

**Master of Data Analytics (Spring 2025)**

**DAMO-611-2 Data Analytics Case Study 3**

Course Project Final Report

**Bank Loan Repayment Analysis – Comprehensive Report**

**Submitted by:**

Heta Chavda - NF1014555

Yash Patel – NF1009944

Devarsh Oza – NF1003776

Joy Ajayi – NF1002698

**Course Instructor – Omid Isfahanialamdari**

**Date: June 21, 2025**

**Table of Contents:**

1. Introduction

2. Project Objectives

3. Statement of Purpose

4. Scope of the project

5. Background Research and Literature

6. Design and Data Collection Method

7. Methodology / Strategies

8. Business Impact and Conclusion

9. Recommendation.

References

**1. Introduction**

In the contemporary financial ecosystem, lending institutions operate under the continuous pressure of minimizing risk while maximizing returns. One of the most critical challenges faced by banks and financial service providers is accurately assessing the creditworthiness of potential borrowers. Poor lending decisions can lead to an increase in non-performing loans (NPLs), which not only impact profitability but also affect regulatory compliance and investor confidence.

As digital transformation drives data availability, institutions are turning to advanced analytics and machine learning models to augment traditional risk assessment frameworks. Predictive modeling has emerged as a powerful technique in loan underwriting processes, offering insights based on historical data and borrower attributes.

This project leverages such techniques—specifically logistic regression—to predict whether a loan will be fully repaid or charged off, using real-world bank loan data.

**2. Project Objectives**

This project aims to achieve the following objectives:

* **Build a predictive model** that classifies loan applications as “Fully Paid” or “Charged Off” using borrower and loan-specific features.
* **Evaluate the impact** of key variables such as interest rate, annual income, debt-to-income (DTI) ratio, and employment length on loan repayment.
* **Support financial decision-making** by identifying risk-prone loan segments and offering actionable recommendations to reduce charge-off rates.
* **Demonstrate the applicability of logistic regression** in financial risk modeling, supported by sound statistical metrics.

**3. Statement of Purpose**

**Background:**

Loan default is a pressing issue in the banking sector, leading to significant losses annually. As institutions expand access to credit, especially in developing economies or digitally enabled microfinance platforms, it becomes increasingly important to adopt robust, data-driven lending strategies.

**Relevance:**

By using predictive analytics, banks can automate loan approval workflows, improve risk segmentation, and ensure compliance with prudential standards. Predictive modeling is especially relevant given the availability of structured financial data and the need for scalable decision systems.

**Goal:**

The goal of this project is to provide a proof of concept for the use of logistic regression in evaluating loan repayment risk. The model will use a historical dataset to identify high-risk borrowers before loan disbursement, thereby optimizing credit risk management.

**4. Scope of the Project**

**Included within scope:**

* Data cleansing and preprocessing of a real-world loan dataset (financial\_loan.csv)
* Exploratory Data Analysis (EDA) to uncover trends and insights
* Feature engineering (e.g., numeric encoding of employment length)
* Logistic regression modeling for binary classification
* Model performance evaluation using ROC-AUC, precision, recall, and confusion matrix
* Business-oriented recommendations based on model results

**Excluded from scope:**

* Use of external credit scores (e.g., FICO)
* Macroeconomic variables (e.g., unemployment rate, inflation)
* Time-series forecasting or longitudinal analysis
* Integration with production systems

**5. Background Research and Literature**

Several academic studies validate the importance and effectiveness of predictive models in credit risk assessment.

According to ***Chris Ekai, July 2023***, Credit risk assessment is an essential component of any lending or financial institution’s operations. It involves evaluating the likelihood that a borrower will default on their debt obligations, helping to protect the lender from potential losses.

Establishing a process for this assessment can be complicated and requires an understanding of both the borrower’s financial situation and current market trends. An effective credit risk assessment helps lenders/financial institutions identify potential defaults and protect themselves from incurring losses while continuing to provide loans.

Credit risk assessment models help evaluate and manage the likelihood of borrower default and potential financial losses. Below are some widely used techniques:

**i. Credit Scoring Models**: These models assign numerical scores that reflects the likelihood of default. It incorporates factors such as payment history, credit utilization, length of credit history, and types of credit accounts.

**ii. Credit Risk Models:** This uses advanced techniques to estimate the likelihood of borrower default and potential losses. They incorporate historical data, financial ratios, and economic indicators. Methods like logistic regression, decision trees, and machine learning help refine these predictions and improve accuracy.

**iii. Credit Valuation Models:** It determines the fair value of credit-related instruments, such as bonds, by evaluating default risk and potential recovery rates. Models like Credit Valuation Adjustment (CVA) and the Merton model help investors and institutions price and manage credit risk effectively.

