Credit Risk Assessment Using Machine Learning

Table of Contents

1. Background

Financial institutions face significant challenges in assessing credit risk and predicting loan defaults. With the increasing availability of data and advanced machine learning techniques, there's an opportunity to develop more accurate risk assessment models. This project focuses on developing a predictive model for loan default risk using various borrower characteristics and loan attributes.

2. Business and Technical challenges

Financial institutions need an automated, accurate system to:

- Predict likelihood of loan default
- Identify high-risk applications early
- Understand key risk factors
- Standardize loan approval decisions

We need to Develop a machine learning model that can:

- Process multiple borrower characteristics
- Handle both numerical and categorical data
- Provide interpretable results
- Maintain high accuracy across different borrower segments

3. Project Objectives

- Develop a predictive model to identify potential loan defaults
- Identify key factors that influence loan default risk
- Create a reliable risk assessment tool for loan approval decisions
- Compare different machine learning models for optimal performance

4. Data Exploration and Preprocessing

4.1 Dataset Overview

Initial dataset: 32,581 records with 12 features

Final dataset after cleaning: 32,416 records

4.2 Feature Description

Numerical Features:

- person_age: Borrower's age
- person_income: Annual income
- person_emp_length: Employment length in years
- loan_amnt: Loan amount requested
- loan_int_rate: Interest rate
- loan_percent_income: Loan amount as percentage of income
- cb_person_cred_hist_length: Credit history length

Categorical Features:

- person_home_ownership: RENT, MORTGAGE, OWN, OTHER
- loan_intent: PERSONAL, EDUCATION, MEDICAL, VENTURE, HOMEIMPROVEMENT, DEBTCONSOLIDATION
- loan_grade: A through G
- cb_person_default_on_file: Y/N
- loan_status: 0 (non-default), 1 (default)

4.3 Data Quality Assessment

Missing Values:

- person_emp_length: 887 records (2.7%)
- loan_int_rate: 3,095 records (9.5%)

Duplicates:

- 165 duplicate records identified and removed

4.4 Data Cleaning Steps

1. Duplicate Removal

   - Identified using full record comparison
   - Removed to prevent model bias

2. Missing Value Treatment

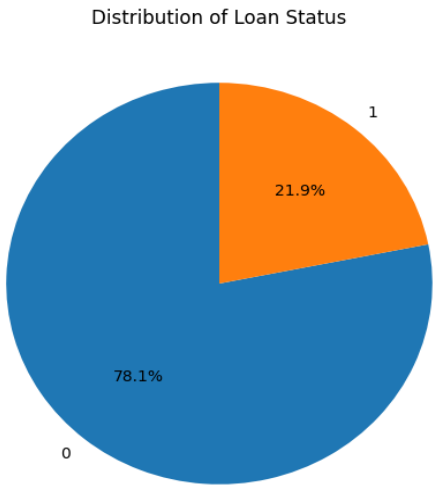
   - Median imputation for numerical features
   - Chosen over mean to minimize outlier impact
   - Verified distribution preservation after imputation

3. Categorical Encoding

   - Label encoding applied to categorical variables
   - Maintained ordinal relationships where applicable
   - Preserved categorical value mapping for deployment

