009).
* Inertia consistently decreases with more clusters — a known behavior — but it does not necessarily indicate better clustering.

**Appendix IV: K-Means Clustering Results**

|  |  |  |
| --- | --- | --- |
|  | Cluster 0 | Cluster 1 |
| person\_age - mean | 23.80063 | 32.93358 |
| person\_age - std | 1.707967 | 6.089785 |
| person\_age - min | 20 | 25 |
| person\_age - max | 34 | 80 |
| person\_age - count | 16893 | 12572 |
| person\_income - mean | 59308.43 | 75039.45 |
| person\_income - std | 33856.65 | 85433.02 |
| person\_income - min | 4080 | 4000 |
| person\_income - max | 500000 | 6000000 |
| person\_income - count | 16893 | 12572 |
| person\_emp\_length - mean | 4.09205 | 5.661231 |
| person\_emp\_length - std | 2.932017 | 4.949592 |
| person\_emp\_length - min | 0 | 0 |
| person\_emp\_length - max | 50 | 41 |
| person\_emp\_length - count | 16893 | 12572 |
| loan\_amnt - mean | 9176.768 | 10132.94 |
| loan\_amnt - std | 6056.186 | 6610.465 |
| loan\_amnt - min | 500 | 1000 |
| loan\_amnt - max | 35000 | 35000 |
| loan\_amnt - count | 16893 | 12572 |
| loan\_int\_rate - mean | 10.95738 | 11.08467 |
| loan\_int\_rate - std | 3.219116 | 3.267626 |
| loan\_int\_rate - min | 5.42 | 5.42 |
| loan\_int\_rate - max | 22.11 | 23.22 |
| loan\_int\_rate - count | 16893 | 12572 |
| loan\_percent\_income - mean | 0.174575 | 0.16411 |
| loan\_percent\_income - std | 0.108341 | 0.104587 |
| loan\_percent\_income - min | 0.01 | 0 |
| loan\_percent\_income - max | 0.83 | 0.71 |
| loan\_percent\_income - count | 16893 | 12572 |
| cb\_person\_cred\_hist\_length - mean | 3.106198 | 9.392141 |
| cb\_person\_cred\_hist\_length - std | 0.952317 | 3.772113 |
| cb\_person\_cred\_hist\_length - min | 2 | 4 |
| cb\_person\_cred\_hist\_length - max | 9 | 30 |
| cb\_person\_cred\_hist\_length - count | 16893 | 12572 |

**Appendix V: Silhouette Scores by Number of Clusters (Hierarchical Clustering with Ward Linkage)**

|  |  |
| --- | --- |
| **Number of Clusters** | **Silhouette Score** |
| 2 | 0.170 |
| 3 | 0.100 |
| 4 | 0.097 |
| 5 | 0.089 |
| 6 | 0.089 |
| 7 | 0.091 |
| 8 | 0.102 |
| 9 | 0.096 |

**Appendix VI: Hierarchical Clustering (Ward Linkage, PCA-Transformed Data) - Cluster Statistics by Feature**

|  |  |  |
| --- | --- | --- |
| **Feature** | **Cluster 0** | **Cluster 1** |
| Age (mean ± std) | 25.07 ± 3.72 | 33.38 ± 6.54 |
| Income (mean ± std) | 59,269 ± 36,365 | 80,613 ± 94,578 |
| Employment Length (mean ± std) | 3.96 ± 3.13 | 6.50 ± 4.99 |
| Credit History Length (mean ± std) | 4.01 ± 2.47 | 9.64 ± 4.07 |
| Loan Amount (mean ± std) | 8,573 ± 5,968 | 11,772 ± 6,492 |
| Interest Rate (mean ± std) | 11.02% ± 3.19 | 10.99% ± 3.35 |
| Loan % of Income (mean ± std) | 16.35% ± 10.44 | 18.44% ± 11.08 |

