AI PROJECT MANAGEMENT LIFECYCLE

Project 01: AI-Powered Loan Default Prediction System for Banking and Fintech

Week 02 Report

Data Processing and Feature Engineering

Dataset Source

Loan Default Dataset [Kaggle]

Prepared By:
Farhan Zarif
ID: E1-23301692-MHS60E1D

Date:
December 12, 2025

Contents

1 Introduction

This report documents the comprehensive data preprocessing and feature engineering steps performed on the Loan Default Dataset obtained from Kaggle. The primary objective of this project is to develop an AI-powered system capable of predicting loan defaults in the banking and fintech sector. Accurate prediction of loan defaults is crucial for financial institutions as it helps in risk assessment, portfolio management, and reducing potential losses.

The dataset contains information about loan applicants including their demographic details, financial attributes, credit history, and employment information. The target variable indicates whether a borrower defaulted on their loan (1) or not (0). Through systematic data preprocessing and thoughtful feature engineering, we aim to prepare this dataset for building robust machine learning models.

This report covers the following key areas: exploratory data analysis to understand the data distribution and patterns, data quality assessment including null value and duplicate handling, identification and removal of irrelevant features, and creation of new meaningful features that could improve model performance.

2 Dataset Overview

The Loan Default dataset consists of 255,347 records with 18 features describing various aspects of loan applicants and their loans. Table 1 provides a detailed description of each feature in the original dataset.

Table 1: Dataset Feature Descriptions

Feature	**Data Type**	**Description**
LoanID	Object	Unique identifier for each loan application
Age	Integer	Age of the loan applicant in years (18-70)
Income	Integer	Annual income of the applicant in currency units
LoanAmount	Integer	The amount of loan requested/approved
CreditScore	Integer	Credit score of the applicant (300-850 range)
MonthsEmployed	Integer	Number of months the applicant has been employed
NumCreditLines	Integer	Number of active credit lines the applicant has
InterestRate	Float	Interest rate applied to the loan (percentage)
LoanTerm	Integer	Duration of the loan in months
DTIRatio	Float	Debt-to-Income ratio of the applicant
Education	Object	Education level (High School, Bachelor's, Master's, PhD)
EmploymentType	Object	Type of employment (Full-time, Part-time, Self-employed, Unemployed)
MaritalStatus	Object	Marital status (Single, Married, Divorced)

Feature	Data Type	Description
HasMortgage	Object	Whether applicant has a mortgage (Yes/No)
HasDependents	Object	Whether applicant has dependents (Yes/No)
LoanPurpose	Object	Purpose of the loan (Home, Auto, Education, Business, Other)
HasCoSigner	Object	Whether loan has a co-signer (Yes/No)
Default	Integer	Target variable - 1 if defaulted, 0 otherwise

3 Exploratory Data Analysis

3.1 Statistical Summary

The exploratory data analysis began with examining the statistical properties of numerical features. Table 2 presents the descriptive statistics for all numerical variables in the dataset.

Table 2: Statistical Summary of Numerical Features

Feature	Min	Max	Mean	Std	Skew
Age	18	70	43.99	15.26	0.00
Income	15,000	149,999	82,473	39,064	0.00
LoanAmount	5,000	234,999	119,928	66,510	0.00
CreditScore	300	850	575.00	158.85	0.00
MonthsEmployed	0	120	59.96	34.66	0.00
NumCreditLines	1	4	2.50	1.12	0.00
InterestRate	2.00	24.99	13.50	6.64	0.00
LoanTerm	12	60	36.00	17.42	0.00
DTIRatio	0.10	0.90	0.50	0.23	0.00

The statistical analysis reveals that the numerical features are well-distributed with minimal skewness, indicating that the data follows approximately uniform distributions. The features span reasonable ranges for their respective domains, such as credit scores ranging from 300 to 850 and ages from 18 to 70.

3.2 Target Variable Distribution

Understanding the distribution of the target variable is essential for classification problems. The analysis of the Default variable reveals the following distribution:

Table 3: Target Variable Distribution

Class	Count	Percentage
Non-Default (0)	225,694	88.39%
Default (1)	29,653	11.61%
Total	**255,347**	**100.00%**

The target variable exhibits class imbalance, with approximately 88.39% of loans being non-defaults and 11.61% being defaults. This imbalance is typical in real-world loan default scenarios and will need to be addressed during model training through techniques such as oversampling (SMOTE), undersampling, or using class weights.

