esade

Business Case:
Lending club

Do Good. Do Better.

01 – Goals Definition

02 – Data Preparation

03 – Model Creation

04 – Model Interpretation

05 – Model Implementation

01

Goals Definition

- Implement supervised classification algorithms to predict loan defaults for Lending Club.

- Analyze confusion matrices and key performance metrics to derive meaningful insights.

- Minimize the Impact of False Predictions

- Identify the most profitable strategy while effectively managing financial risk.

02

Data Preparation

Handling a Large Dataset: Initial Approach

- **Dataset Characteristics:** 2,029,950 rows and 140 columns → Large and difficult to process.
- **Sampling Strategy:** To ensure efficiency, we take a 5% chunk (~101,500 rows).
- **Feature Selection Challenge:**

1. Dropped post-issuance data and applied filtering techniques:
2. Removed columns with >75% missing values.
3. Dropped low variance features.
4. Used Random Forest Regressor to drop features with <1% importance.

- **Issue:** Too many features were removed, leaving only 15 out of 140.

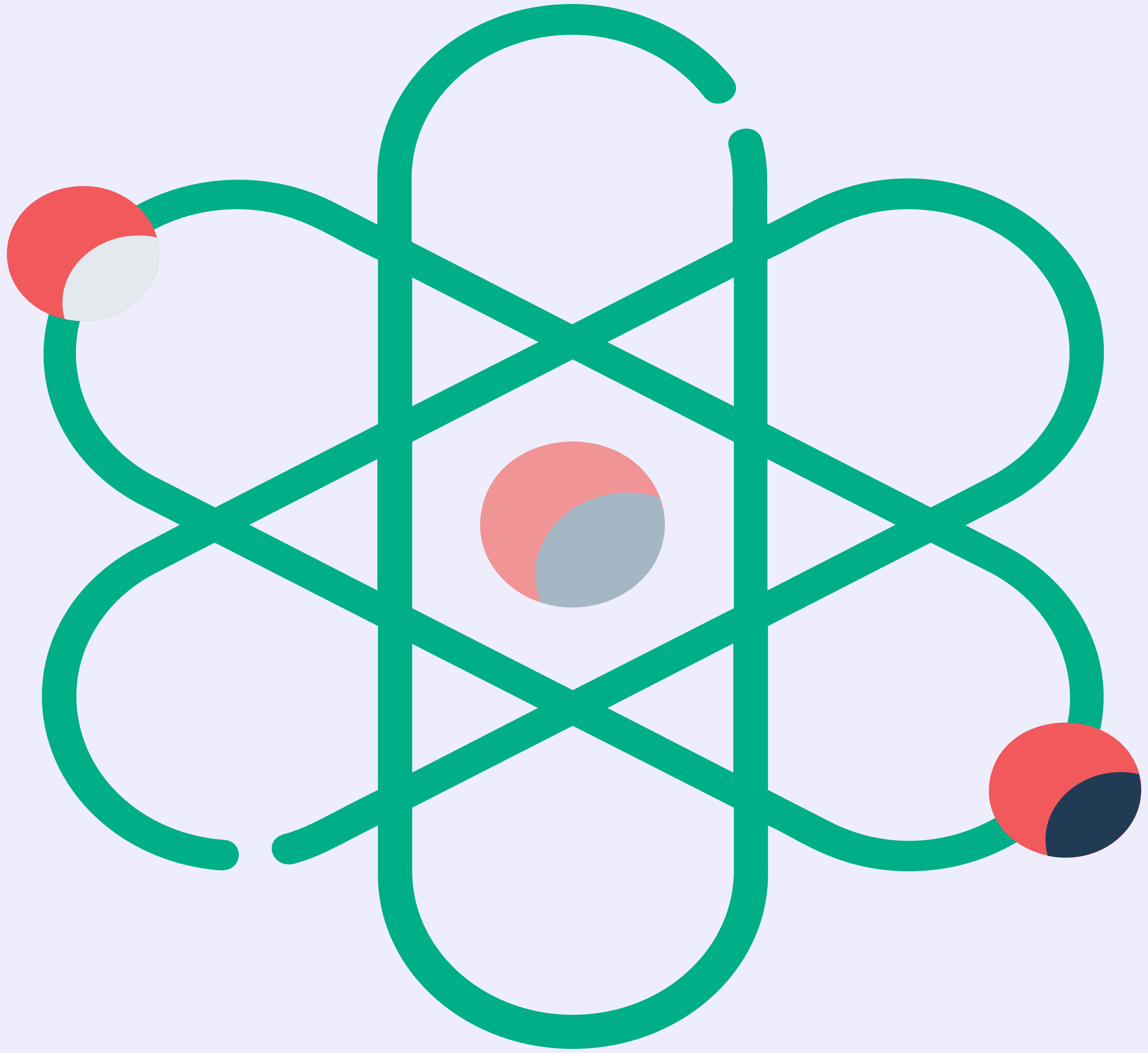
·**Result:** Start all over again ☹

Refining Feature Selection for a Balanced Dataset

- **New Strategy:** Instead of removing features based on missing values and variance, we:

  1. Reviewed feature descriptions to assess relevance.
  2. Removed post-issuance and unreliable features.
  3. Ensured critical variables were retained for model performance.

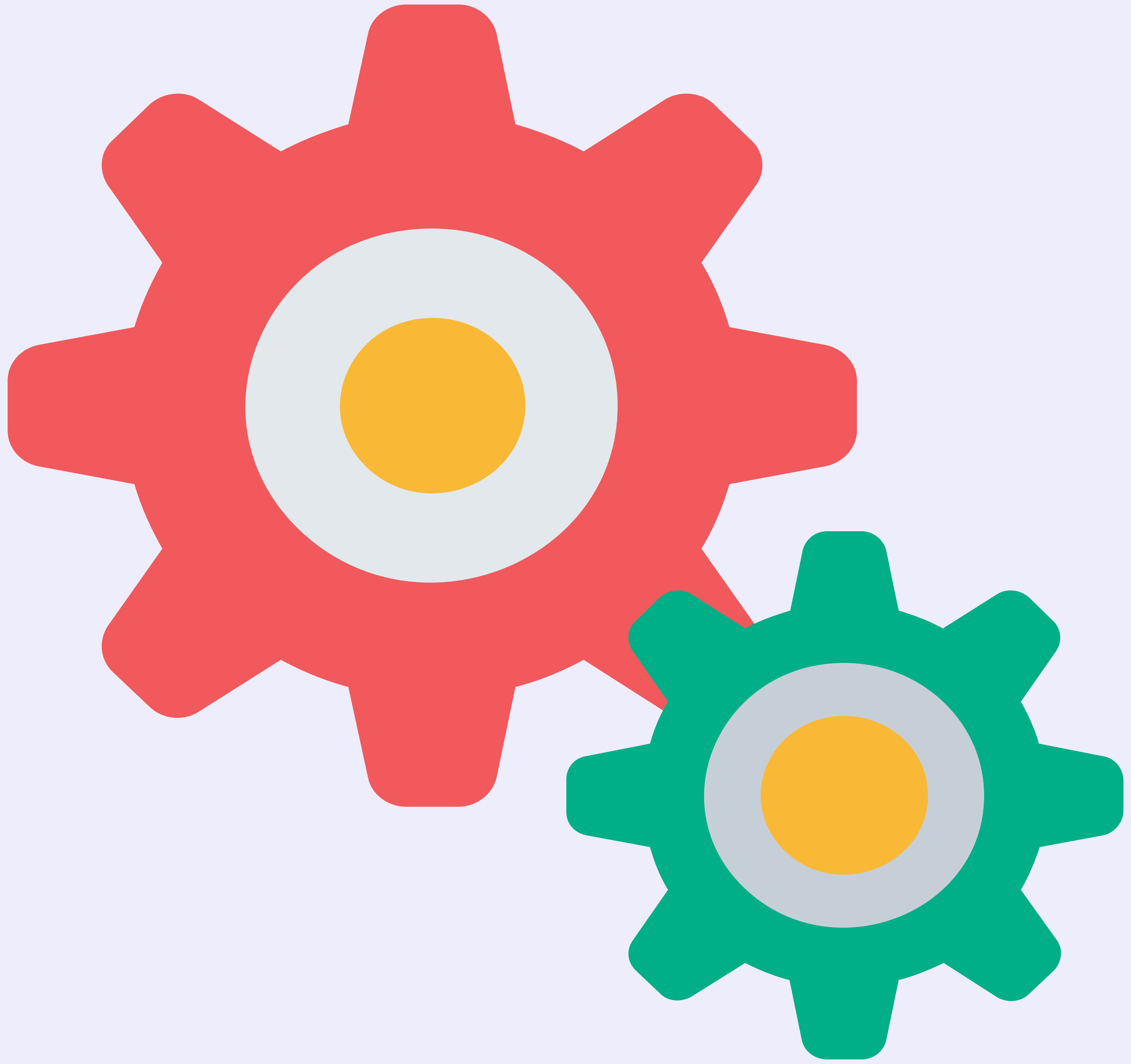
- **Hardship Columns:** Removed (only 3% of samples affected).
- **Goal:** Improve model efficiency while retaining essential features.