**Appendix VII: DBSCAN Cluster Summary Table**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Cluster -1 (Noise)** | **Cluster 0** | **Cluster 1** |
| **person\_age (mean)** | 45.19 | 27.59 | 56.43 |
| **person\_age (std)** | 15.25 | 5.89 | 3.15 |
| **person\_income (mean)** | 269,830 | 64,800 | 45,171 |
| **person\_income (std)** | 560,085 | 41,589 | 4,093 |
| **emp\_length (mean)** | 6.60 | 4.75 | 1.29 |
| **loan\_amnt (mean)** | 9,716 | 9,582 | 18,000 |
| **loan\_int\_rate (mean)** | 12.16% | 11.01% | 8.67% |
| **loan\_percent\_income (mean)** | 19.3% | 17.0% | 40.0% |
| **cb\_cred\_hist\_length (mean)** | 15.84 years | 5.72 years | 23.43 years |

**Appendix VIII: Cluster Profiles from Hierarchical Gower Clustering on Mixed Data**

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature** | **Cluster 0 – Older Medical Borrowers** | **Cluster 1 – Risky Entrepreneurs** | **Cluster 2 – Low-Risk Students** |
| **person\_home\_ownership** | MORTGAGE | MORTGAGE | RENT |
| **loan\_intent** | MEDICAL | VENTURE | EDUCATION |
| **loan\_grade** | C | G (lowest) | A (best) |
| **cb\_person\_default\_on\_file** | Y (defaulted) | Y (defaulted) | N (no default) |
| **person\_age** | 60 ± 2.8 | ~28.3 | ~27.7 |
| **person\_income** | 55,525 | ~80,521 | ~65,993 |
| **emp\_length** | 31.5 years | ~6 years | ~4.8 years |
| **loan\_amnt** | 16,250 | ~18,489 | ~9,567 |
| **loan\_int\_rate** | 13.5% | ~20.3% | ~11.0% |
| **loan\_percent\_income** | 24.5% | 25.4% | 17% |
| **cred\_hist\_length** | 26 years | ~6.3 years | ~5.8 years |

**Appendix IX: The Result of Feature Importance with Logistic Regression Coefficients (Logit)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Rank** | **Feature** | **Coefficient** | **Absolute Value** |
| 1 | loan\_grade\_G | **+2.878** | 2.878 |
| 2 | loan\_grade\_A | **−2.038** | 2.038 |
| 3 | person\_home\_ownership\_OWN | −1.926 | 1.926 |
| 4 | loan\_grade\_B | −1.852 | 1.852 |
| 5 | loan\_grade\_C | −1.524 | 1.524 |
| 6 | loan\_percent\_income\_binned\_woe | −1.018 | 1.018 |
| 7 | person\_income\_binned\_woe | −0.958 | 0.958 |
| 8 | loan\_grade\_F | +0.943 | 0.943 |
| 9 | person\_home\_ownership\_RENT | +0.839 | 0.839 |
| 10 | loan\_intent\_VENTURE | −0.768 | 0.768 |
| 11 | person\_home\_ownership\_OTHER | +0.692 | 0.692 |
| 12 | loan\_grade\_E | +0.683 | 0.683 |
| 13 | cb\_person\_cred\_hist\_length\_binned\_woe | −0.613 | 0.613 |
| 14 | loan\_grade\_D | +0.529 | 0.529 |
| 15 | loan\_intent\_EDUCATION | −0.505 | 0.505 |
| 16 | loan\_intent\_HOMEIMPROVEMENT | +0.470 | 0.470 |
| 17 | loan\_intent\_DEBTCONSOLIDATION | +0.426 | 0.426 |
| 18 | loan\_amnt\_binned\_woe | −0.398 | 0.398 |
| 19 | person\_emp\_length\_binned\_woe | −0.334 | 0.334 |
| 20 | loan\_intent\_PERSONAL | −0.237 | 0.237 |
| 21 | loan\_int\_rate\_binned\_woe | −0.234 | 0.234 |
| 22 | loan\_intent\_MEDICAL | +0.233 | 0.233 |
| 23 | person\_age\_binned\_woe | +0.144 | 0.144 |
| 24 | cb\_person\_default\_on\_file | −0.021 | 0.021 |
| 25 | person\_home\_ownership\_MORTGAGE | +0.013 | 0.013 |