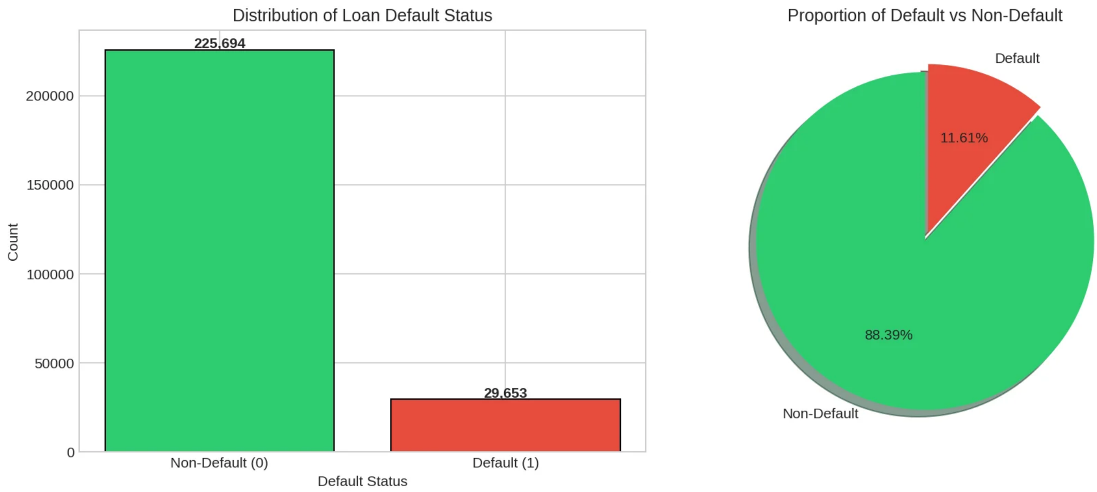

Figure 1: Distribution of Loan Default Status showing class imbalance

3.3 Numerical Features Distribution

Figure 2 shows the distribution of all numerical features, separated by default status. The visualizations reveal interesting patterns in how defaults and non-defaults are distributed across different feature ranges.

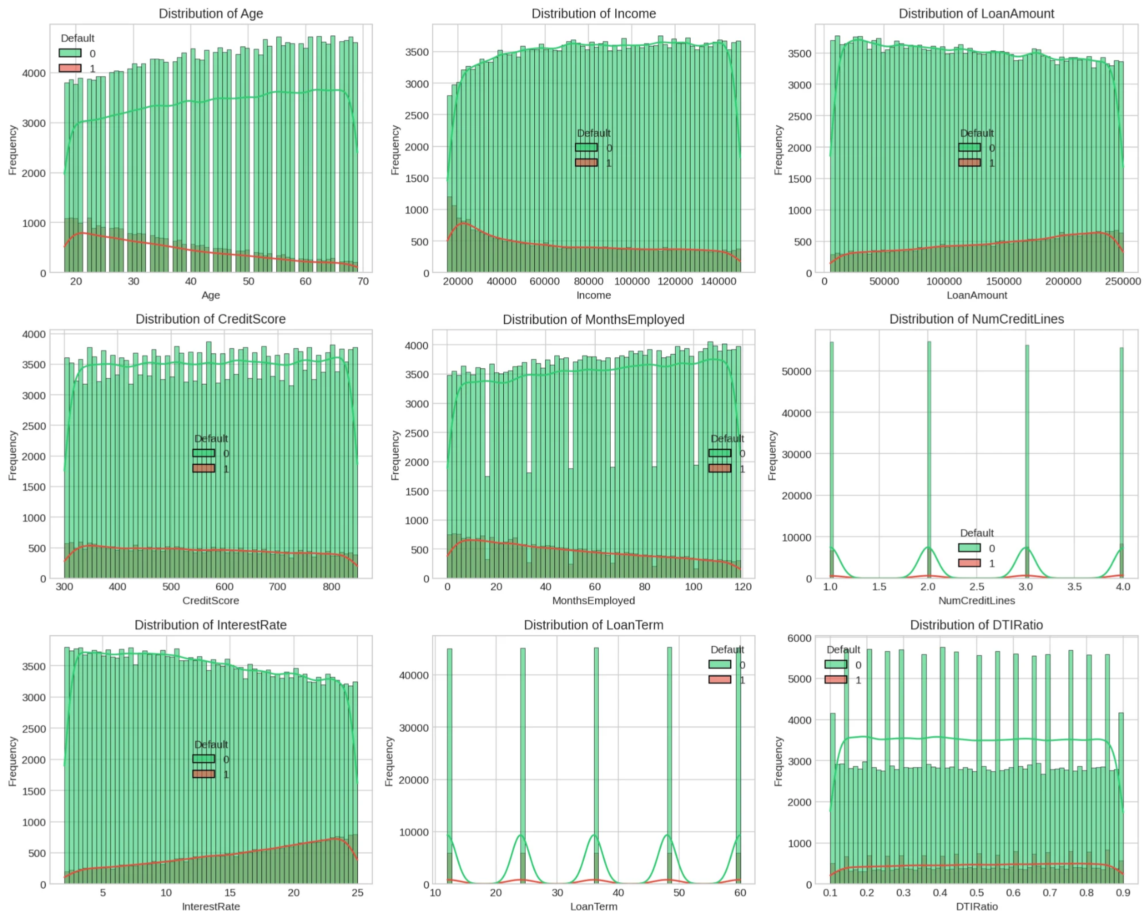

Figure 2: Distribution of Numerical Features by Default Status

Key observations from the numerical distributions include the fact that Age shows higher default rates among younger applicants, Income demonstrates that lower income applicants have higher default tendencies, InterestRate exhibits a positive relationship with defaults where higher rates correlate with more defaults, and MonthsEmployed reveals that less employment tenure correlates with higher default risk.