Optimizing Data Quality for Model Performance

- **Manual Column Dropping:** Removed features irrelevant to our goals.
- **Handling Missing Values:**
  - Replaced float NaNs with 0.
  - Replaced object NaNs with ‘Unknown.’

- **Categorical Encoding Strategy:**
- **Binary Features:** One-hot encoding.
- **Categorical with ≤10 values:** Label encoding.
- **Categorical with >10 values:** Custom preprocessing (e.g., removing ‘%’ and converting to float).



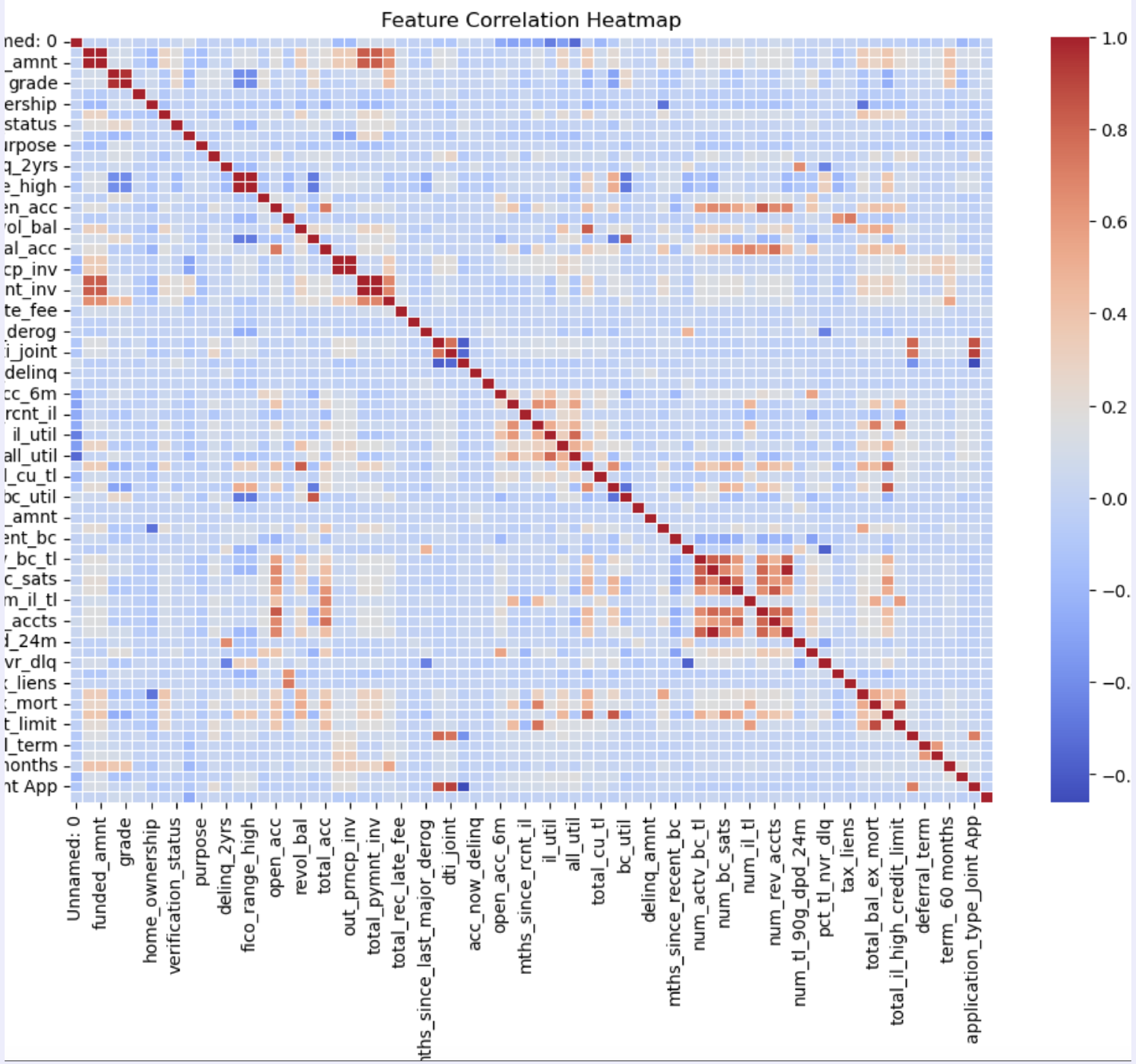
Final Feature Set & Ensuring Data Integrity

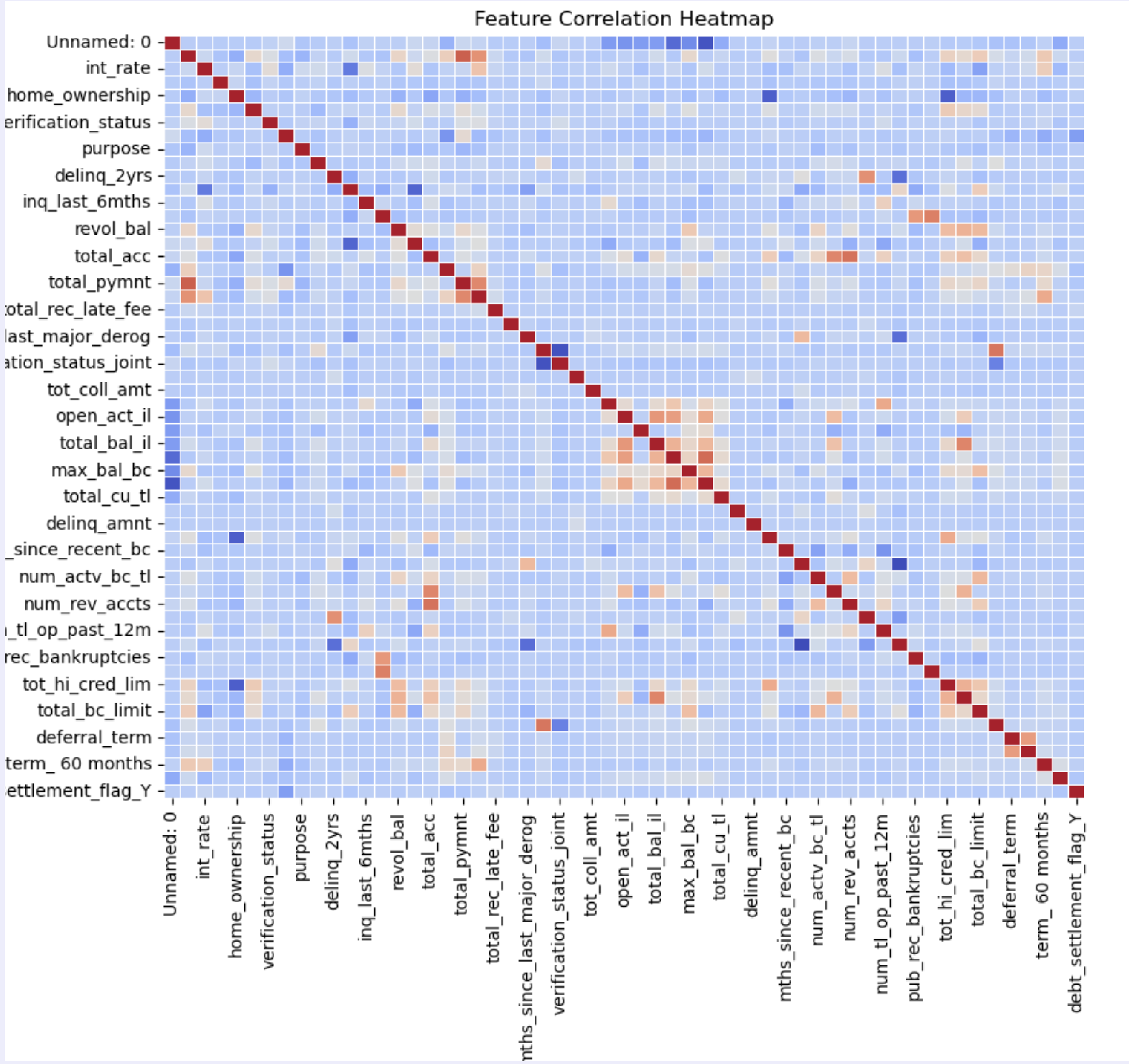
- **Correlation Analysis:**

- Computed correlation matrix.
- Created a heatmap to visualize feature relationships.
- Removed features with >80% correlation to avoid redundancy.

- **Final Outcome:** Retained 56 out of 140 original features.
- **Key Decision:** Did **not** create new features from existing ones.
- **Result:** A more balanced and efficient dataset for model training.

Before vs After





03

Model Creation

Filtering and Creating Target Variable

- **Filtering the Dataset:**
- Focus on closed loans (Fully Paid, Charged Off, Default).
- Removed ongoing loans for accurate classification.

- **Defining Target Variable (y):**
- Good Loan (Fully Paid) → 1
- Bad Loan (Charged Off, Default) → 0

- **Features (X):**
- Dropped loan_status after mapping to binary labels.