5. Exploratory Data Analysis and Visualization

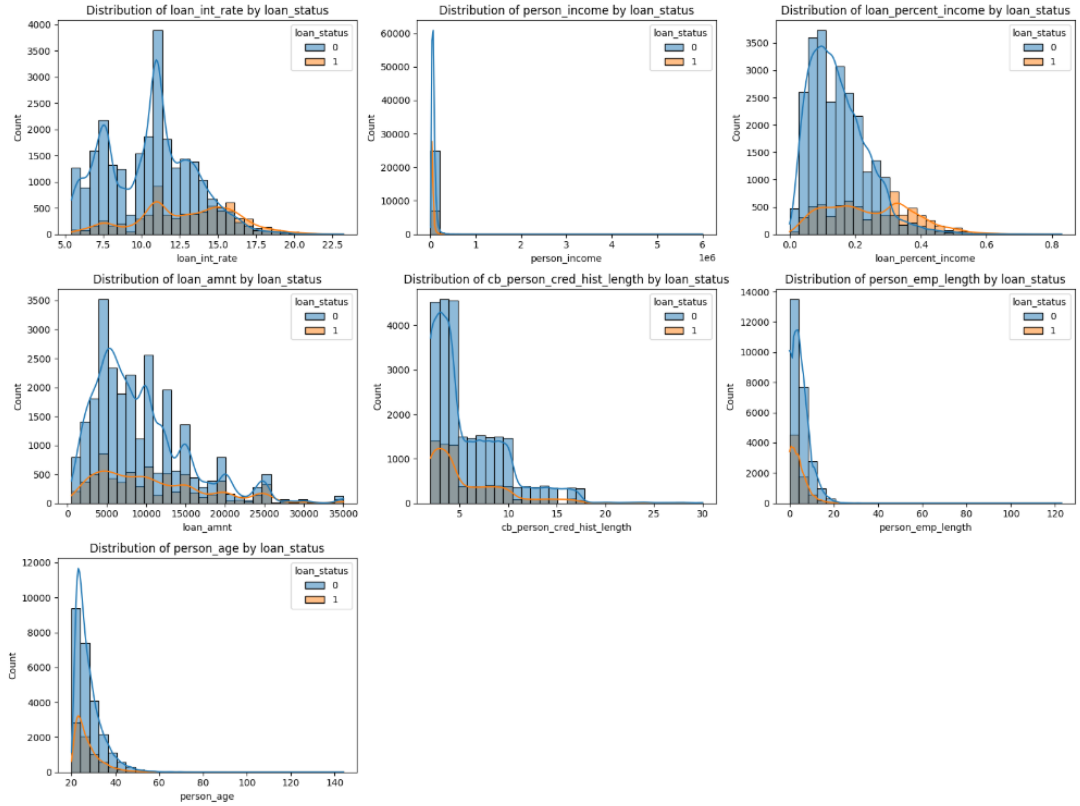
Our exploratory data analysis utilized various visualization techniques to uncover patterns and relationships within the loan data. Each visualization provided unique insights into different aspects of loan default behavior.



6.1 Loan Status Distribution (Pie Chart Analysis) We began our analysis with a fundamental pie chart visualization of loan outcomes. The chart revealed that among our dataset of 32,416 loans:

- 78.2% of loans were successfully repaid (non-default)
- 21.8% resulted in defaults

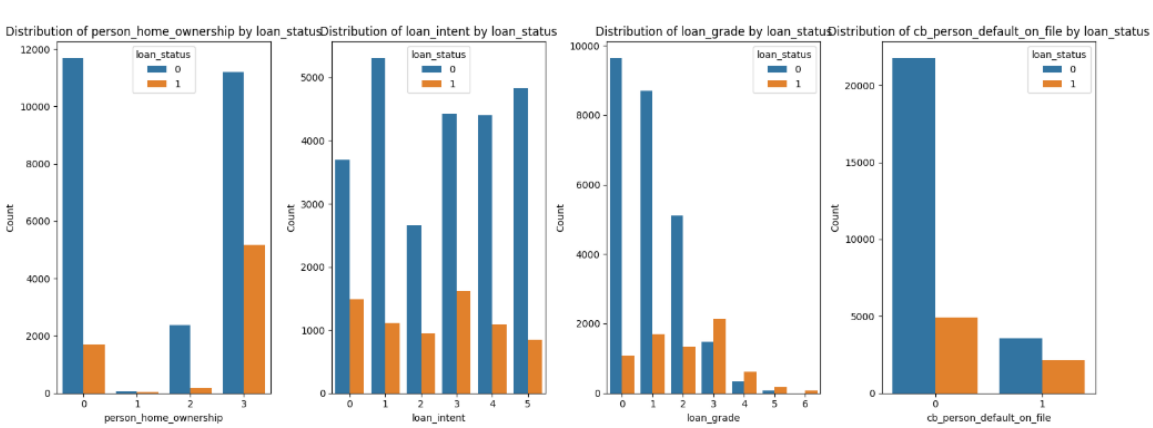
5.2 Numerical Features Distribution by Loan Status (Histograms) We created a comprehensive set of histograms for each numerical feature, separated by loan status. These visualizations revealed several critical patterns:



- Lower income, higher loan amount, and shorter credit history seem to be associated with a higher likelihood of default.
- There's a potential trend of higher default rates for younger borrowers with shorter employment history.
- The bimodality in the loan interest rate distribution might be due to different loan types or risk categories.