3.4 Categorical Features Analysis

The dataset contains seven categorical features. Figure 3 shows the default rates across different categories for each categorical feature.

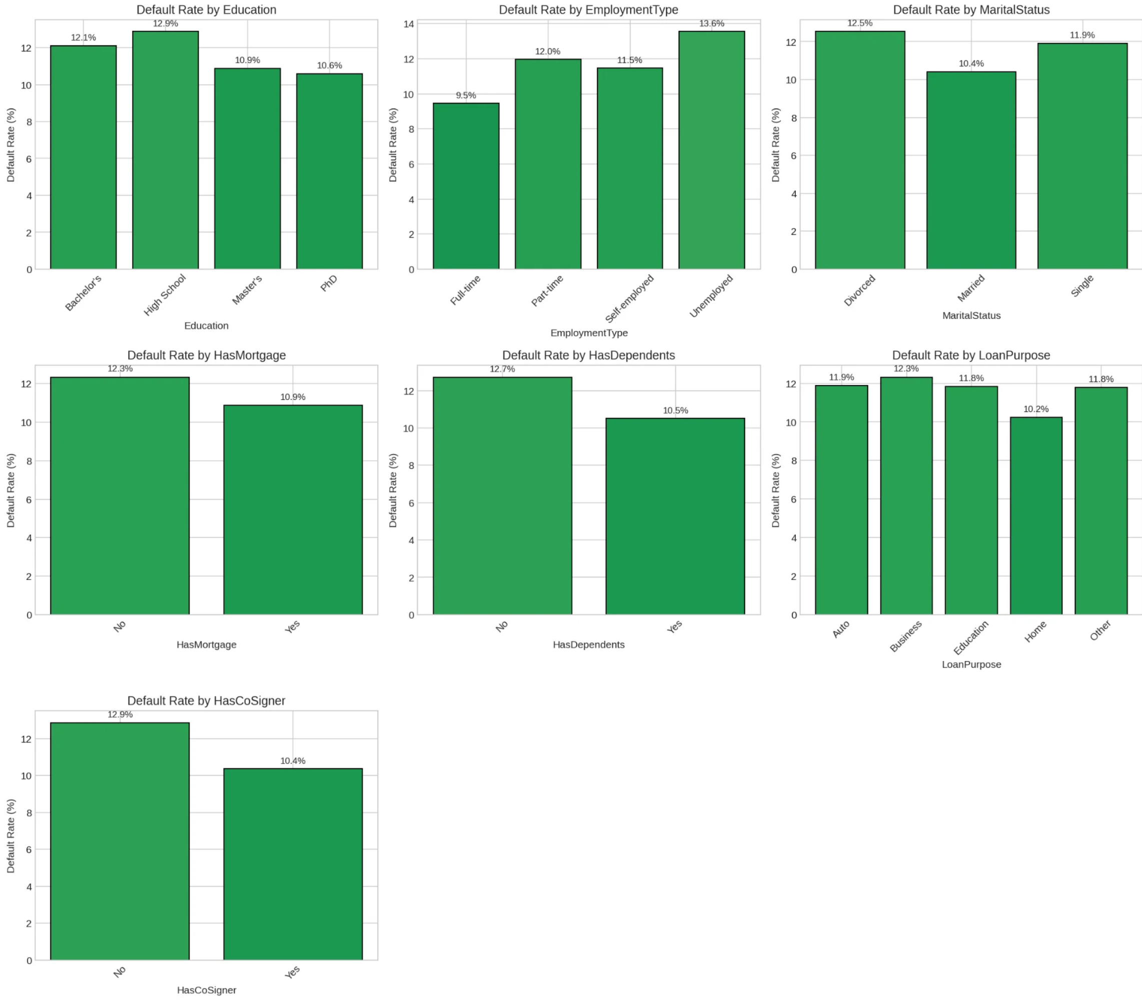

Figure 3: Default Rates by Categorical Features

Table 4 summarizes the key findings from categorical feature analysis.

Table 4: Categorical Features - Key Default Rate Findings

Feature	Highest Default Category	Rate (%)
Education	High School	12.9
EmploymentType	Unemployed	13.6
MaritalStatus	Divorced	12.5
HasMortgage	No	12.3
HasDependents	No	12.7
LoanPurpose	Business	12.3
HasCoSigner	No	12.9

The analysis reveals meaningful patterns in default rates: unemployed individuals have the highest default rate (13.6%), those without co-signers default more frequently (12.9%), High School educated applicants show higher default rates (12.9%), and divorced individuals have elevated default risk (12.5%).