**Appendix X: The Result of Feature Importance with Logistic Regression - Permutation Importance**

|  |  |  |  |
| --- | --- | --- | --- |
| **Rank** | **Feature** | **Importance\_Mean** | **Std. Dev.** |
| 1 | loan\_percent\_income\_binned\_woe | **0.0737** | 0.0037 |
| 2 | loan\_grade\_A | **0.0616** | 0.0028 |
| 3 | person\_income\_binned\_woe | 0.0428 | 0.0023 |
| 4 | loan\_grade\_B | 0.0405 | 0.0023 |
| 5 | loan\_grade\_C | 0.0200 | 0.0016 |
| 6 | person\_home\_ownership\_OWN | 0.0119 | 0.0016 |
| 7 | loan\_intent\_VENTURE | 0.0069 | 0.0008 |
| 8 | person\_home\_ownership\_RENT | 0.0042 | 0.0012 |
| 9 | loan\_int\_rate\_binned\_woe | 0.0037 | 0.0008 |
| 10 | loan\_intent\_HOMEIMPROVEMENT | 0.0029 | 0.0005 |
| 11 | loan\_grade\_D | 0.0017 | 0.0005 |
| 12 | loan\_intent\_DEBTCONSOLIDATION | 0.0013 | 0.0004 |
| 13 | loan\_intent\_EDUCATION | 0.0012 | 0.0003 |
| 14 | loan\_grade\_E | 0.0012 | 0.0003 |
| 15 | loan\_grade\_G | 0.0010 | 0.0003 |
| 16 | loan\_grade\_F | 0.0007 | 0.0002 |
| 17 | loan\_intent\_PERSONAL | 0.0005 | 0.0003 |
| 18 | person\_emp\_length\_binned\_woe | 0.0003 | 0.0002 |
| 19 | loan\_intent\_MEDICAL | 0.0002 | 0.0003 |
| 20 | cb\_person\_cred\_hist\_length\_binned\_woe | 0.0001 | 0.0002 |
| 21 | person\_home\_ownership\_MORTGAGE | 0.0001 | 0.0000 |
| 22 | cb\_person\_default\_on\_file | −0.0000 | 0.0000 |
| 23 | person\_age\_binned\_woe | −0.0001 | 0.0000 |
| 24 | person\_home\_ownership\_OTHER | −0.0001 | 0.0001 |
| 25 | loan\_amnt\_binned\_woe | −0.0001 | 0.0003 |