- **Target Variable Distribution:**
- Display count and percentage of Good vs. Bad loans.

```python
# Create the target variable
y = df['loan_status'].map({
    3: 1,  # Fully Paid -> Good
    1: 1,  # Current -> Good
    0: 0,  # Charged Off -> Bad
    6: 0,  # Late (31-120 days) -> Bad
    4: 0,  # In Grace Period -> Bad
    5: 0,  # Late (16-30 days) -> Bad
    2: 0   # Default -> Bad
})
```

Handling Imbalanced Data with SMOTE & Undersampling

- **Splitting Data:**
- 70% Training, 30% Testing.

- **Addressing Imbalance:**
- Used **SMOTE** to oversample the minority class.
- Applied **Random Under sampling** to balance majority class.

- **Post-Balancing Target Distribution:**
- Ensured even representation of Good vs. Bad loans.

Evaluating Classification Models

- **Trained Three Models:**

- **Logistic Regression**
- **Random Forest Classifier**
- **Support Vector Machine (SVM)**

- **Performance Metrics:**
- Training & Testing Accuracy.
- AUC-ROC Score (for imbalanced classification).
- Precision-Recall AUC (important for minority class).

- **Best Model Selection:**
- Model with the highest **AUC-ROC** used for further analysis.
- **Confusion Matrix Visualization:**
- Show prediction performance of the selected model.

04

Model Interpretation

Choosing the Right Metrics for Model Assessment

- **Precision:** Measures how many predicted positive cases were actually positive.
  - High precision reduces false positives (incorrectly predicting bad loans as good).

- **Recall:** Measures how many actual positive cases were correctly predicted.
  - High recall reduces false negatives (missing bad loans).

- **F1 Score:** Harmonic mean of precision and recall.
  - Balances false positives and false negatives.

- **AUC-ROC Score**: Measures model’s ability to distinguish between classes.
  - Higher AUC → Better classification ability.

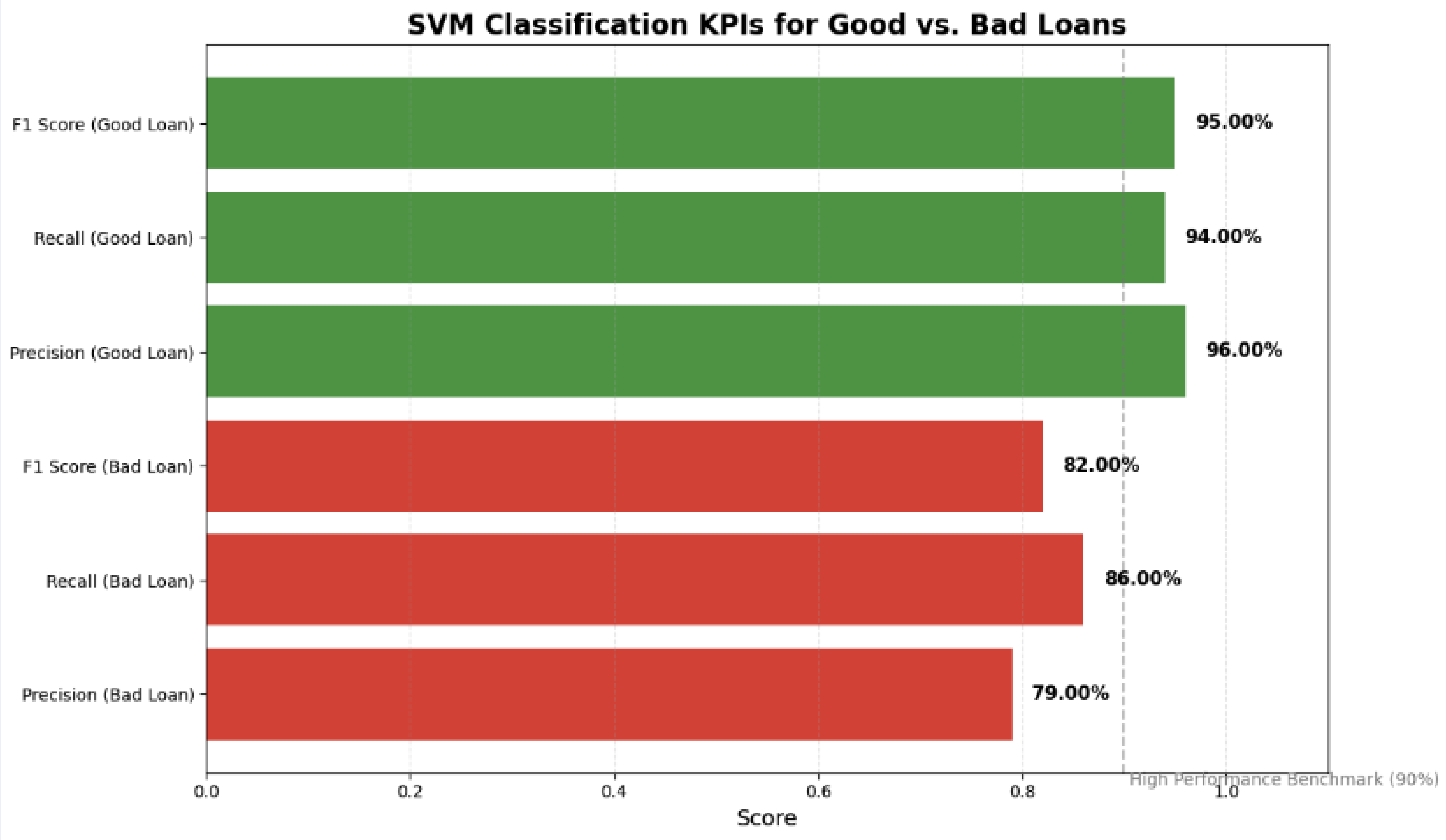
- **Precision-Recall AUC:** Best for imbalanced datasets where false negatives are costly.

Analysis of the results of the best model

1. High Precision for Fully Paid Loans (0.96)
• The model is very confident in predicting good loans (Class 1) correctly, meaning fewer false approvals of risky loans.

2. Strong Recall for Defaulted Loans (0.86)
• The model catches 86% of bad loans correctly, which is essential for risk management and minimizing defaults.

3. Overall Accuracy of 93%
• The model performs well, but the macro-averaged recall is slightly lower (0.88), indicating a small trade-off in detecting minority classes (defaulted loans).



Model Performance: Confusion Matrix Comparison

• **Random Forest:**

- High TP (18,059) and TN (3,324) → Strong at identifying both fully paid (good) and defaulted (bad) loans.
- Moderate FP (1,376) and FN (538) → Some defaulted loans misclassified as good loans.

