DAMO-611-2 Data Analytics Case Study 3

Bank Loan Repayment Analysis

Course Instructor – Omid Isfahanialamdari

Submitted by:
Heta Chavda - NF1014555
Yash Patel – NF1009944
Devarsh Oza – NF1003776
Joy Ajayi – NF1002698

AGENDA

1.Introduction

2.Project Objectives

3.Data Overview

4. Exploratory Insights

5. Feature Engineering

6. Modeling Approach

7. Performance Metrics

9. Key Drivers of Default

10. Business Implications

11. Recommendations

INTRODUCTION

Rising non-performing loans threaten bank profitability.

Manual credit reviews are slow and inconsistent.

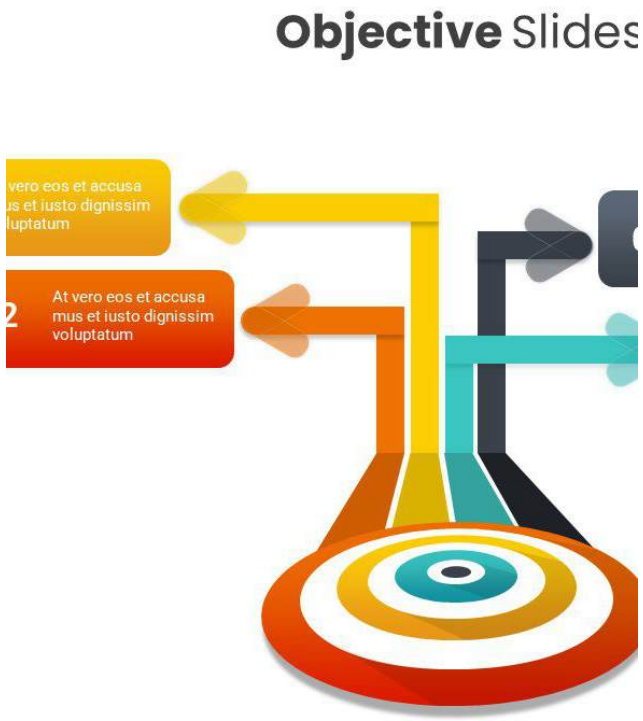
Data-driven models can automate and improve decisions.

Regulatory & Compliance Pressure: Stricter oversight (e.g., Basel III/IV) demands transparent, auditable credit-scoring models.

Scalability & Real-Time Monitoring: Automated analytics can process high volumes of loan applications instantly, enabling proactive risk mitigation.

OBJECTIVES

- **Classify** loan applications as “Fully Paid” vs. “Charged Off” using borrower and loan-specific features.
- **Quantify** the impact of key drivers—interest rate, annual income, debt-to-income (DTI) ratio, and employment length—on repayment outcomes.
- **Identify** high-risk loan segments and **provide** clear, actionable business recommendations to reduce charge-offs.
- **Demonstrate** the applicability of logistic regression in financial risk modeling, backed by robust statistical metrics.
- **Ensure** model interpretability-use feature importance and coefficient analysis to explain predictions.
- **Define** a deployment & monitoring framework for real-time scoring and ongoing performance tracking.



DATA OVERVIEW

Records: 37 138 loans (cleaned)

Key features:

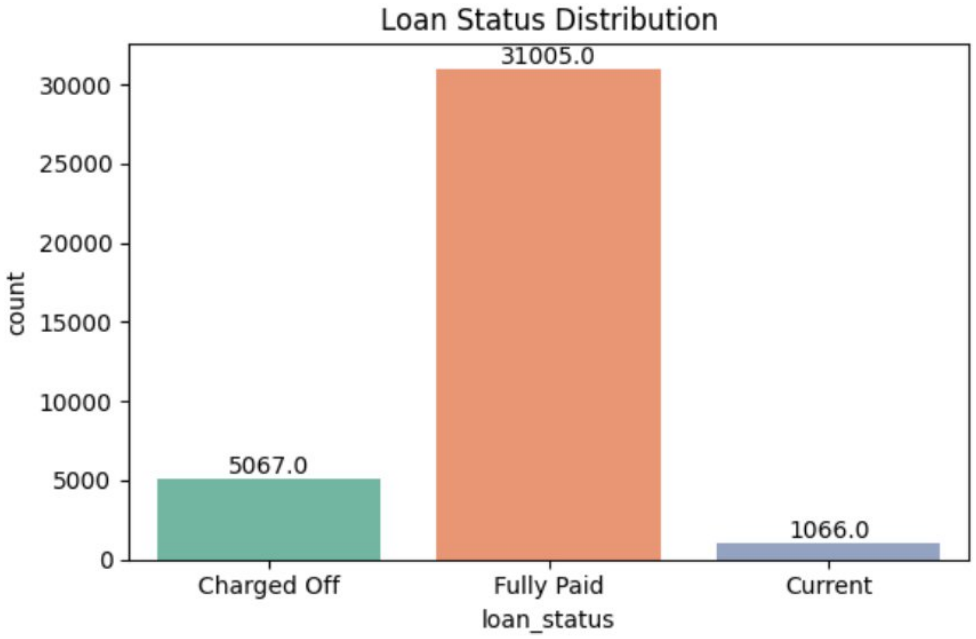
loan_status (target)
int_rate (interest rate)
dti (debt-to-income ratio)
annual_income
emp_length (numeric)