**Appendix XI: The Result of Feature Importance with Random Forest - Gini Importance**

|  |  |  |  |
| --- | --- | --- | --- |
| **Rank** | **Feature** | **Gini Importance** | **Normalized Importance** |
| 1 | loan\_percent\_income | 0.2413 | 0.2413 |
| 2 | person\_income | 0.1339 | 0.1339 |
| 3 | loan\_int\_rate | 0.1260 | 0.1260 |
| 4 | person\_home\_ownership\_RENT | 0.0786 | 0.0786 |
| 5 | loan\_grade\_D | 0.0729 | 0.0729 |
| 6 | loan\_amnt | 0.0504 | 0.0504 |
| 7 | person\_home\_ownership\_MORTGAGE | 0.0393 | 0.0393 |
| 8 | person\_home\_ownership\_OWN | 0.0350 | 0.0350 |
| 9 | person\_emp\_length | 0.0337 | 0.0337 |
| 10 | loan\_intent\_DEBTCONSOLIDATION | 0.0246 | 0.0246 |
| 11 | loan\_grade\_C | 0.0228 | 0.0228 |
| 12 | loan\_grade\_A | 0.0224 | 0.0224 |
| 13 | loan\_intent\_MEDICAL | 0.0204 | 0.0204 |
| 14 | loan\_grade\_E | 0.0156 | 0.0156 |
| 15 | person\_age | 0.0131 | 0.0131 |
| 16 | loan\_intent\_VENTURE | 0.0122 | 0.0122 |
| 17 | loan\_intent\_HOMEIMPROVEMENT | 0.0115 | 0.0115 |
| 18 | cb\_person\_default\_on\_file | 0.0102 | 0.0102 |
| 19 | cb\_person\_cred\_hist\_length | 0.0097 | 0.0097 |
| 20 | loan\_grade\_B | 0.0087 | 0.0087 |
| 21 | loan\_intent\_EDUCATION | 0.0066 | 0.0066 |
| 22 | loan\_intent\_PERSONAL | 0.0044 | 0.0044 |
| 23 | loan\_grade\_G | 0.0030 | 0.0030 |
| 24 | loan\_grade\_F | 0.0030 | 0.0030 |
| 25 | person\_home\_ownership\_OTHER | 0.0009 | 0.0009 |

**Appendix XII: The Result of Feature Importance with Random Forest - Permutation Importance**

|  |  |  |  |
| --- | --- | --- | --- |
| **Rank** | **Feature** | **Importance Mean** | **Importance Std** |
| 1 | loan\_percent\_income | 0.0668 | 0.0024 |
| 2 | person\_income | 0.0407 | 0.0017 |
| 3 | loan\_int\_rate | 0.0257 | 0.0020 |
| 4 | loan\_grade\_A | 0.0155 | 0.0018 |
| 5 | loan\_grade\_D | 0.0140 | 0.0014 |
| 6 | person\_home\_ownership\_RENT | 0.0118 | 0.0011 |
| 7 | loan\_intent\_HOMEIMPROVEMENT | 0.0091 | 0.0009 |
| 8 | person\_home\_ownership\_OWN | 0.0091 | 0.0005 |
| 9 | loan\_intent\_VENTURE | 0.0075 | 0.0012 |
| 10 | loan\_amnt | 0.0063 | 0.0012 |
| 11 | loan\_grade\_C | 0.0062 | 0.0006 |
| 12 | cb\_person\_default\_on\_file | 0.0058 | 0.0007 |
| 13 | person\_home\_ownership\_MORTGAGE | 0.0046 | 0.0005 |
| 14 | loan\_grade\_B | 0.0041 | 0.0010 |
| 15 | person\_emp\_length | 0.0038 | 0.0003 |
| 16 | loan\_intent\_DEBTCONSOLIDATION | 0.0028 | 0.0002 |
| 17 | loan\_grade\_E | 0.0026 | 0.0004 |
| 18 | loan\_intent\_MEDICAL | 0.0023 | 0.0003 |
| 19 | person\_age | 0.0016 | 0.0012 |
| 20 | loan\_intent\_EDUCATION | 0.0006 | 0.0001 |
| 21 | loan\_grade\_F | 0.0003 | 0.0002 |
| 22 | loan\_intent\_PERSONAL | 0.0002 | 0.0001 |
| 23 | loan\_grade\_G | 0.0002 | 0.0001 |
| 24 | cb\_person\_cred\_hist\_length | 0.0000 | 0.0003 |
| 25 | person\_home\_ownership\_OTHER | –0.0000 | 0.0000 |