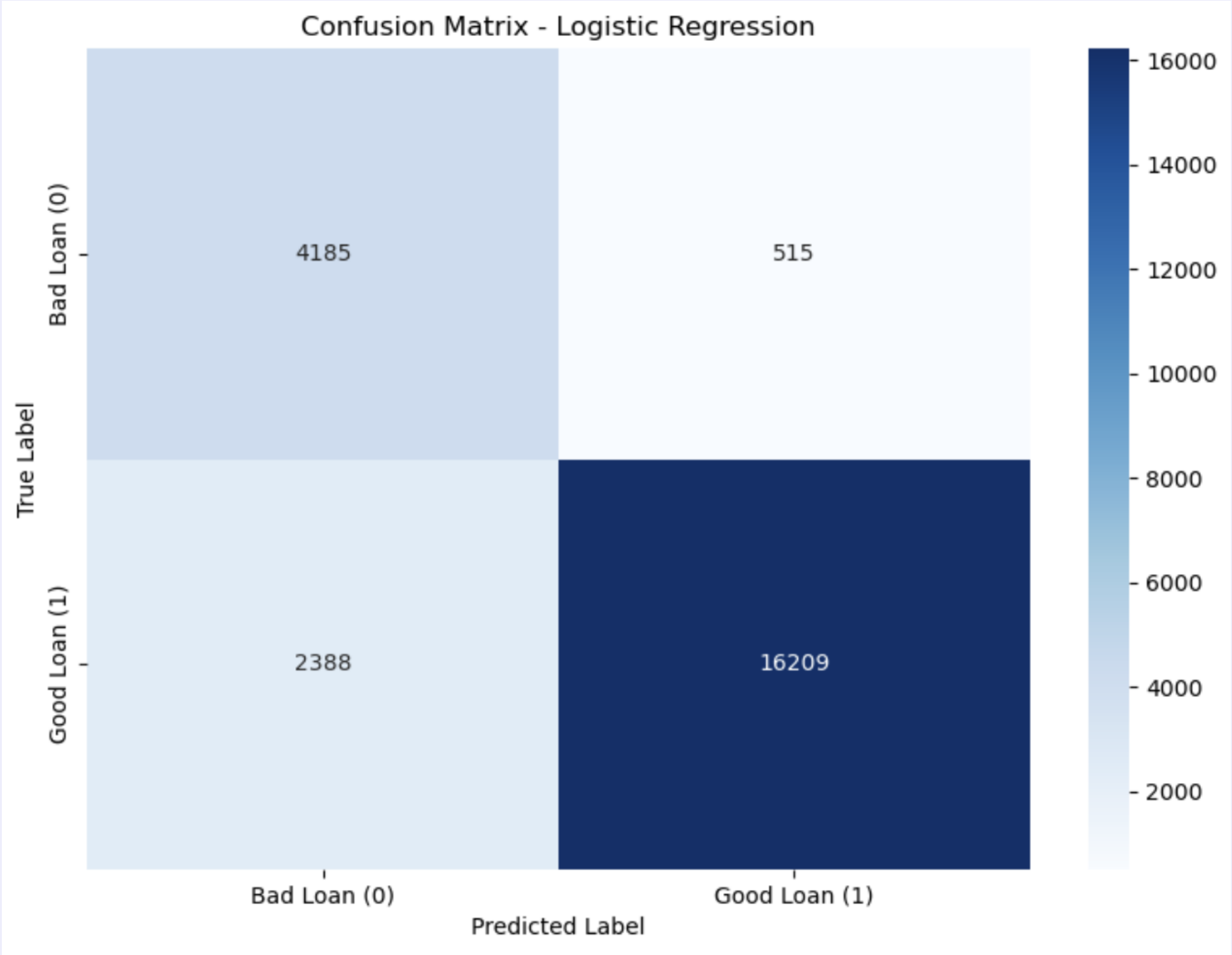
• **Logistic Regression:**

- Higher TN (4,185) → Best at identifying defaulted loans.
- Higher FN (2,388) → More defaulted loans incorrectly classified as fully paid.

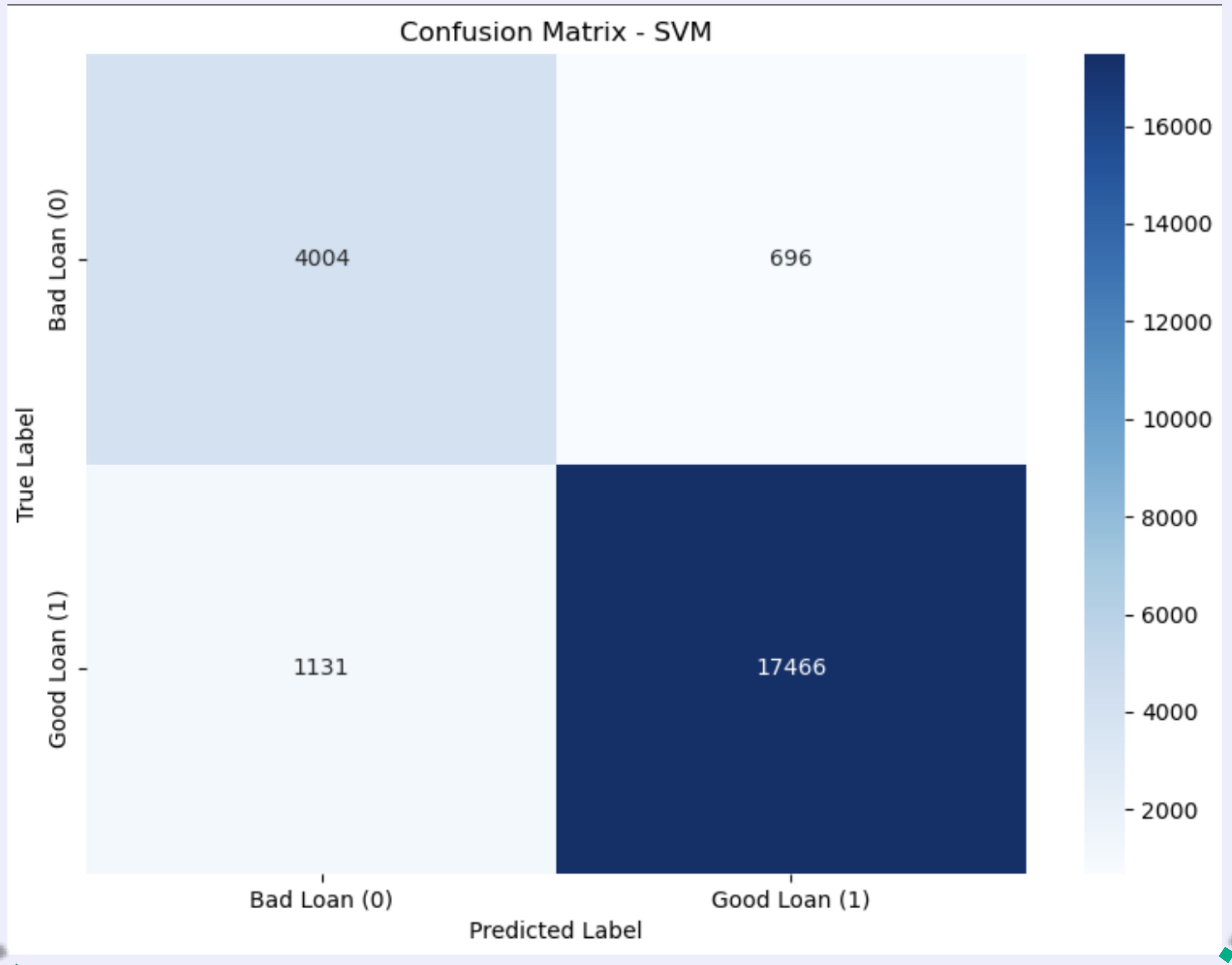
• **SVM:**

- Balanced TP (17,466) and TN (4,004) → Overall best recall for both classes.
- Moderate FP (696) and FN (1,131) → Better precision than Logistic Regression.

Understanding Model Predictions: Confusion Matrix



Best Model



- **4700 actual bad loans → 4025 correctly classified (86% recall).**
- **18597 actual good loans → 17538 correctly classified (94% recall).**
- **Precision for bad loans: 79% (some misclassifications occur).**

05

Model Implementation

How Model Errors Affect Loan Decisions

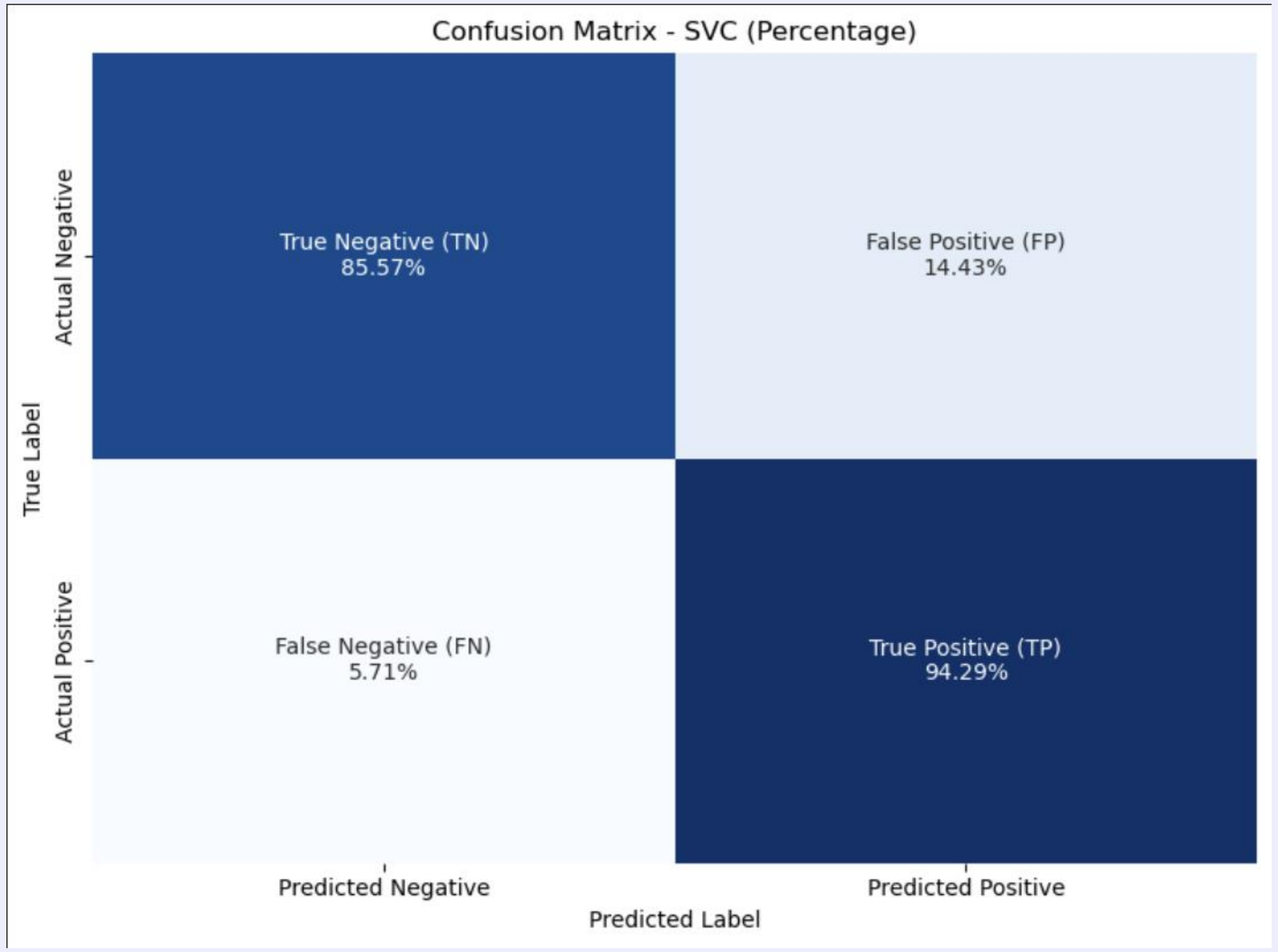
Understanding Model Decisions:

•**True Positives (TP):** Correctly predicted fully paid loans → Generates expected profit.