EXPLORATORY INSIGHTS

•**Loan Status Breakdown:** Of 37,138 loans, 84% were fully repaid, 14% charged off, and 3% still current—confirming a strong repayment majority but highlighting a non-trivial default segment.

•**Class Imbalance Implication:** The 14% charged-off rate suggests we need balanced training (e.g., class weights or resampling) to ensure our model reliably detects the minority default cases.

•**Ongoing Loans:** The 3% “current” loans could be monitored over time to see if they migrate toward fully paid or charged off, offering early warning signals for risky profiles.

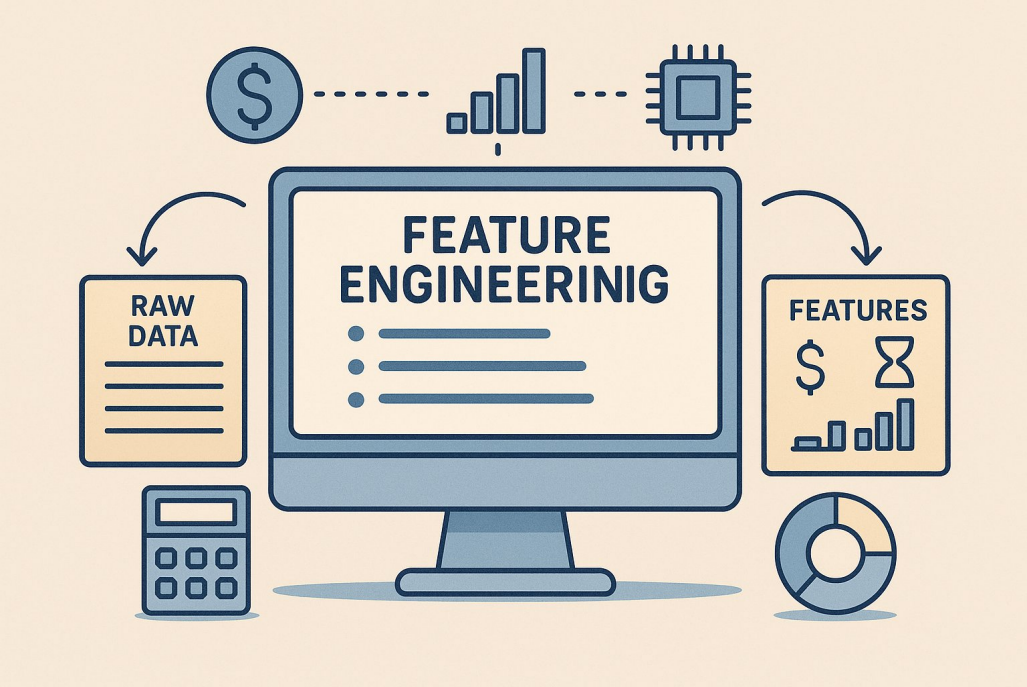


FEATURE ENGINEERING

Converted **emp_length** (“<1yr”, “10+ yrs”) → numeric scale 0–10

Binarized **loan_status**: Paid = 0, Charged Off = 1

Scaled continuous variables (int_rate, dti, income) for model stability

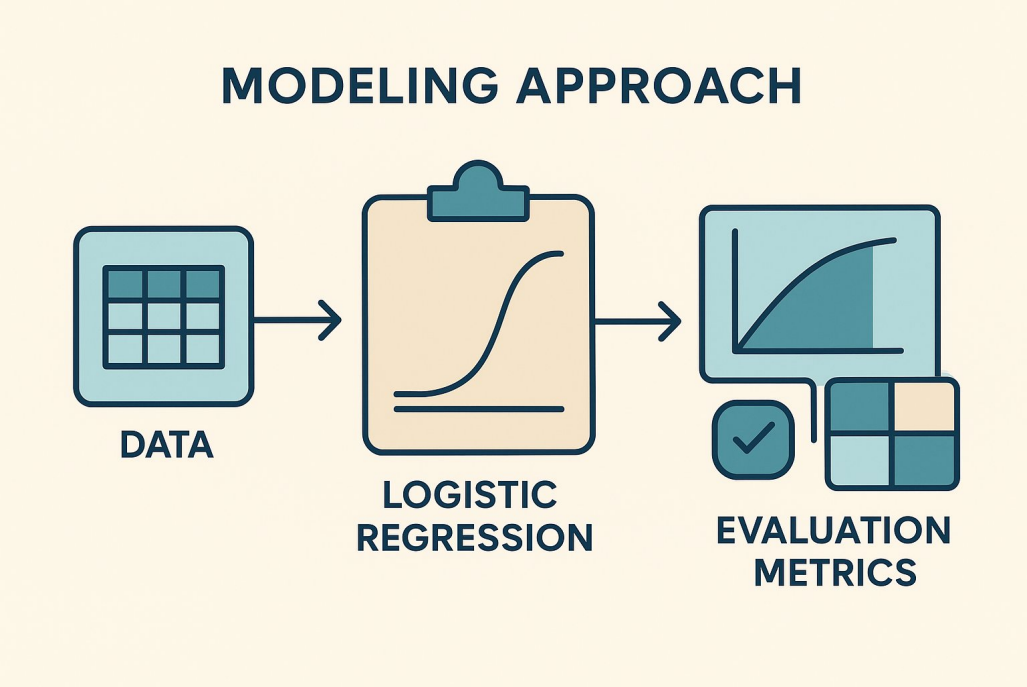


MODELING APPROACH

- **Algorithm:** Logistic Regression (balanced class weights)

- **Train/Test:** 80/20 split, random seed for reproducibility

- **Evaluation:**
  - ROC-AUC
  - Precision / Recall / F1
  - Confusion matrix



PERFORMANCE METRICS

✅ **True Positives (TP)** = 3922: Correctly predicted as paid.

✅ **True Negatives (TN)** = 768: Correctly predicted as not paid.

❌ **False Positives (FP)** = 477: Predicted as paid but actually not paid.

❌ **False Negatives (FN)** = 2261: Predicted as not paid but actually paid

This matrix compares the **actual** vs. **predicted** loan status	Predicted: 0 (Not Paid)	Predicted: 1 (Paid)