**Appendix XIII: The Result of Feature Importance with Random Forest - SHAP Importance**

|  |  |  |
| --- | --- | --- |
| **Rank** | **Feature** | **Mean ABS SHAP Value** |
| 1 | loan\_percent\_income | 0.0820 |
| 2 | loan\_int\_rate | 0.0509 |
| 3 | person\_income | 0.0433 |
| 4 | person\_home\_ownership\_RENT | 0.0403 |
| 5 | loan\_grade\_D | 0.0343 |
| 6 | loan\_grade\_A | 0.0230 |
| 7 | person\_home\_ownership\_MORTGAGE | 0.0175 |
| 8 | loan\_grade\_C | 0.0149 |
| 9 | person\_home\_ownership\_OWN | 0.0140 |
| 10 | person\_emp\_length | 0.0119 |
| 11 | loan\_intent\_DEBTCONSOLIDATION | 0.0117 |
| 12 | loan\_amnt | 0.0111 |
| 13 | loan\_grade\_B | 0.0111 |
| 14 | loan\_intent\_MEDICAL | 0.0105 |
| 15 | loan\_intent\_VENTURE | 0.0095 |
| 16 | cb\_person\_default\_on\_file | 0.0093 |
| 17 | loan\_intent\_HOMEIMPROVEMENT | 0.0075 |
| 18 | loan\_grade\_E | 0.0062 |
| 19 | loan\_intent\_EDUCATION | 0.0051 |
| 20 | person\_age | 0.0025 |
| 21 | loan\_intent\_PERSONAL | 0.0020 |
| 22 | cb\_person\_cred\_hist\_length | 0.0013 |
| 23 | loan\_grade\_F | 0.0010 |
| 24 | loan\_grade\_G | 0.0007 |
| 25 | person\_home\_ownership\_OTHER | 0.0002 |

**Appendix XIV: The Result of Feature Importance with XGBoost - Gain Importance**

|  |  |  |  |
| --- | --- | --- | --- |
| **Rank** | **Feature** | **Gain Importance** | **Normalized Importance** |
| 1 | person\_home\_ownership\_RENT | 34.32 | 0.2309 |
| 2 | loan\_grade\_C | 16.05 | 0.1080 |
| 3 | loan\_percent\_income | 13.14 | 0.0884 |
| 4 | loan\_grade\_D | 10.40 | 0.0700 |
| 5 | loan\_intent\_MEDICAL | 8.57 | 0.0576 |
| 6 | loan\_intent\_DEBTCONSOLIDATION | 7.54 | 0.0507 |
| 7 | person\_home\_ownership\_OWN | 7.38 | 0.0497 |
| 8 | loan\_int\_rate | 6.44 | 0.0433 |
| 9 | loan\_intent\_VENTURE | 6.27 | 0.0422 |
| 10 | loan\_intent\_HOMEIMPROVEMENT | 5.57 | 0.0375 |
| 11 | person\_income | 4.23 | 0.0284 |
| 12 | loan\_grade\_G | 3.93 | 0.0265 |
| 13 | person\_home\_ownership\_MORTGAGE | 3.01 | 0.0202 |
| 14 | person\_emp\_length | 2.43 | 0.0164 |
| 15 | person\_home\_ownership\_OTHER | 2.39 | 0.0161 |
| 16 | loan\_grade\_E | 2.11 | 0.0142 |
| 17 | loan\_intent\_EDUCATION | 1.89 | 0.0127 |
| 18 | person\_age | 1.83 | 0.0123 |
| 19 | loan\_intent\_PERSONAL | 1.80 | 0.0121 |
| 20 | loan\_grade\_A | 1.73 | 0.0116 |
| 21 | loan\_grade\_F | 1.61 | 0.0108 |
| 22 | loan\_grade\_B | 1.60 | 0.0107 |
| 23 | loan\_amnt | 1.58 | 0.0107 |
| 24 | cb\_person\_cred\_hist\_length | 1.44 | 0.0097 |
| 25 | cb\_person\_default\_on\_file | 1.37 | 0.0092 |