•**True Negatives (TN):** Correctly predicted defaults → Avoids high-risk investments.

•**False Positives (FP):** Predicted fully paid but actually defaulted → Financial loss.

•**False Negatives (FN):** Predicted default but actually paid → Missed investment opportunities.



Business Impact:

•**FP impact:** Loss of principal (mitigated by recovery rate).

•**FN impact:** Lost interest revenue and over-conservative lending.

•**Solution:** Adjust the decision threshold to optimize risk vs. return.

Optimizing the Decision Threshold

Goal:

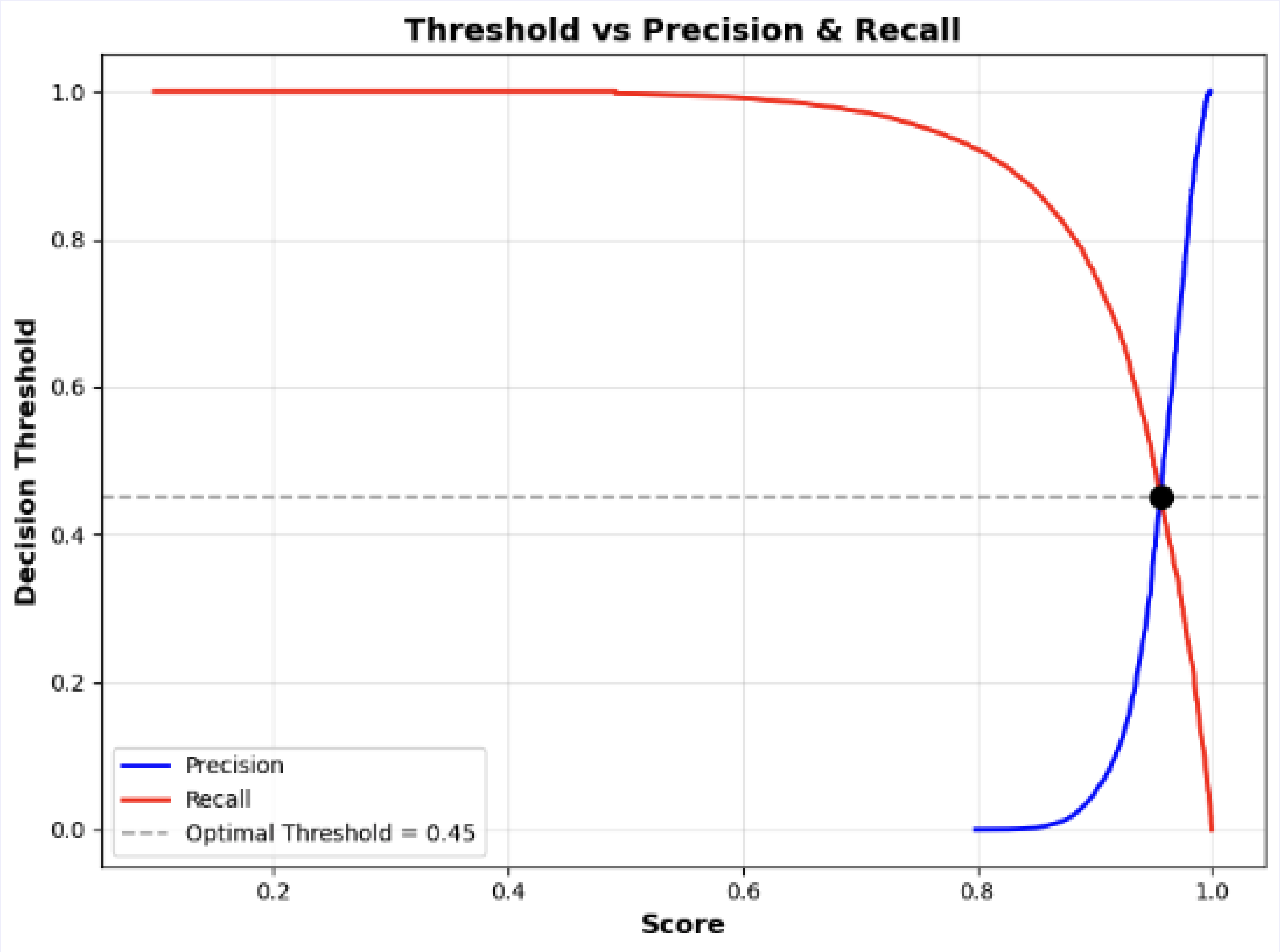
Adjust the classification threshold to balance **approval rate** (recall) and **risk control** (precision).

Visual Insight:

In the graph, we flipped the axes to better observe how **threshold** impacts both metrics.

• **Precision decreases** as threshold lowers: more loans are approved, but risk increases.

• **Recall increases** as threshold lowers: more good loans are captured, but also more bad ones.



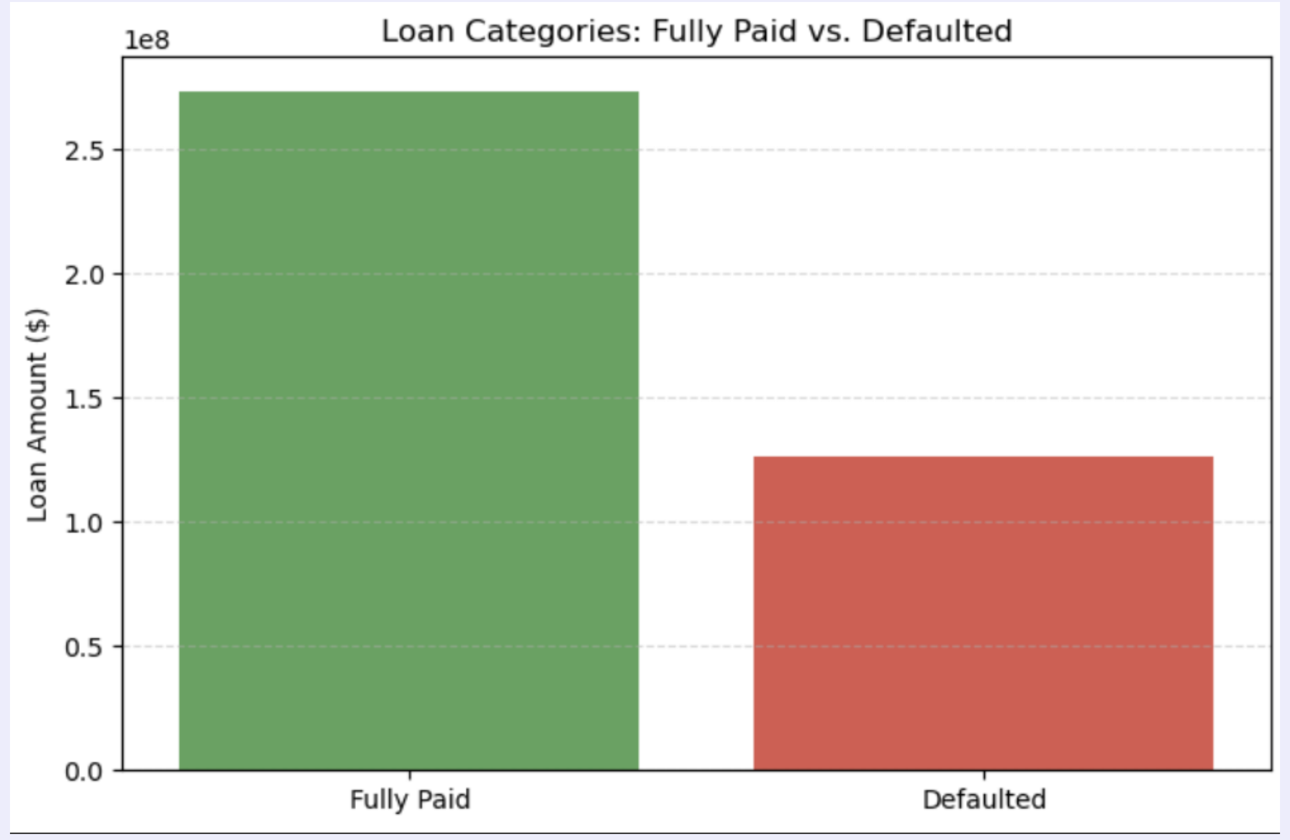
Optimal Threshold ≈ 0.45:

Where **precision ≈ recall**. This is the **best balance** between avoiding risky loans and not missing profitable ones.