5.3 Categorical Variable Impact

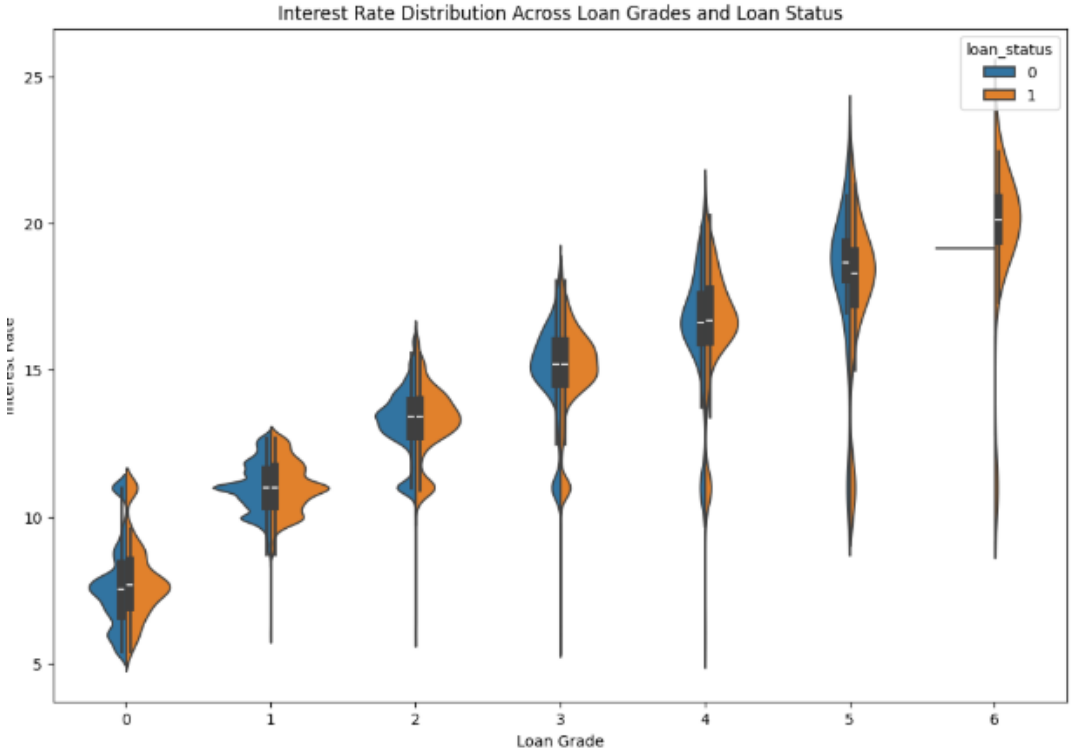
These visualizations provide insights into the distribution of categorical features across different loan statuses. The patterns observed can help identify potential risk factors and inform credit decision-making processes.



- Borrowers with no prior defaults (cb_person_default_on_file = 0) and higher loan grades are less likely to default.
- There's a potential association between renting and defaulting on loans.
- Debt consolidation as a loan intent might be associated with a slightly higher risk of default.

5.4 Relationship Between Interest Rates, Loan Grades, and Loan Default Risk

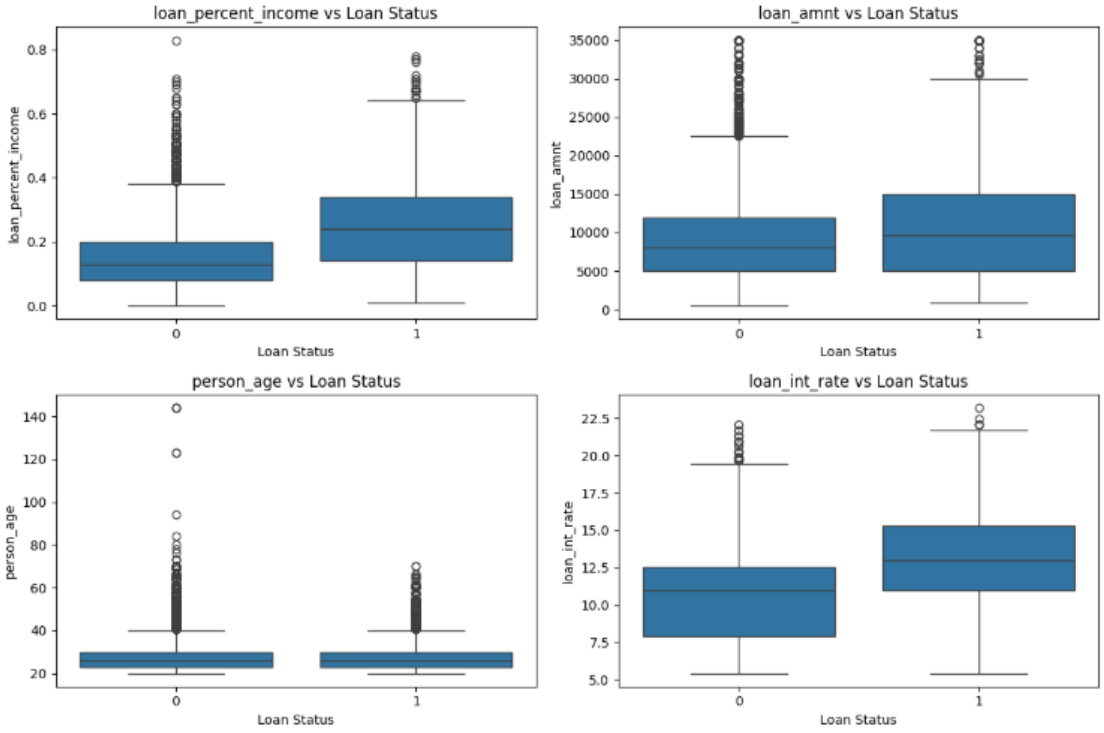
This violin plot shows the distribution of interest rates across different loan grades and loan statuses.