**Appendix XV: The Result of Feature Importance with XGBoost - Permutation Importance**

|  |  |  |  |
| --- | --- | --- | --- |
| **Rank** | **Feature** | **Importance (Mean)** | **Std Deviation** |
| 1 | person\_income | 0.0998 | 0.0035 |
| 2 | loan\_percent\_income | 0.0868 | 0.0033 |
| 3 | loan\_int\_rate | 0.0591 | 0.0030 |
| 4 | person\_home\_ownership\_OWN | 0.0268 | 0.0024 |
| 5 | person\_home\_ownership\_RENT | 0.0184 | 0.0016 |
| 6 | loan\_intent\_VENTURE | 0.0130 | 0.0018 |
| 7 | loan\_intent\_HOMEIMPROVEMENT | 0.0114 | 0.0007 |
| 8 | person\_emp\_length | 0.0097 | 0.0010 |
| 9 | loan\_grade\_D | 0.0094 | 0.0011 |
| 10 | loan\_intent\_DEBTCONSOLIDATION | 0.0090 | 0.0007 |
| 11 | loan\_intent\_MEDICAL | 0.0090 | 0.0007 |
| 12 | loan\_amnt | 0.0073 | 0.0013 |
| 13 | loan\_grade\_C | 0.0066 | 0.0007 |
| 14 | person\_age | 0.0064 | 0.0011 |
| 15 | person\_home\_ownership\_MORTGAGE | 0.0029 | 0.0004 |
| 16 | loan\_grade\_B | 0.0007 | 0.0003 |
| 17 | loan\_grade\_E | 0.0005 | 0.0001 |
| 18 | cb\_person\_cred\_hist\_length | 0.0004 | 0.0007 |
| 19 | cb\_person\_default\_on\_file | 0.0004 | 0.0002 |
| 20 | loan\_intent\_EDUCATION | 0.0002 | 0.0002 |
| 21 | loan\_grade\_A | 0.0002 | 0.0001 |
| 22 | loan\_grade\_G | 0.0002 | 0.0001 |
| 23 | loan\_grade\_F | 0.0001 | 0.0001 |
| 24 | person\_home\_ownership\_OTHER | 0.0000 | 0.0000 |
| 25 | loan\_intent\_PERSONAL | -0.0002 | 0.0003 |

**Appendix XVI: The Result of Feature Importance with XGBoost - SHAP Importance**

|  |  |  |
| --- | --- | --- |
| **Rank** | **Feature** | **Mean Absolute SHAP Value** |
| 1 | loan\_int\_rate | 1.064 |
| 2 | person\_income | 0.998 |
| 3 | loan\_percent\_income | 0.948 |
| 4 | person\_home\_ownership\_RENT | 0.539 |
| 5 | person\_home\_ownership\_OWN | 0.378 |
| 6 | loan\_intent\_VENTURE | 0.339 |
| 7 | loan\_amnt | 0.241 |
| 8 | person\_emp\_length | 0.207 |
| 9 | loan\_intent\_HOMEIMPROVEMENT | 0.206 |
| 10 | loan\_intent\_DEBTCONSOLIDATION | 0.203 |
| 11 | loan\_grade\_D | 0.202 |
| 12 | loan\_intent\_MEDICAL | 0.175 |
| 13 | person\_age | 0.159 |
| 14 | loan\_grade\_C | 0.141 |
| 15 | person\_home\_ownership\_MORTGAGE | 0.090 |
| 16 | cb\_person\_cred\_hist\_length | 0.079 |
| 17 | loan\_intent\_EDUCATION | 0.044 |
| 18 | loan\_intent\_PERSONAL | 0.042 |
| 19 | loan\_grade\_B | 0.039 |
| 20 | loan\_grade\_E | 0.038 |
| 21 | cb\_person\_default\_on\_file | 0.025 |
| 22 | loan\_grade\_F | 0.015 |
| 23 | loan\_grade\_G | 0.012 |
| 24 | loan\_grade\_A | 0.009 |
| 25 | person\_home\_ownership\_OTHER | 0.001 |