**iv. Management Quality Assessment:** This involves evaluating the borrower’s management team to gauge credit risk. Factors such as experience, competence, and past performance are assessed. Strong management usually indicates better decision-making and risk management, which can lower the risk of default.

**v. Industry and Economic Analysis:** This technique examines industry trends and economic conditions to understand how they affect a borrower’s credit risk. By evaluating factors like economic cycles and regulatory changes, institutions can better assess how external conditions might impact a borrower’s ability to repay loans.

Predictive modelling in credit risk involves the use of statistical and machine learning techniques to assess the likelihood of borrowers defaulting on loans or credit. These models help financial institutions make informed decisions by predicting the probability of default, loss given default, and exposure at default. Various methodologies are employed in credit risk modelling, including traditional approaches, statistical models like logistic regression and decision trees, as well as machine learning algorithms such as support vector machines and neural networks. ***Atul K. Gupta, February 2024***

A diagram of a credit risk management

AI-generated content may be incorrect.

*Source: Credit Risk Assessment Guide: Strategy for Effective Credit Risk Monitoring*

These foundational works underscore the relevance of using logistic regression in this project and establish its practical utility in the financial services industry.

**6. Design and Data Collection Methods**

**Data Source:**

The dataset used in this project was sourced from a bank’s internal lending records and is provided in CSV format (financial\_loan.csv). The dataset includes key borrower attributes and loan characteristics.

**Key Variables:**

* loan\_status: Loan repayment outcome
* int\_rate: Interest rate on the loan
* dti: Debt-to-income ratio
* annual\_income: Applicant’s self-reported income
* emp\_length: Length of employment

**Data Preparation Steps:**

* Removed records with missing values in critical columns (emp\_title, loan\_status, emp\_length)
* Encoded emp\_length into a numeric field (emp\_length\_clean)
* Created a binary target variable (loan\_status\_binary) for classification
* Discretized income into quantiles for better visualization

**Data Visualization**

* Countplot of loan statuses: revealed class imbalance (majority class dominates)
* Boxplots: dti and int\_rate are higher for defaulted loans
* Tableau dashboards provided further exploration of loan risk by income, state, and loan amount (Insert screenshots below if available)

**a. Loan Status Distribution:**  
plt.figure(figsize=(6, 4))

ax = sns.countplot(data=df, x='loan\_status', palette='Set2')

for p in ax.patches:

ax.annotate(f'{p.get\_height()}',

(p.get\_x() + p.get\_width() / 2, p.get\_height()),

ha='center', va='bottom', fontsize=10)

plt.title('Loan Status Distribution')

plt.tight\_layout()

plt.show()

A graph of a loan status distribution

AI-generated content may be incorrect.

Out of 37,138 loans, 31,005 (84%) have been fully repaid, 5,067 (14%) were charged off, and 1,066 (3%) remain current. This updated chart reaffirms that most borrowers successfully complete repayments, with defaults and ongoing loans comprising only a small share.

**b. Loan Status by Employment Length:**   
plt.figure(figsize=(10, 5))

ax = sns.countplot(data=df, x='emp\_length', hue='loan\_status', palette='coolwarm')

for p in ax.patches:

ax.annotate(f'{p.get\_height()}',

(p.get\_x() + p.get\_width() / 2, p.get\_height()),

ha='center', va='bottom', fontsize=9)

plt.title('Loan Status by Employment Length')

plt.xticks(rotation=45)

plt.tight\_layout()

plt.show()A graph of a loan status

AI-generated content may be incorrect.Across all tenure groups, fully paid loans remain dominant—particularly among borrowers with 10+ years of employment (7,157 fully paid vs. 1,322 charged off). Default rates (charged-off ÷ fully paid) stay around 12–15% for most categories, indicating consistent risk regardless of employment length.

**c. DTI by Loan Status:**   
plt.figure(figsize=(8, 5))

sns.boxplot(data=df, x='loan\_status', y='dti', palette='Set3')

plt.title('DTI vs Loan Status')

plt.tight\_layout()

plt.show()**A diagram of a different color scheme

AI-generated content may be incorrect.**

Borrowers who charged off have a slightly higher median DTI (14.5%) than those who fully repaid (13%), while current loans sit at the highest median (15%). Although the distributions overlap substantially, defaulters tend to carry marginally greater debt burdens than repayers.

**d. Interest Rate by Loan Status:**

plt.figure(figsize=(8, 5))

sns.boxplot(data=df, x='loan\_status', y='int\_rate', palette='magma')

plt.title('Interest Rate vs Loan Status')

plt.tight\_layout()

plt.show()A graph showing different colored squares

AI-generated content may be incorrect.