3.5 Correlation Analysis

The correlation analysis examines the linear relationships between numerical features and the target variable. Understanding these correlations helps identify potentially predictive features and detect multicollinearity issues.

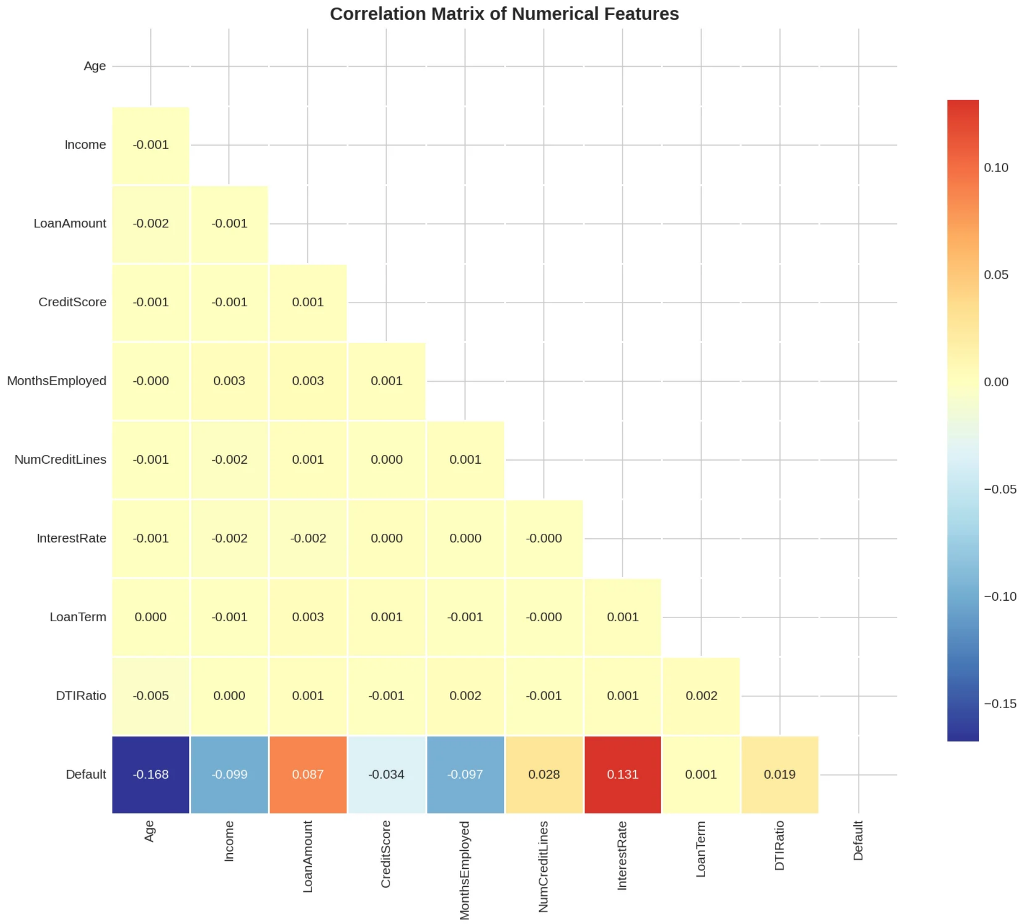

Figure 4: Correlation Matrix of Numerical Features with Default

Table 5: Correlation with Target Variable (Default)

Feature	**Correlation**	**Interpretation**
Age	-0.168	Older applicants less likely to default
InterestRate	+0.131	Higher rates correlate with defaults
Income	-0.099	Higher income reduces default risk
MonthsEmployed	-0.097	Longer employment reduces default
LoanAmount	+0.087	Larger loans slightly increase risk
CreditScore	-0.034	Better scores reduce default risk
NumCreditLines	+0.028	Minimal positive correlation
DTIRatio	+0.019	Higher DTI slightly increases risk
LoanTerm	+0.001	Negligible correlation

The correlation analysis reveals that Age has the strongest negative correlation (-0.168) with default, suggesting younger applicants are more likely to default. InterestRate shows the strongest positive correlation (+0.131), indicating that higher interest rates are associated with higher default risk. Income and MonthsEmployed also show meaningful negative correlations with default.

4 Data Quality Assessment

4.1 Null Value Analysis

A comprehensive examination of missing values was conducted across all features. Data quality is paramount for building reliable machine learning models, and missing values can significantly impact model performance if not properly handled.