Applying the Model to Open Loans

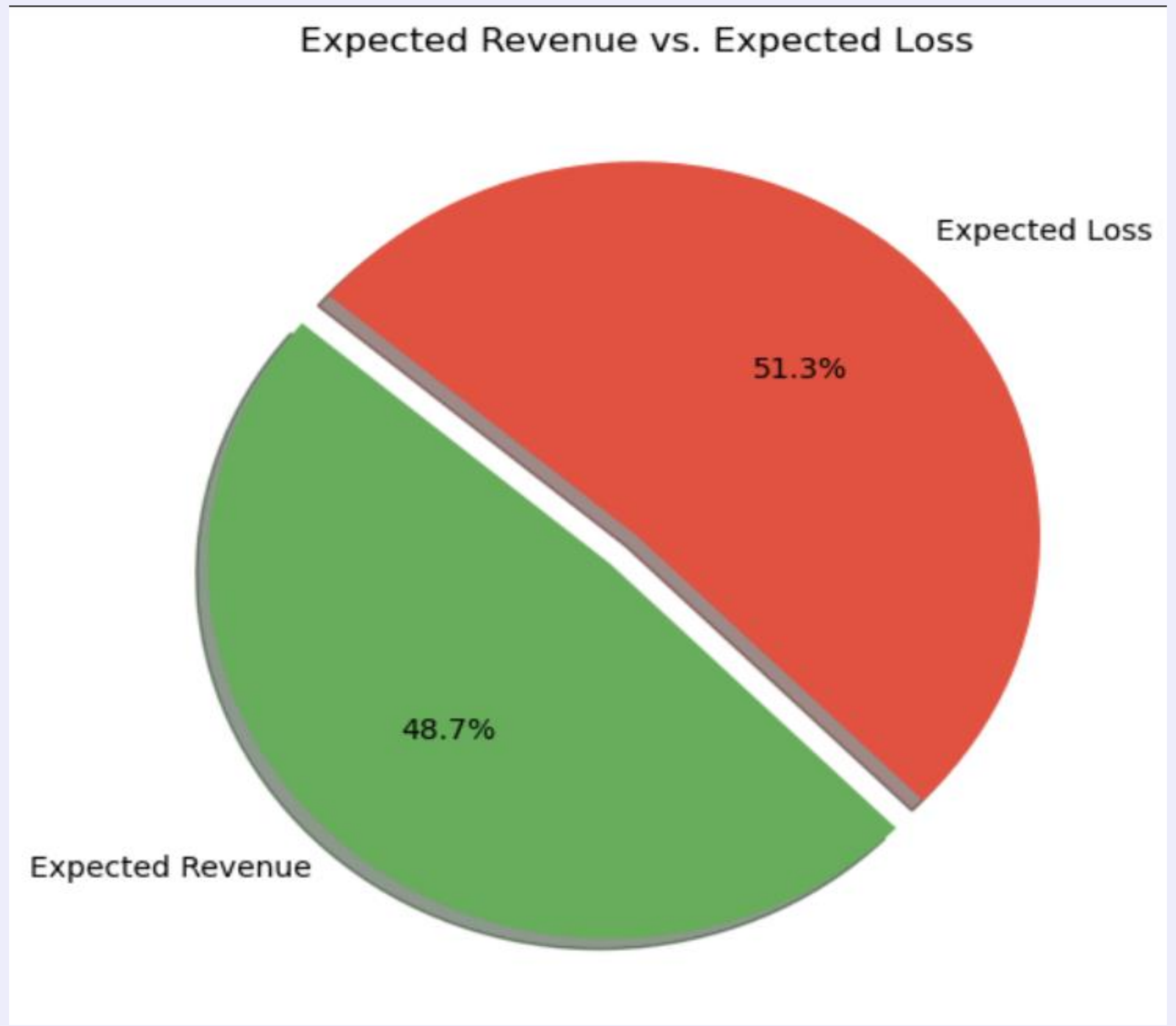
Open Loan Portfolio Impact (Post-Prediction):

- **Total open loans:** $399,902,000.00
- **Predicted Fully Paid:** 17,236 loans | $273,466,900.00
- **Predicted Defaults:** 6,606 loans | $126,435,100.00
- **Expected Revenue:** $32,603,035.95
- **Expected Loss (after recovery):** $34,390,347.20
- **Net Return:** $-1,787,311.25



Assumptions & Context:

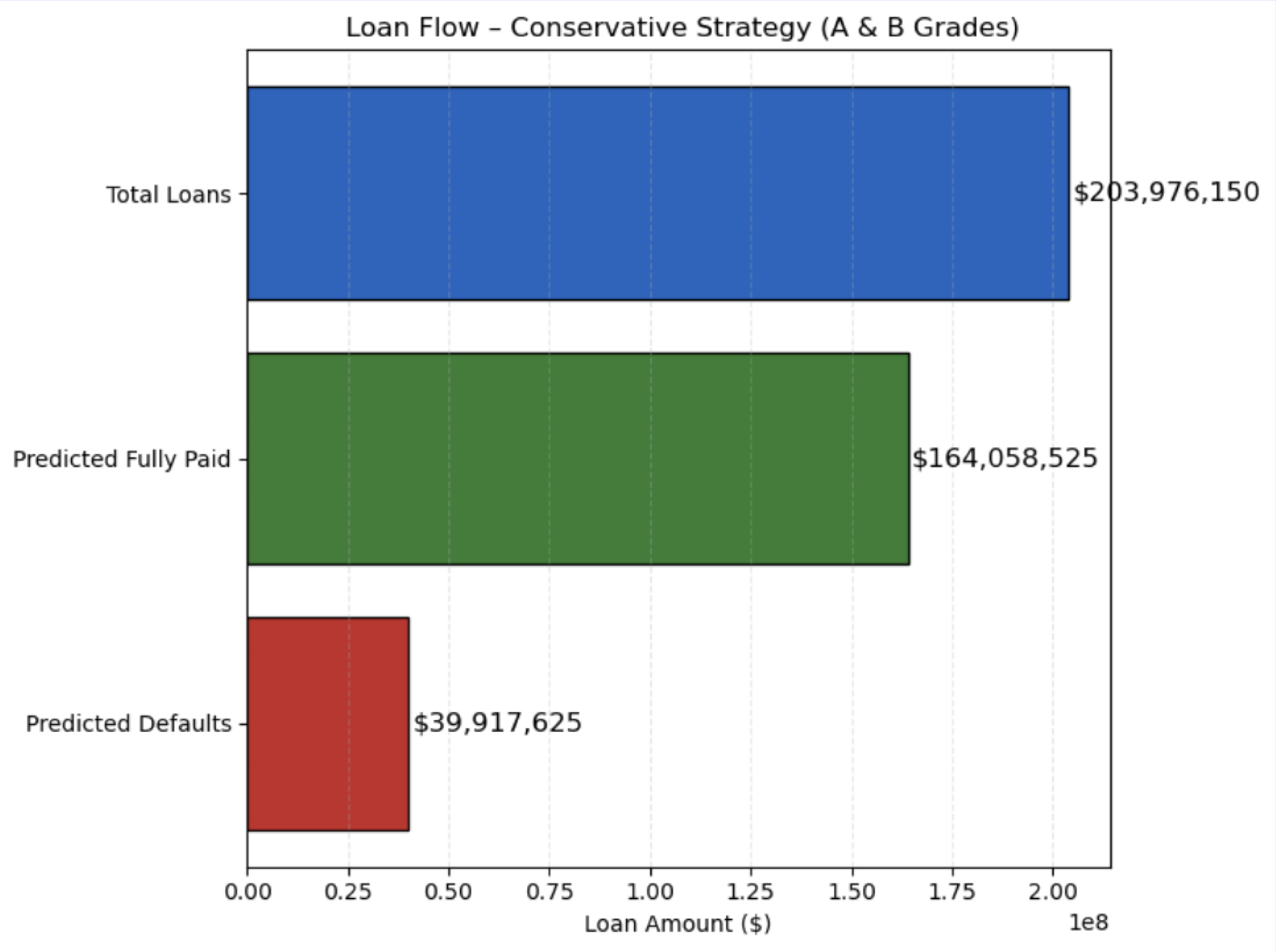
- Since the **2008 financial crisis**, loans are only issued as **A-grade first-lien loans**, meaning they are **secured by collateral** (e.g., properties, other assets, or guarantees from third parties).
- Based on research, the **expected recovery rate** for such loans in the **U.S. is 72.8%** (Source: **S&P Global**).