Higher loan grades are associated with higher interest rates. Default loans tend to have a higher concentration in the higher interest rate regions. This suggests a correlation between interest rates and default risk.

5.5 Numerical Features and Loan Default Risk

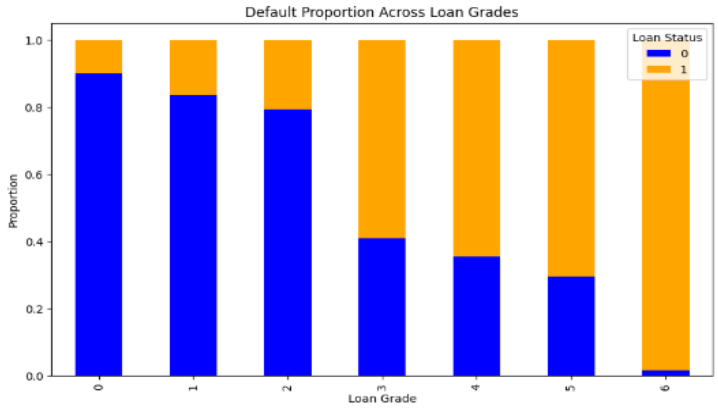
These box plots show the distribution of numerical features by loan status.



Higher loan amounts, higher loan percent income, and higher interest rates are associated with increased default risk. Age doesn't seem to be a strong predictor.

5.6 Default Proportion Across Loan Grades

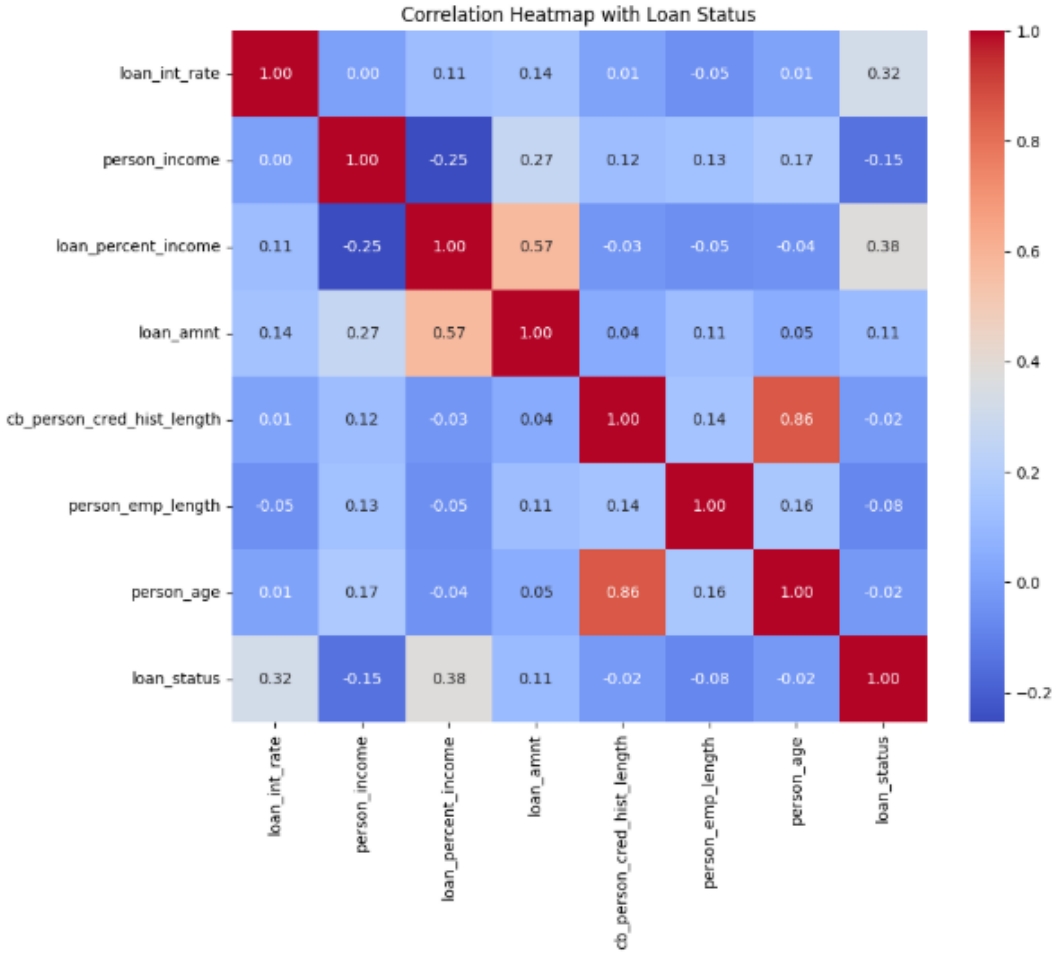
This stacked bar chart shows the proportion of default and non-default loans for each loan grade.