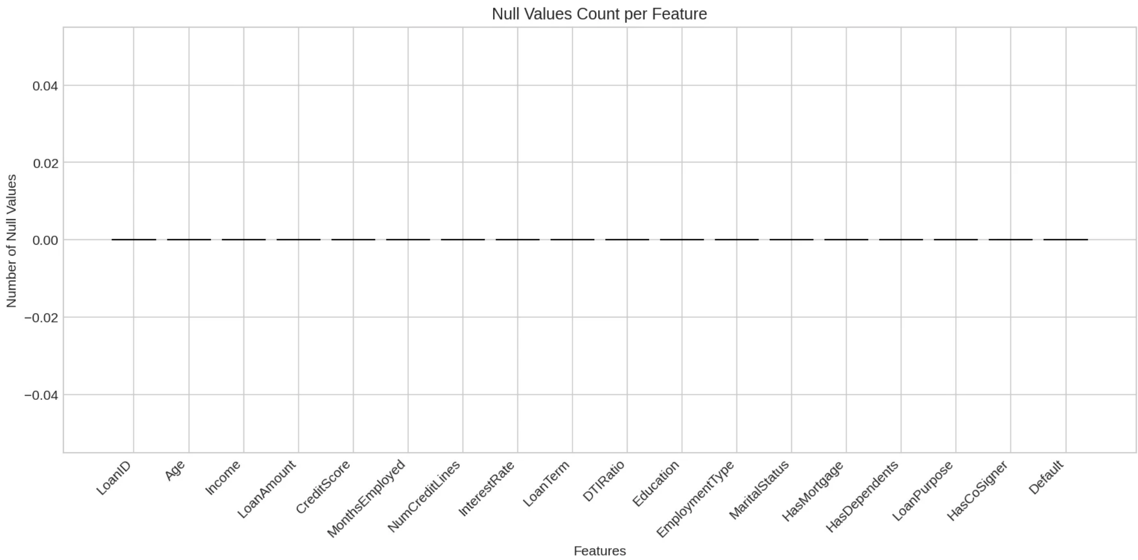

Figure 5: Null Values Count per Feature

As shown in Figure 5, the analysis confirms that the dataset contains no missing values across any of the 18 features. This eliminates the need for imputation strategies and ensures that all records can be used for model training without data loss.

Finding: The dataset has **zero null values**. No imputation or missing value handling is required.

4.2 Duplicate Detection and Removal

Duplicate records can bias model training and lead to data leakage. Three types of duplicate checks were performed: exact duplicates (records where all column values are identical), LoanID duplicates (records with the same loan identifier), and content duplicates (records with identical content but different LoanIDs).

Table 6: Duplicate Analysis Results

Duplicate Type	Count
Exact Duplicates	0
LoanID Duplicates	0
Content Duplicates (excluding LoanID)	0

Finding: The dataset contains **no duplicate records**. All 255,347 records are unique and can be retained for analysis.

4.3 Irrelevant Feature Identification

Identifying and removing irrelevant features is crucial for model efficiency and interpretability. Features were evaluated based on the following criteria: identifier columns (features that serve only as unique identifiers and have no predictive value), constant columns (features with zero variance), and near-constant columns (features where more than 99% of values are the same).

Table 7: Irrelevant Feature Analysis

Feature	Reason for Removal	Action
LoanID	Unique identifier with no predictive value (255,347 unique values)	**DROP**

Decision: The `LoanID` column was identified as the only irrelevant feature and removed from the dataset. This column serves purely as a record identifier and has 255,347 unique values (one per record), making it unsuitable for prediction.

After removal: Dataset shape changed from 255,347 $\times$ 18 to 255,347 $\times$ 17 columns.

5 Feature Engineering

Feature engineering is the process of creating new features from existing ones to capture additional patterns and improve model performance. Based on domain knowledge in lending and credit risk assessment, the following new features were engineered.

5.1 Financial Ratio Features

5.1.1 Loan-to-Income Ratio (LTI)

The Loan-to-Income ratio measures the loan amount relative to the applicant's annual income. Higher values indicate greater financial burden.

$$\text{LoanToIncomeRatio} = \frac{\text{LoanAmount}}{\text{Income}} \tag{1}$$

Rationale: This ratio is a fundamental metric in credit risk assessment. Borrowers with high LTI ratios may struggle to repay their loans, increasing default risk.

5.1.2 Monthly Payment Estimate

An estimated monthly payment was calculated based on loan terms:

$$\text{MonthlyPaymentEstimate} = \frac{\text{LoanAmount} \times (1 + \frac{\text{InterestRate}}{100})}{\text{LoanTerm}} \tag{2}$$

Rationale: The monthly payment obligation directly affects a borrower's ability to meet payment schedules. Higher monthly payments relative to income increase default probability.

5.1.3 Payment-to-Income Ratio

This ratio measures monthly payment as a percentage of monthly income:

$$\text{PaymentToIncomeRatio} = \frac{\text{MonthlyPaymentEstimate}}{\text{Income}/12} \times 100 \tag{3}$$

Rationale: This metric captures the proportion of monthly income dedicated to loan repayment, providing insight into financial stress levels.