Conservative Strategy: Low Risk, Stable Returns

- **Target:** A & B grade loans (Grades 0, 1)
- **Total Loans:** 12,355
- **Total Loan Amount:** $203,976,150.00
- **Predicted Fully Paid:** 10,430 loans | $164,058,525.00
- **Predicted Defaults:** 1,925 loans | $39,917,625.00
- **Expected Revenue:** $14,465,143.69
- **Expected Loss (after recovery 72.8%):** $10,857,594.00
- **Net Return:** $3,607,549.69

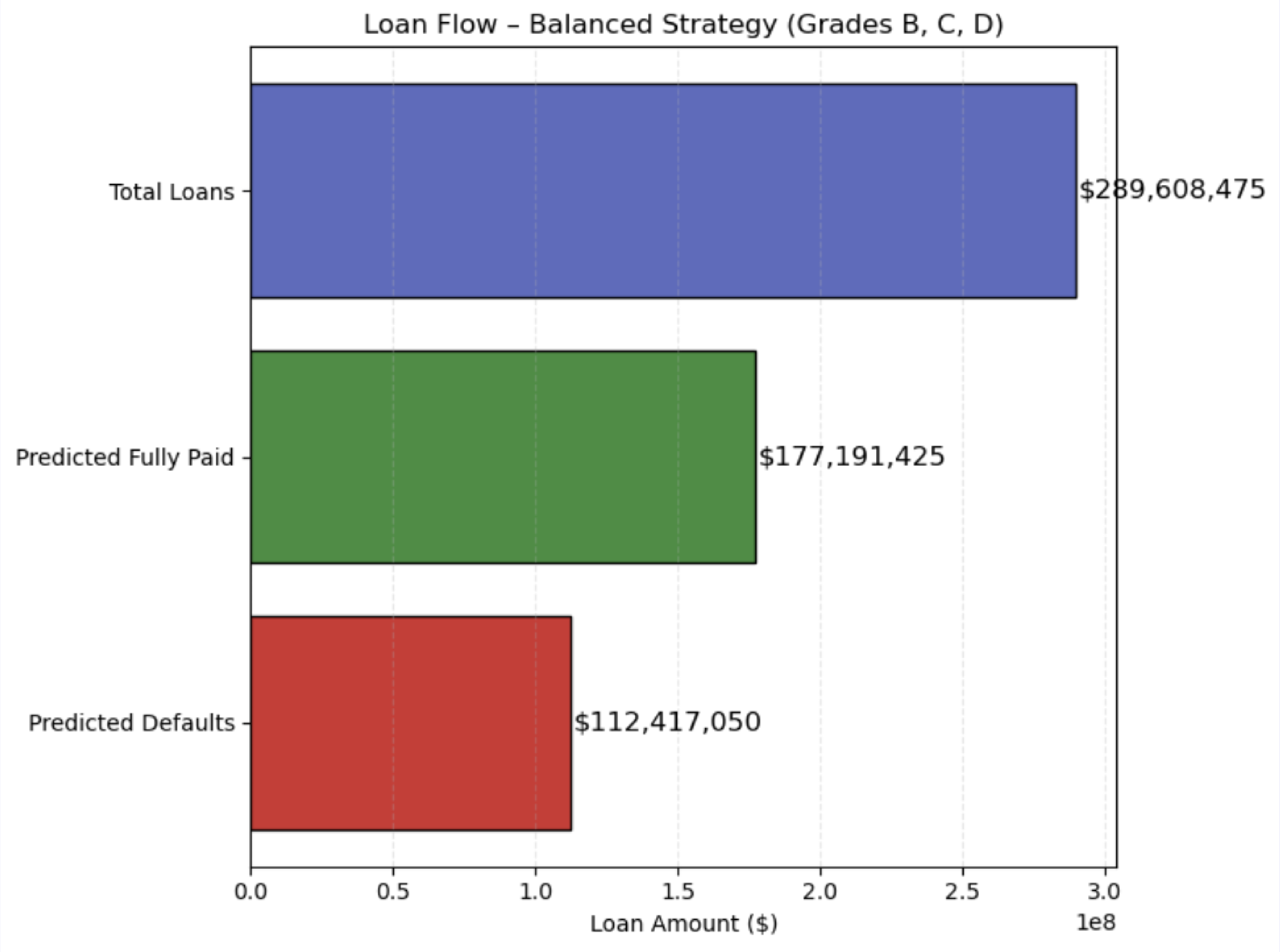
Conclusion: Stable and profitable, with minimal default risk.



Balanced Strategy: Moderate Risk, Optimized Returns

- **Target:** B, C, D grade loans (Grades 1, 2, 3)
- **Total Loans:** 17,129
- **Total Loan Amount:** $289,608,475.00
- **Predicted Fully Paid:** 11,318 loans | $177,191,425.00
- **Predicted Defaults:** 5,811 loans | $112,417,050.00
- **Expected Revenue:** $23,442,312.06
- **Expected Loss (after recovery 72.8%):** $30,577,437.60
- **Net Return: $-7,135,125.54**

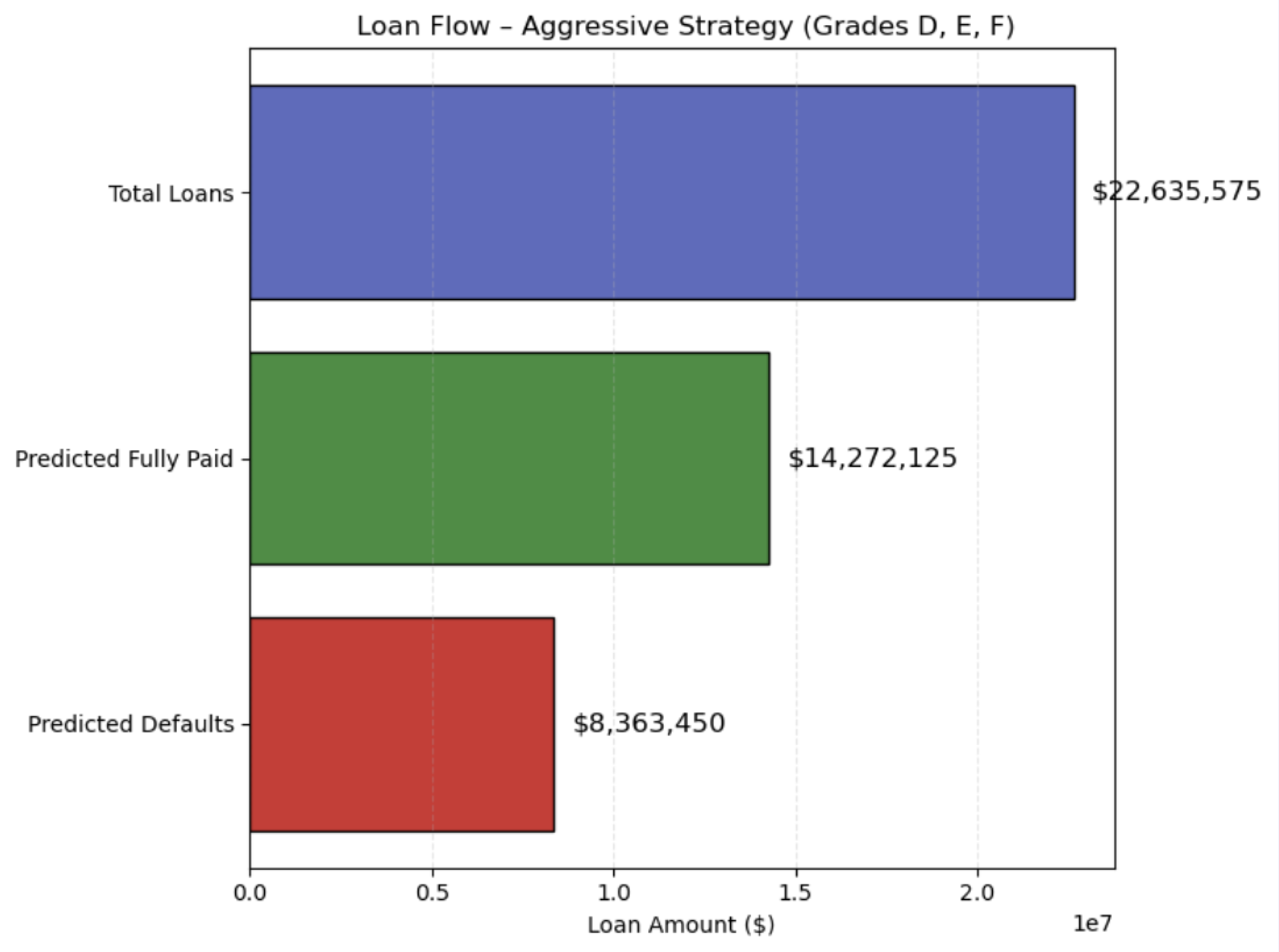
Conclusion: Higher revenue potential, but excessive losses result in negative net return.