Higher loan grades have a significantly higher proportion of defaults, indicating that loan grade is a strong predictor of default risk.

5.7 Correlation Heatmap

This heatmap visualizes the correlation between numerical features and loan status.



The strongest correlations with loan status are observed for loan interest rate and loan percent income. These features might be significant predictors of default risk.

6. Feature Analysis and Selection for Loan Approval

6.1 Features Created:

Several features were created to better understand and predict loan approval outcomes:

1. **Effective Interest Burden**: (loan_int_rate * loan_amnt) / person_income
   - Captures the financial burden of the loan on borrowers.
2. **Grade Income Ratio**: loan_grade * loan_percent_income
   - Combines risk grade with income commitment, addressing the relationship between grade and income.
3. **Monthly Payment Ratio**: loan_amnt / (person_income / 12)
   - Provides a monthly perspective on the borrower's payment burden.

4. **Debt-to-Income**: loan_amnt / person_income
    - Represents the ratio of loan amount to total income, indicating financial sustainability.
5. **Credit Years per Age**: cb_person_cred_hist_length / person_age
    - Measures relative credit experience based on age.
6. **Employment Ratio**: person_emp_length / person_age
    - Assesses employment stability relative to age.
7. **Stability Score**: (person_emp_length * person_income) / (loan_amnt * loan_percent_income)
    - Combines key stability indicators for a holistic view of financial stability.

6.2 Feature Selection

Selected features based on:

- Correlation with target > 0.4
- Business significance
- Low multicollinearity
- Predictive power in initial models

7. Methodology

7.1 Data Preparation

To prepare the data for analysis, we conducted a train-test split to ensure a robust evaluation of the models. The training dataset comprises 80% of the data, amounting to 25,932 records, while the remaining 20%, or 6,484 records, was used for testing. This stratified split ensures that the distribution of the target variable, loan_status, is consistent across both sets.

Additionally, feature scaling was applied using the StandardScaler to standardize the features. This process guarantees that all features are on a comparable scale, which improves the model's convergence and enhances prediction accuracy.

7.2 Model Selection Approach

Three popular machine learning algorithms were tested:

1. **Logistic Regression**
    - This served as a baseline model, offering interpretability of features and fast training times.
    - It is effective in providing insights into the relationship between features and the target variable.

2. **Random Forest**

- Random Forest is adept at handling non-linear relationships and provides valuable insights into feature importance.
- Its robustness to outliers and ability to manage complex interactions between variables make it a strong choice for predictive tasks.

3. **XGBoost**

- XGBoost utilizes a gradient boosting approach and excels in handling imbalanced data.
- It has demonstrated high performance in many machine learning challenges, particularly when rapid predictions are required.

8. Results and Analysis

8.1 Model Performance Comparison

Random Forest

- **Overall Accuracy**: 89%
- **Class-wise Performance**:
  - Non-default: 91% precision, 95% recall
  - Default: 79% precision, 66% recall

Random Forest provides the best balance between precision and recall, which is essential for both non-default and default prediction. Its feature importance insights further enhance its utility in understanding key drivers of loan default.