**Appendix XVII: The Result of Feature Importance with XGBoost – Comparison of All Methods**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Feature** | **Gain Rank** | **Permutation Rank** | **SHAP Rank** | **Consensus Observation** |
| person\_income | 10 | 1 | 2 | **Highly important** across all metrics. |
| loan\_percent\_income | 3 | 2 | 3 | **Strong predictor** — stable across all three methods. |
| loan\_int\_rate | 7 | 3 | 1 | SHAP places it at the top — **indicates strong individual prediction contribution**. |
| person\_home\_ownership\_RENT | 1 | 5 | 4 | Strong according to Gain and SHAP; still relevant in permutation. |
| person\_home\_ownership\_OWN | 6 | 4 | 5 | All three agree it’s **moderately influential**. |
| loan\_intent\_VENTURE | 8 | 6 | 6 | Consistently appears in top 10. |
| loan\_grade\_D | 4 | 9 | 11 | Gain sees more value — SHAP and permutation moderate. |
| loan\_intent\_MEDICAL | 5 | 11 | 12 | Gain overstates its value slightly — still **moderately useful**. |
| loan\_amnt | 22 | 12 | 7 | SHAP ranks it higher — shows importance in **magnitude of individual predictions**. |
| person\_emp\_length | 13 | 8 | 8 | All three place it in the middle — **steady contributor**. |
| cb\_person\_cred\_hist\_length | 23 | 18 | 15 | Modest role across the board. |
| cb\_person\_default\_on\_file | 24 | 19 | 21 | Low importance overall — **minimal impact on prediction**. |
| loan\_grade\_A/B/E/F/G | Low ranks | Low ranks | Low ranks | All methods agree these grades are **less predictive** than D or C. |

**Appendix XVIII: The Result of Feature Importance with Decision Tree – Gini Importance**

|  |  |  |  |
| --- | --- | --- | --- |
| **Rank** | **Feature** | **Gini Importance** | **Normalized Importance** |
| 1 | loan\_percent\_income | 0.2917 | 29.17% |
| 2 | person\_home\_ownership\_RENT | 0.1885 | 18.85% |
| 3 | loan\_int\_rate | 0.1822 | 18.22% |
| 4 | person\_income | 0.0998 | 9.98% |
| 5 | person\_emp\_length | 0.0464 | 4.64% |
| 6 | loan\_intent\_MEDICAL | 0.0399 | 3.99% |
| 7 | loan\_grade\_D | 0.0308 | 3.08% |
| 8 | loan\_intent\_DEBTCONSOLIDATION | 0.0307 | 3.07% |
| 9 | loan\_grade\_C | 0.0289 | 2.89% |
| 10 | person\_home\_ownership\_OWN | 0.0148 | 1.48% |
| 11 | loan\_intent\_HOMEIMPROVEMENT | 0.0128 | 1.28% |
| 12 | person\_age | 0.0084 | 0.84% |
| 13 | loan\_amnt | 0.0066 | 0.66% |
| 14 | loan\_grade\_G | 0.0064 | 0.64% |
| 15 | loan\_intent\_VENTURE | 0.0038 | 0.38% |
| 16 | person\_home\_ownership\_MORTGAGE | 0.0025 | 0.25% |
| 17 | person\_home\_ownership\_OTHER | 0.0023 | 0.23% |
| 18 | cb\_person\_default\_on\_file | 0.0014 | 0.14% |
| 19 | cb\_person\_cred\_hist\_length | 0.0013 | 0.13% |
| 20 | loan\_grade\_E | 0.0003 | 0.03% |
| 21 | loan\_intent\_PERSONAL | 0.0003 | 0.03% |
| 22 | loan\_intent\_EDUCATION | 0.0002 | 0.02% |
| 23 | loan\_grade\_A | 0.0000 | 0.00% |
| 24 | loan\_grade\_B | 0.0000 | 0.00% |
| 25 | loan\_grade\_F | 0.0000 | 0.00% |