5.2 Categorical Binning Features

5.2.1 Credit Score Category

Credit scores were binned into interpretable categories based on standard credit rating classifications:

Table 8: Credit Score Categorization

Score Range	Category	Default Rate (%)
< 580	Poor	12.5
580 – 669	Fair	11.4
670 – 739	Good	10.6
740 – 799	Very Good	10.5
≥ 800	Excellent	9.8

Rationale: Categorical credit scores capture non-linear relationships and align with industry-standard credit tier classifications used by lenders.

5.2.2 Age Group

Age was categorized into meaningful demographic groups:

Table 9: Age Group Categorization

Age Range	Category	Default Rate (%)
< 25	Young	21.0
25 – 34	Young Adult	16.6
35 – 49	Middle Aged	11.1
50 – 64	Senior	6.7
≥ 65	Elderly	4.8

Rationale: Different age groups have varying financial stability levels and risk profiles. The data clearly shows younger applicants have significantly higher default rates (21.0% for Young vs 4.8% for Elderly).

5.2.3 Employment Stability

Months of employment was transformed into stability categories:

Table 10: Employment Stability Categorization

Months Employed	Category	Default Rate (%)
0	Unemployed	18.1
1 – 11	New	16.9
12 – 35	Moderate	15.0
36 – 59	Stable	12.2
≥ 60	Very Stable	9.0

Rationale: Employment tenure is a strong indicator of income stability and job security. The data shows a clear trend: unemployed applicants have a default rate of 18.1% compared to only 9.0% for very stable employees.

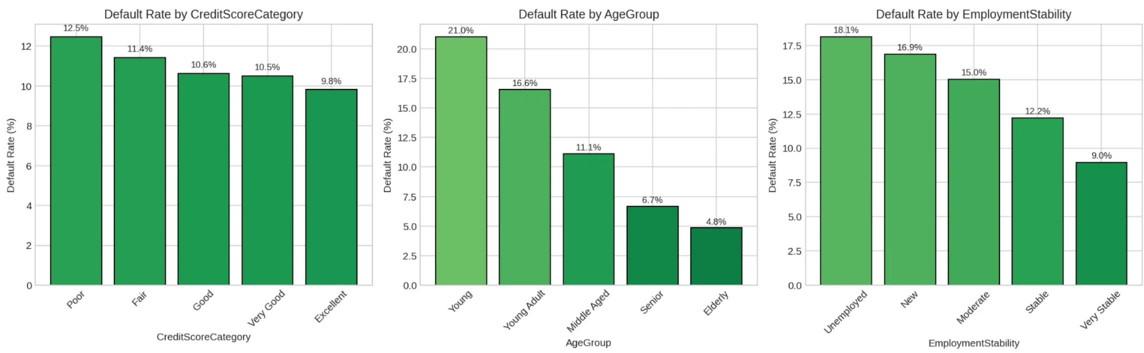

Figure 6: Default Rates by Engineered Categorical Features

Figure 6 demonstrates the effectiveness of the engineered categorical features. All three binning features (CreditScoreCategory, AgeGroup, EmploymentStability) show clear monotonic relationships with default rates, confirming their predictive value.

5.3 Composite Risk Score

A composite risk score was created by normalizing and combining multiple risk factors:

$$\text{RiskScore} = \text{mean}\left(\text{Normalize}(\text{DTIRatio}), \text{Normalize}(\text{InterestRate}), \text{Normalize}(\text{LTI})\right) \times 100 \tag{4}$$

This score ranges from 0 to 100, where higher values indicate higher risk profiles.

Rationale: A composite score combines multiple risk indicators into a single metric, simplifying risk assessment and potentially capturing interaction effects.

5.4 Binary Indicator Features

5.4.1 Has Multiple Credit Lines

A binary indicator was created for applicants with more than one credit line:

$$\text{HasMultipleCreditLines} = \begin{cases} 1 & \text{if NumCreditLines} > 1 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

Rationale: Multiple credit lines can indicate both credit experience (positive) or over-leverage (negative), making it a potentially informative feature.

5.5 Summary of Engineered Features

Table 11 summarizes all newly created features:

Table 11: Summary of Engineered Features

Feature Name	Type	Description
LoanToIncomeRatio	Continuous	Loan amount relative to annual income
MonthlyPaymentEstimate	Continuous	Estimated monthly loan payment
PaymentToIncomeRatio	Continuous	Monthly payment as % of monthly income
CreditScoreCategory	Categorical	Credit score tier (Poor to Excellent)
AgeGroup	Categorical	Age-based demographic grouping
EmploymentStability	Categorical	Employment tenure classification
RiskScore	Continuous	Composite risk indicator (0-100)
HasMultipleCreditLines	Binary	Multiple credit lines indicator

6 Additional Data Operations

6.1 Categorical Variable Encoding

Machine learning algorithms typically require numerical inputs. Therefore, categorical variables were encoded using appropriate strategies based on their nature.