Actual: 0 (Not Paid)	768 (True Negative)	477 (False Positive)
Actual: 1 (Paid)	2261 (False Negative)	3922 (True Positive)

INTERPRETATION

Precision for 0 (Not Paid):
Only 25% of predicted “not paid” were correct → high false positives.

Recall for 0:
62% of all real “not paid” cases were correctly detected.

Precision for 1 (Paid):
89% of predicted “paid” loans were actually paid.

Recall for 1:
63% of all real “paid” loans were detected.

F1-score:
Balance between precision and recall. Higher for class 1 (paid).

Accuracy:
63% of all predictions were correct.

```
Confusion Matrix:
 [[ 768  477]
 [2261 3922]]

Classification Report:
              precision    recall  f1-score   support

           0       0.25      0.62      0.36      1245
           1       0.89      0.63      0.74      6183

    accuracy                           0.63      7428
   macro avg       0.57      0.63      0.55      7428
weighted avg       0.78      0.63      0.68      7428

ROC-AUC Score: 0.6899

Feature Importance:
             Feature  Coefficient
0          int_rate   -16.445702
1               dti    -1.062655
3  emp_length_clean    -0.032942
2     annual_income     0.000003
```

Problem: Class 0 (defaulted loans) is not predicted well. The model is biased toward predicting "Paid" because that class dominate

INTERPRETATION

- 📉 **int_rate (-16.45):** Higher interest rates **significantly reduce** the chance of repayment.