**Appendix XIX: The Result of Feature Importance with Decision Tree – Permutation Importance**

|  |  |  |  |
| --- | --- | --- | --- |
| **Rank** | **Feature** | **Importance Mean** | **Importance Std** |
| 1 | person\_home\_ownership\_RENT | 0.1327 | 0.0050 |
| 2 | loan\_percent\_income | 0.1275 | 0.0037 |
| 3 | loan\_int\_rate | 0.1160 | 0.0041 |
| 4 | person\_income | 0.0686 | 0.0022 |
| 5 | loan\_grade\_C | 0.0413 | 0.0026 |
| 6 | loan\_grade\_D | 0.0255 | 0.0014 |
| 7 | person\_home\_ownership\_OWN | 0.0155 | 0.0027 |
| 8 | loan\_intent\_VENTURE | 0.0148 | 0.0027 |
| 9 | loan\_intent\_HOMEIMPROVEMENT | 0.0108 | 0.0007 |
| 10 | person\_emp\_length | 0.0102 | 0.0010 |
| 11 | person\_home\_ownership\_MORTGAGE | 0.0097 | 0.0012 |
| 12 | loan\_intent\_DEBTCONSOLIDATION | 0.0090 | 0.0009 |
| 13 | loan\_intent\_MEDICAL | 0.0087 | 0.0010 |
| 14 | cb\_person\_default\_on\_file | 0.0036 | 0.0014 |
| 15 | person\_age | 0.0019 | 0.0005 |
| 16 | cb\_person\_cred\_hist\_length | 0.0013 | 0.0007 |
| 17 | person\_home\_ownership\_OTHER | 0.0008 | 0.0001 |
| 18 | loan\_amnt | 0.0007 | 0.0007 |
| 19 | loan\_grade\_E | 0.0003 | 0.0001 |
| 20 | loan\_grade\_G | 0.0002 | 0.0001 |
| 21 | loan\_intent\_PERSONAL | 0.0001 | 0.0002 |
| 22 | loan\_intent\_EDUCATION | 0.0000 | 0.0001 |
| 22 | loan\_grade\_A | 0.0000 | 0.0000 |
| 22 | loan\_grade\_B | 0.0000 | 0.0000 |
| 22 | loan\_grade\_F | 0.0000 | 0.0000 |

**Appendix XX: The Result of Feature Importance with Decision Tree – SHAP Importance**

|  |  |  |
| --- | --- | --- |
| **Rank** | **Feature** | **Mean ABS SHAP Value** |
| 1 | loan\_percent\_income | 0.1051 |
| 2 | loan\_int\_rate | 0.0790 |
| 3 | person\_home\_ownership\_RENT | 0.0596 |
| 4 | person\_income | 0.0434 |
| 5 | loan\_grade\_D | 0.0237 |
| 6 | loan\_grade\_C | 0.0227 |
| 7 | person\_emp\_length | 0.0201 |
| 8 | loan\_intent\_MEDICAL | 0.0200 |
| 9 | loan\_intent\_DEBTCONSOLIDATION | 0.0194 |
| 10 | loan\_intent\_HOMEIMPROVEMENT | 0.0130 |
| 11 | loan\_intent\_VENTURE | 0.0118 |
| 12 | cb\_person\_default\_on\_file | 0.0088 |
| 13 | person\_home\_ownership\_OWN | 0.0062 |
| 14 | person\_age | 0.0040 |
| 15 | person\_home\_ownership\_MORTGAGE | 0.0035 |
| 16 | loan\_amnt | 0.0033 |
| 17 | loan\_grade\_G | 0.0013 |
| 18 | cb\_person\_cred\_hist\_length | 0.0010 |
| 19 | person\_home\_ownership\_OTHER | 0.0005 |
| 20 | loan\_intent\_PERSONAL | 0.0004 |
| 21 | loan\_grade\_E | 0.0002 |
| 22 | loan\_grade\_A | 0.0000 |
| 23 | loan\_grade\_B | 0.0000 |
| 24 | loan\_grade\_F | 0.0000 |