Current loans carry the highest median interest rates, charged-off loans sit in the middle, and fully repaid loans exhibit the lowest rates. This pattern reinforces that higher borrowing costs tend to coincide with reduced repayment likelihood.

**Tableau Dashboard**A screenshot of a computer screen

AI-generated content may be incorrect.

**7. Methodology / Strategies**

**Modeling Technique:**

A **Logistic Regression** model was selected for its interpretability and statistical reliability in binary outcome prediction.

**Steps Taken:**

1. **Feature Selection**: Used int\_rate, dti, annual\_income, and emp\_length\_clean
2. **Target Encoding**: Transformed loan\_status into a binary variable
3. **Train-Test Split**: 80% training, 20% testing
4. **Model Training**: Used class\_weight='balanced' to address class imbalance
5. **Evaluation Metrics**:
   * **Confusion Matrix**: Provided insights on classification errors
   * **Classification Report**: Included precision, recall, F1-score
   * **ROC-AUC Score**: 0.9154, indicating strong predictive performance

**Key Findings:**

* **High interest rates** and **high DTI** are strong predictors of loan default.
* **Higher income** and **longer employment** correlate positively with repayment.

**8. Business Impact and Conclusion**

**Potential Business Benefits:**

* **Risk Mitigation**: Early identification of high-risk applicants reduces potential loan losses.
* **Portfolio Optimization**: Banks can rebalance lending portfolios to favor lower-risk borrowers.
* **Automation**: Integration with digital lending platforms can enable near-instantaneous loan decisions.
* **Customer Targeting**: Enables personalized interest rates and credit terms based on risk segmentation.

**Conclusion:**

The logistic regression model demonstrated strong performance in predicting loan repayment outcomes. Its simplicity and accuracy make it a suitable choice for operational deployment in risk management workflows. It allows banks to make informed, data-driven lending decisions while ensuring compliance with regulatory standards.

**9. Recommendations**

Based on the findings of this project, the following actions are recommended:

i. **Deploy the model in pre-loan screening**

**Expanded:**  
Integrate the logistic regression model into the bank's loan origination platform to flag applications with a high probability of default. This can be done via API endpoints, enabling real-time risk scoring during the approval workflow. Establish threshold cutoffs (e.g., probability > 0.70) to trigger manual review or stricter underwriting.

*Action item:*  
Develop a dashboard that allows loan officers to visualize risk probabilities and the key drivers (e.g., high DTI or low income) for each flagged application.

ii. **Retrain the model quarterly**

**Expanded:**  
Borrower behavior and macroeconomic conditions evolve—so should your model. Schedule quarterly retraining pipelines using newly approved loan data. This ensures performance metrics such as precision and recall don’t degrade over time due to data drift.

*Action item:*  
Set up an automated pipeline using tools like Python + Airflow or Azure ML, incorporating data version control and retraining alerts based on performance thresholds.

iii. **Add more variables for enhanced performance**

**Expanded:**  
Incorporate features such as:

* *Credit history* (e.g., number of delinquencies, account age)
* *Loan purpose* (e.g., debt consolidation, home improvement)
* *Past payment patterns* (e.g., missed payments, partial repayments)

These features often have strong predictive power and help improve model discrimination.

*Action item:*  
Partner with credit bureaus or internal data teams to access and engineer these new features, ensuring privacy and compliance protocols are respected.

iv. **Test ensemble models**

**Expanded:**  
While logistic regression is interpretable, ensemble methods like *Random Forest* or *Gradient Boosting* (e.g., XGBoost) might provide better accuracy, especially in handling complex feature interactions. Running side-by-side model comparisons could validate whether interpretability or precision matters more in your use case.

*Action item:*  
Run k-fold cross-validation on ensemble models and benchmark them against the current logistic model using ROC-AUC, precision, recall, and lift charts.

v. **Collaborate with data governance and compliance teams**

**Expanded:**  
Ensure that model usage aligns with consumer protection laws (like the Fair Lending Act or GDPR, depending on jurisdiction). You’ll need documentation covering:

* Model logic and audit trails
* Fairness and bias testing (e.g., checking for disparate impact)
* Explainability techniques (like SHAP values)

*Action item:*  
Initiate a cross-functional committee with Risk, Compliance, and Legal teams to formally approve the model and monitor for regulatory shifts.

**References**:

1. Atul K. Gupta, Credit Risk Predictive Modeling, February 24, 2024. <https://Linkedin.com/pulse/credit-risk-predictive-modelling-atul-k-gupta-gyq6f>

2. Chris Ekai, July 15 2023, Risk Publishing, Credit Risk Assessment: A Comprehensive Guide for Lenders and Financial Institutions. <https://riskpublishing.com/credit-risk-assessment-a-comprehensive-guide-for-lenders-and-financial-institutions/>