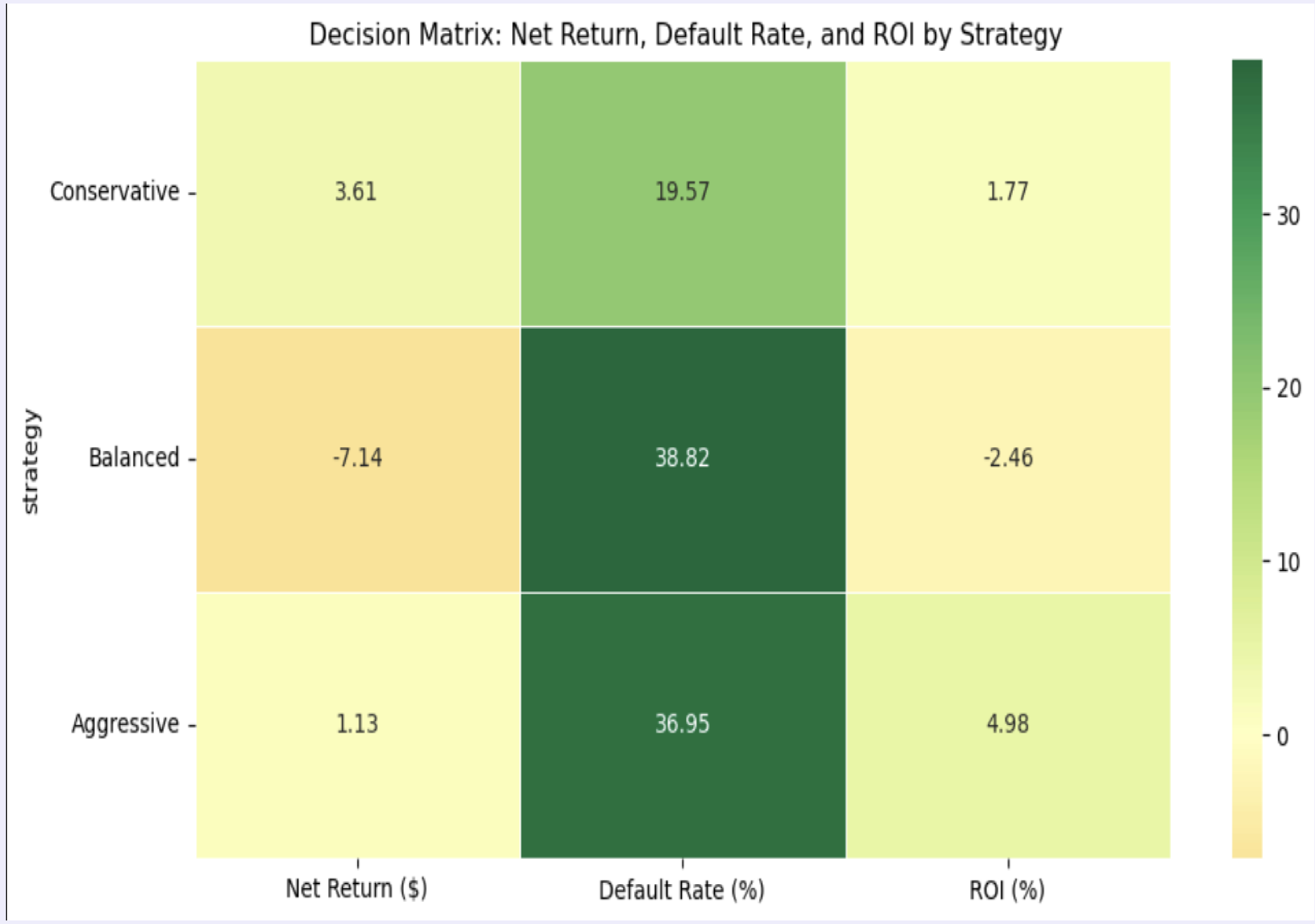
Aggressive Strategy: High Risk, Maximum Yield

- **Target:** D, E, F grade loans (Grades 4, 5, 6)
- **Total Loans:** 1,275
- **Total Loan Amount:** $22,635,575.00
- **Predicted Fully Paid:** 774 loans | $14,272,125.00
- **Predicted Defaults:** 501 loans | $8,363,450.00
- **Expected Revenue:** $3,401,057.34
- **Expected Loss (after recovery 72.8%):** $2,274,858.40
- **Net Return:** $1,126,198.94

Conclusion: Higher risk, but still profitable with a positive net return.



Analyzing KPI’s



1. **Conservative Strategy (Low Risk, Stable Returns)**

• **Net Return:** $3.61M (Highest profitability)

• **Default Rate:** 19.57% (Lowest risk)

• **ROI:** 1.77% (Stable but low)

• **Insight:** This strategy yields the most consistent and low-risk returns, making it ideal for risk-averse investors.

2. **Balanced Strategy (Medium Risk, Poor Returns)**

• **Net Return:** -$7.14M (Significant financial loss)

• **Default Rate:** 38.82% (Extremely high)

• **ROI:** -2.46% (Worst investment outcome)

• **Insight:** This strategy fails to optimize returns due to excessive default rates, leading to negative profitability.

3. **Aggressive Strategy (High Risk, Volatile Returns)**

• **Net Return:** $1.13M (Positive but lower than Conservative)

• **Default Rate:** 36.95% (Very high)

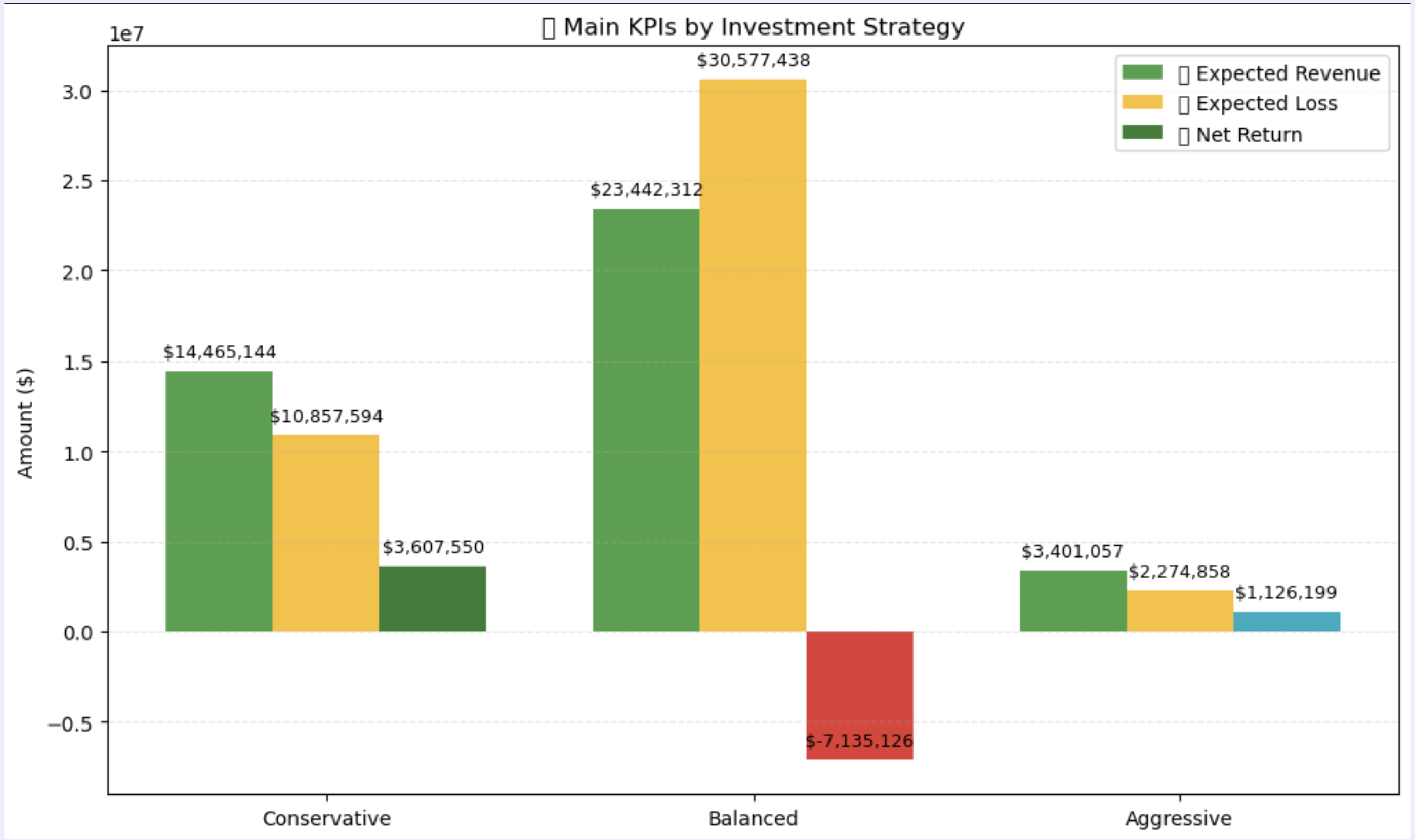
• **ROI:** 4.98% (Best return efficiency)

• **Insight:** While this strategy maximizes ROI, the high default rate threatens sustainability. It is only viable for high-risk investors.

Final Strategy Selection & Business Recommendation

Best Strategy: Conservative Strategy (Grades A & B)

- **Why?** Most stable return, positive net profit, and lowest default risk.
- **Balanced and Aggressive** strategies are not sustainable due to high default rates.
- **Optimization:** Focus on A & B loans while testing small portions of C grade loans to optimize profit.



esade

Do Good. Do Better.