6.1.1 Binary Encoding

Binary categorical variables (Yes/No) were converted to numerical format:

Table 12: Binary Variable Encoding

Feature	Encoding
HasMortgage	Yes → 1, No → 0
HasDependents	Yes → 1, No → 0
HasCoSigner	Yes → 1, No → 0

6.1.2 Ordinal Encoding

Variables with inherent ordering were encoded to preserve their natural hierarchy:

Table 13: Ordinal Variable Encoding

Feature	Encoding Order
Education	High School (0) < Bachelor's (1) < Master's (2) < PhD (3)
CreditScoreCategory	Poor (0) < Fair (1) < Good (2) < Very Good (3) < Excellent (4)
EmploymentStability	Unemployed (0) < New (1) < Moderate (2) < Stable (3) < Very Stable (4)
AgeGroup	Young (0) < Young Adult (1) < Middle Aged (2) < Senior (3) < Elderly (4)

6.1.3 One-Hot Encoding

Nominal categorical variables (without inherent order) were one-hot encoded. Note that one category from each feature was dropped (drop_first=True) to avoid multicollinearity in linear models.

Table 14: One-Hot Encoded Variables

Original Feature	Generated Dummy Columns
EmploymentType	EmploymentType_Part-time, EmploymentType_Self-employed, EmploymentType_Unemployed
MaritalStatus	MaritalStatus_Married, MaritalStatus_Single
LoanPurpose	LoanPurpose_Business, LoanPurpose_Education, LoanPurpose_Home, LoanPurpose_Other

6.2 Final Correlation Analysis

After feature engineering and encoding, the final correlation matrix was computed to examine relationships between all processed features and the target variable.