XGBoost

- **Overall Accuracy**: 88%
- **Class-wise Performance**:
  - Non-default: 90% precision, 95% recall
  - Default: 77% precision, 64% recall

XGBoost delivers strong performance, particularly in handling imbalanced datasets, but slightly lags behind Random Forest in overall accuracy and balance between classes.

Logistic Regression

- **Overall Accuracy**: 84%
- **Class-wise Performance**:
  - Non-default: 86% precision, 95% recall
  - Default: 69% precision, 45% recall

Although Logistic Regression offers interpretability and simplicity, its overall performance and recall rates for defaults are lower compared to Random Forest and XGBoost.

After evaluating multiple machine learning models, we selected Random Forest as our primary model based on its superior performance metrics:

Random Forest Performance:

- Highest overall accuracy at 89%
- Superior handling of non-linear relationships
- Better balance between precision and recall
- Robust feature importance capabilities

8.2 Feature Importance Analysis

Using the trained Random Forest model, we conducted a detailed analysis to identify the most influential features in predicting loan defaults. The findings reveal several key predictors:

- **Loan Percent Income** emerged as the most important factor, with an importance score of 0.145. This highlights the significance of debt-to-income ratios in assessing default risk.
- **Effective Interest Burden** (0.141) plays a crucial role in evaluating the financial strain on borrowers and its relationship to default probabilities.
- **Personal Income** (0.138) directly reflects financial stability and inversely correlates with default risk, emphasizing repayment capacity.
- **Grade Income Ratio** (0.135) is significant, as it provides insights into borrower classification and risk levels.
- **Loan Interest Rate** (0.132) further emphasizes how pricing risk impacts loan outcomes, with higher rates linked to increased default probabilities.

These insights provide a deeper understanding of the factors contributing to default risk and guide more precise risk assessment strategies.

9. Conclusions and Recommendations

9.1 Key Findings

Random Forest has proven to be the most effective model for predicting loan defaults, offering a strong balance between accuracy and interpretability. It performs exceptionally well in both non-default and default cases, with a particular strength in handling non-default predictions. However, there is still room for improvement in accurately forecasting defaults, highlighting the need for ongoing refinement.

Loan-related features, particularly income and debt ratios, have been identified as key factors in assessing default risk. These insights emphasize the importance of maintaining manageable debt levels in relation to income and understanding borrowers' long-term financial stability. By focusing on these aspects, institutions can better predict and manage risk.

9.2 Implementation Recommendations

To optimize performance, it is recommended to use Random Forest as the primary model for loan default prediction. By incorporating probability scores, institutions can implement tiered decision-making, allowing for more nuanced risk assessment. Regular updates and retraining will ensure the model remains accurate and adaptable over time.

A strong emphasis on income verification and managing loan burden relative to income will further enhance the model’s effectiveness. Additionally, developing products tailored to specific risk segments and adjusting pricing strategies accordingly will help meet the diverse needs of borrowers. An early warning system for potential defaults can also aid institutions in proactively managing risk.

9.3 Future Improvements

To continuously improve Random Forest’s performance, collecting additional behavioral data can provide deeper insights into borrower behaviours and risk factors. Exploring ensemble methods and implementing real-time monitoring will further enhance the model’s predictive capabilities.

Automation of data collection and streamlining verification processes can reduce manual effort, increasing overall efficiency. API-based decision-making can facilitate faster, automated outcomes.

For comprehensive risk management, implementing portfolio-level monitoring and developing risk-based pricing models are essential. Additionally, early intervention strategies will effectively mitigate risks associated with loan defaults, ensuring better risk management and borrower support.

10. References

- PYTHON. (2019, May 29). Python. Python.org; Python.org. https://www.python.org/
- Scikit-learn. (n.d.). scikit-learn: Machine Learning in Python. Scikit-Learn.org. https://scikit-learn.org/stable/
- Loan Approval Prediction using Machine Learning. (2022, September 23). GeeksforGeeks. https://www.geeksforgeeks.org/loan-approval-prediction-using-machine-learning/
- FINRA. (2017). A vibrant market is at its best when it works for everyone. | FINRA.org. Finra.org. https://www.finra.org/
- Tanisha.Digital. (2024, December 13). The Essentials of Data Manipulation and Analysis in Data Science. Medium; Gen AI Adventures. https://medium.com/gen-ai-adventures/the-essentials-of-data-manipulation-and-analysis-in-data-science-a58343838031