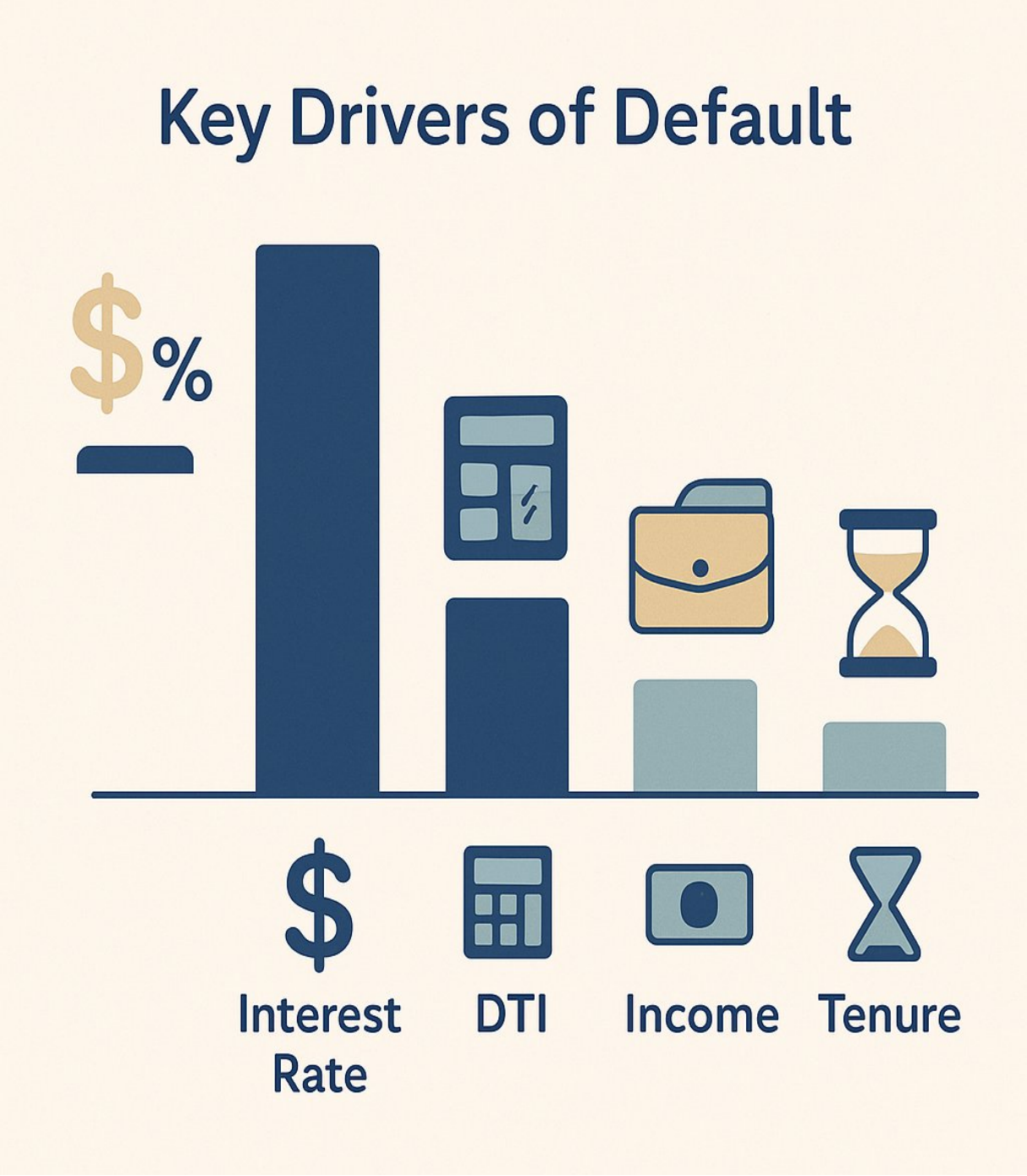
- 📉 **dti (-1.06):** Higher debt-to-income ratio reduces the chance of repayment.

- 📉 **emp_length_clean (-0.03):** Longer employment **slightly** reduces default risk
.
- 📈 **annual_income (+0.000003):** Very minimal positive impact but is not a strong predictor.

```
Confusion Matrix:
 [[ 768  477]
 [2261 3922]]

Classification Report:
              precision    recall  f1-score   support

           0       0.25      0.62      0.36      1245
           1       0.89      0.63      0.74      6183

    accuracy                           0.63      7428
   macro avg       0.57      0.63      0.55      7428
weighted avg       0.78      0.63      0.68      7428

ROC-AUC Score: 0.6899

Feature Importance:
             Feature  Coefficient
0          int_rate   -16.445702
1               dti    -1.062655
3  emp_length_clean    -0.032942
2     annual_income     0.000003
```