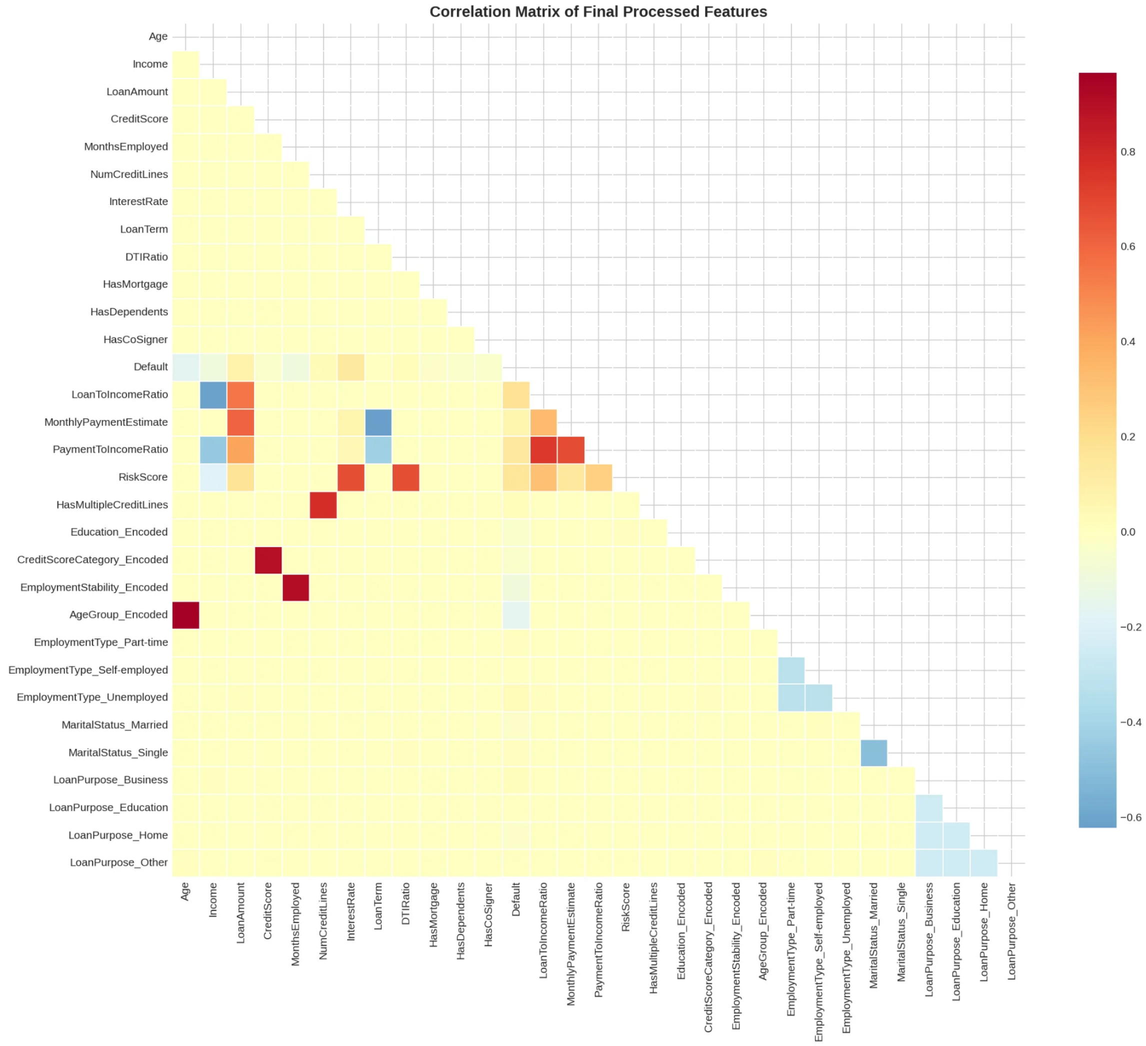

Figure 7: Correlation Matrix of Final Processed Features

The final correlation matrix (Figure 7) reveals several important patterns. The engineered features show stronger correlations with the target variable than many original features. The encoded categorical features (CreditScoreCategory_Encoded, EmploymentStability_Encoded, AgeGroup_Encoded) show meaningful negative correlations with default, confirming their predictive value. Some engineered features like PaymentToIncomeRatio and RiskScore show moderate positive correlations with related original features, as expected.

6.3 Final Dataset Structure

After preprocessing and feature engineering, the final processed dataset contains 28 features organized as follows:

Original Numerical Features (9): Age, Income, LoanAmount, CreditScore, MonthsEmployed, NumCreditLines, InterestRate, LoanTerm, DTIRatio

Encoded Binary Features (3): HasMortgage, HasDependents, HasCoSigner

Engineered Numerical Features (5): LoanToIncomeRatio, MonthlyPaymentEstimate, PaymentToIncomeRatio, RiskScore, HasMultipleCreditLines

Ordinal Encoded Features (4): Education_Encoded, CreditScoreCategory_Encoded, EmploymentStability_Encoded, AgeGroup_Encoded

One-Hot Encoded Features (6): EmploymentType dummies (3), MaritalStatus dummies (2), LoanPurpose dummies (4), minus reference categories

Target Variable (1): Default

7 Summary and Conclusions

This report presented a comprehensive data preprocessing and feature engineering pipeline for the Loan Default dataset. The key findings and actions are summarized below.

7.1 Data Quality Summary

Table 15: Data Quality Summary

Quality Aspect	Finding	Action Taken
Missing Values	No null values found	None required
Duplicate Records	No duplicates detected	None required
Irrelevant Features	LoanID identified	Removed from dataset
Data Types	Appropriate types assigned	No conversion needed

7.2 Key Insights from Analysis

The exploratory data analysis revealed several important patterns that will inform model development:

1. **Class Imbalance:** The target variable shows significant imbalance (88.39% non-default vs 11.61% default), requiring specialized handling during model training.
2. **Age is Highly Predictive:** Younger applicants show dramatically higher default rates (21.0% for under-25 vs 4.8% for 65+), making age-based features particularly valuable.
3. **Employment Stability Matters:** Default rates decrease monotonically with employment tenure (18.1% for unemployed vs 9.0% for very stable).
4. **Interest Rate Correlation:** Higher interest rates correlate with higher default rates (+0.131 correlation), suggesting risk-based pricing in the original data.

7.3 Feature Engineering Summary

Eight new features were engineered based on domain knowledge in credit risk assessment. The engineered categorical features (CreditScoreCategory, AgeGroup, EmploymentStability) show clear monotonic relationships with default rates, confirming their predictive value. The financial ratio features (LoanToIncomeRatio, PaymentToIncomeRatio) capture important debt burden metrics.

7.4 Final Dataset Specifications

Table 16: Final Dataset Specifications

Specification	Value
Total Records	255,347
Original Features	18
Features After Preprocessing	28
Features Dropped	1 (LoanID)
New Features Created	8
Target Variable Imbalance	88.39% : 11.61%

7.5 Files Generated

The following files were generated as part of this preprocessing pipeline:

1. `Loan_default_processed.csv` – Final encoded dataset ready for model training
2. `Loan_default_engineered.csv` – Dataset with categorical labels preserved for interpretability
3. Visualization files (.png) for exploratory analysis documentation

7.6 Recommendations for Next Steps

Based on the preprocessing analysis, the following recommendations are made for subsequent model development:

1. **Address Class Imbalance:** Implement techniques such as SMOTE, random undersampling, or class weights during model training given the 88:12 imbalance ratio.
2. **Feature Selection:** Consider using feature importance from tree-based models to select the most predictive features. The engineered features (particularly AgeGroup, EmploymentStability) show strong potential.
3. **Model Selection:** Given the tabular nature of the data and meaningful feature correlations, ensemble methods (Random Forest, XGBoost, LightGBM) are recommended.
4. **Cross-Validation:** Use stratified k-fold cross-validation to ensure reliable model evaluation given the class imbalance.

References

1. Kaggle. (n.d.). Loan Default Dataset. Retrieved from Kaggle Datasets.
2. McKinney, W. (2017). Python for Data Analysis. O'Reilly Media.

3. Scikit-learn Documentation. (n.d.). Preprocessing Data. Retrieved from scikit-learn.org.

4. Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media.