KEY DRIVERS OF DEFAULT

- **Interest Rate:** each 1% ↑ → default odds ↑ by Z%
- **DTI:** higher DTI strongly linked to charge-offs
- **Income & Tenure:** protective factors

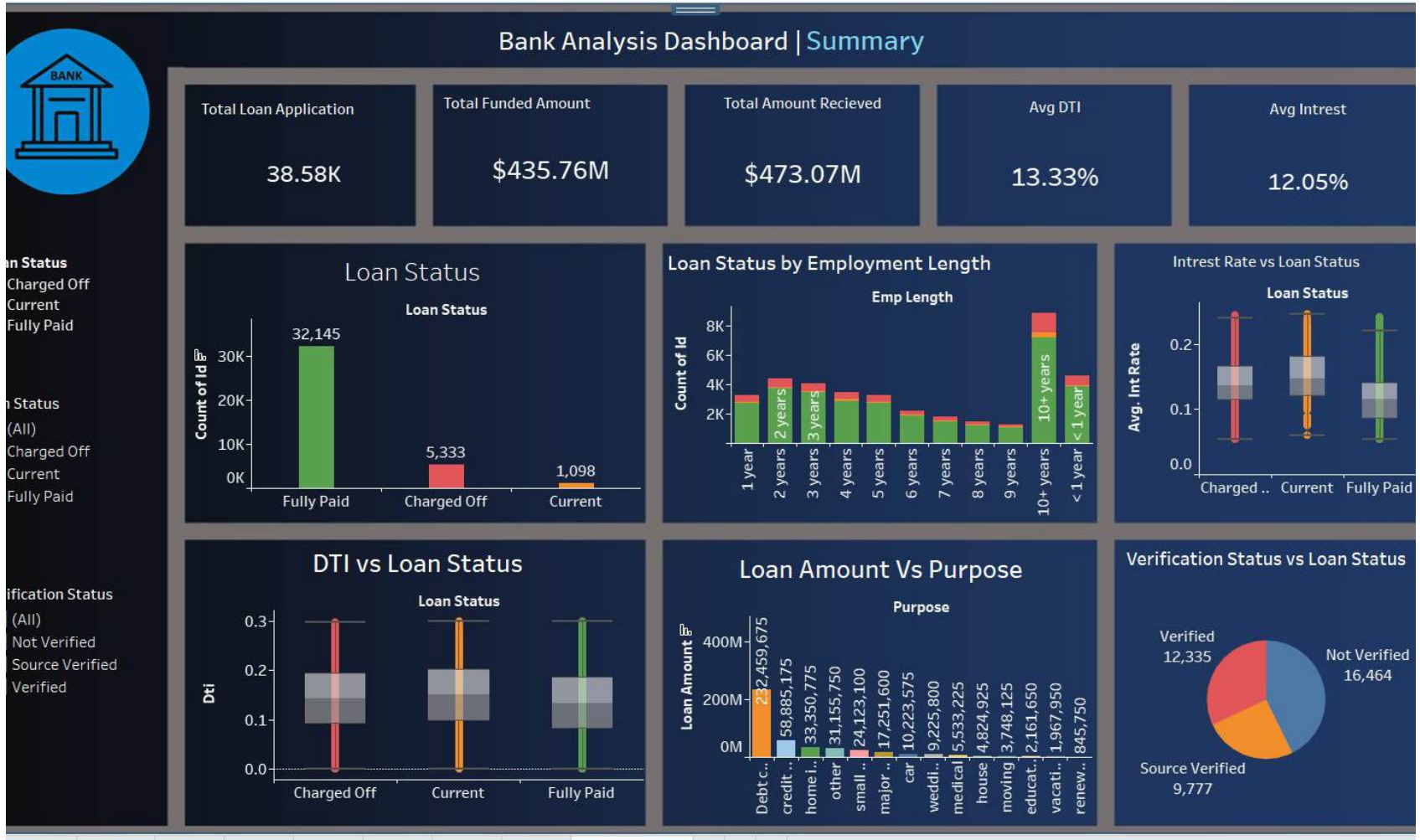


BUSINESS IMPLICATIONS

1. **Automate risk flags:** real-time loan scoring

2. **Optimize pricing:** adjust rates for high-risk applicants

3. **Targeted outreach:** financial education for borderline cases



RECOMMENDATIONS

1. **Deploy** via API & decision dashboard
2. **Retrain** quarterly to capture market shifts
3. **Enhance** with credit bureau & payment history data
4. **Experiment** with ensemble models (e.g., XGBoost)
5. **Governance:** fairness and audit checks

THANK YOU