# End-to-End Credit Risk Modeling and Dashboarding for ID/X Partners

Dwi Budi Setyonugroho
Data Scientist Intern
ID/X Partners
Email: setyonugrohodwibudi@gmail.com

***Abstract*—This paper presents a complete and production-oriented credit risk analytics solution for ID/X Partners based on a historical loan portfolio (2007–2014). The work addresses three core challenges in lending analytics: severe class imbalance, high missingness/noise in raw operational data, and data leakage from post-origination variables. We design a rigorous preprocessing pipeline that reduces the initial 466,285-record dataset into a high-confidence, leakage-free modeling table of 238,913 records. A Balanced Logistic Regression model is selected as the primary classifier after comparative evaluation against baseline alternatives, prioritizing risky-loan recall to align with financial-loss minimization objectives. The optimized model improves risky-class recall from 8% to 66% and F1-score from 0.14 to 0.45. To operationalize insights for non-technical stakeholders, a two-page Power BI dashboard is developed to monitor portfolio health, regional outliers, and risk-segment contribution. The resulting framework demonstrates how disciplined data governance, interpretable modeling, and business intelligence integration can jointly improve credit risk decision quality.**

***Index Terms*—Credit risk, default prediction, class imbalance, data leakage, logistic regression, feature engineering, business intelligence, Power BI.**

## I. Introduction

Credit risk management is central to lending profitability and portfolio stability. Misclassification of high-risk applicants leads directly to non-performing loans and elevated loss reserves, while overly conservative approval strategies reduce revenue opportunities. For ID/X Partners, the core business need is to identify potential defaults as early as the application stage, using only information available at loan origination.

Real-world lending data presents practical difficulties that often degrade model reliability: incomplete fields, target ambiguity, inconsistent metadata, and hidden leakage from outcome-linked variables. In addition, default events typically represent a minority class, causing conventional models to over-optimize global accuracy while under-detecting true risk.

This project develops an end-to-end pipeline that combines data quality triage, leakage-safe feature engineering, class-imbalance-aware modeling, interpretability analysis, and executive dashboarding. The objective is not merely predictive performance in isolation, but measurable business value through better detection of financially risky loans.

## II. Business Problem and Project Objectives

### A. Business Context

ID/X Partners operates in a lending environment where undetected defaults create substantial downstream loss. Historical lending records reveal a material default segment (approximately 22%) that must be identified during underwriting. The company needs an approach that is both technically sound and deployable by risk teams.

TABLE I
Raw Data Summary

| Attribute | Value |
|---|---|
| Records (initial) | 466,285 |
| Columns (initial) | 75 |
| Time horizon | 2007–2014 |
| Target field | `loan_status` |
| Target mapping | Safe vs Risky |

### B. Problem Statement

The project addresses two high-impact risks:

1) **Class Imbalance Bias**: Standard classifiers favor majority "Safe" outcomes, producing low risky-class recall.
2) **Data Leakage**: Post-origination variables (payments, recoveries, outstanding principal) can inflate offline metrics and fail in production.

### C. Project Objectives

The solution is designed to:

- Build a leakage-free supervised learning pipeline using application-time features only.
- Optimize risky-loan detection (Recall/F1) rather than raw accuracy.
- Provide interpretable risk drivers for policy and underwriting decisions.
- Deliver an interactive dashboard for executive monitoring and analyst deep dive.

## III. Data Understanding and Dictionary Analysis

### A. Source Data

The project uses two provided assets: `LoanStats.csv` and `LCDataDictionary.xlsx`. Initial dataset size is 466,285 rows by 75 columns.

### B. Critical Dictionary Findings

Initial schema review surfaced notable inconsistencies:

- **Missing FICO fields**: fico_range_low and fico_range_high were documented but absent in the dataset, likely due to privacy controls.
- **Description mismatches**: selected dictionary descriptions required validation against actual distributions.
- **Leakage indicators**: variables such as `total_pymnt`, `recoveries`, and `out_prncp` encode post-origination outcomes.

### C. Data Quality Issues

Three classes of quality problems were observed:

1) **Structural missingness**: 17 columns were fully empty.
2) **High missingness**: multiple columns exceeded 50% null values.
3) **Ambiguous labels**: statuses like "Current" and "In Grace Period" do not provide final repayment outcomes.

## IV. Preprocessing and Data Engineering Methodology

### A. Initial Assessment

Raw input consists of 466,285 records with heterogeneous completeness. Early profiling focused on identifying columns that were either non-informative (empty), unstable (high missingness), or unavailable at prediction time (leakage).

### B. Cleaning Pipeline

The cleaning strategy follows conservative principles for production reliability:

- Removed 22 empty/noisy columns to reduce dimensional noise.
- Filtered target classes to completed outcomes only:
  - Excluded: Current, In Grace Period, Late ($< 31$ days).
  - Retained: Fully Paid, Charged Off, Default.
- Applied robust imputation:
  - Numeric features: median imputation.
  - Categorical features: mode imputation.
- Standardized `emp_length` values before imputation (e.g., "10+" and "<1" conversion).

### C. Leakage Prevention as a Mandatory Constraint

To simulate true underwriting conditions, all post-origination variables were removed. Key excluded fields are total_pymnt, recoveries, out_prncp, last_pymnt_d, and total_rec_int. This design prevents "future information" contamination and protects model validity during deployment.

### D. Final Engineered Dataset

After filtering and cleaning, the dataset becomes 238,913 rows with 32 features and a risky-class prevalence of 21.84%.

TABLE II
Data Engineering Output

| Metric | Raw | Cleaned |
|---|---|---|
| Rows | 466,285 | 238,913 |
| Columns | 75 | 32 |
| Label certainty | Mixed | High confidence |
| Leakage status | Present | Removed |
| Default rate | N/A | 21.84% |

TABLE III
Baseline vs Optimized Model Performance

| Metric | Baseline | Optimized |
|---|---|---|
| Accuracy | 79.0% | 64.2% |
| Recall (Risky) | 8.0% | 66.0% |
| F1-Score (Risky) | 0.14 | 0.45 |
| Precision (Risky) | 56.0% | 34.0% |

## V. Modeling Strategy

### A. Baseline and Optimization Path

A baseline Logistic Regression model is established to quantify default class under-detection under standard class weighting. Because business impact is dominated by missed risky loans, optimization emphasizes risky-class Recall and F1-score.

### B. Class-Imbalance Handling

A class-balanced learning setup is adopted to increase minority-class sensitivity. This adjustment intentionally accepts lower global accuracy in exchange for materially higher default capture.

### C. Model Candidates

Two model families were compared in the project workflow:

- Balanced Logistic Regression (primary candidate)
- Balanced Random Forest (alternative benchmark)

Balanced Logistic Regression delivered stronger minority-class F1 in this dataset and superior interpretability for stakeholder communication.

## VI. Results and Performance Evaluation

### A. Quantitative Performance

Table III reports final metrics. Figure 1 compares discriminative behavior between the baseline and balanced Logistic Regression models across threshold settings. Figure 2 highlights the precision–recall trade-off of the optimized classifier under class imbalance. Figure 3 summarizes class-level prediction outcomes for the balanced model.

The ROC curves are nearly overlapping with an AUC of approximately 0.71 for both models, indicating similar overall rank-order discrimination. This result shows that class balancing does not materially change global separability, but it supports better operating-point behavior for risky-loan capture when combined with threshold and class-weight decisions.

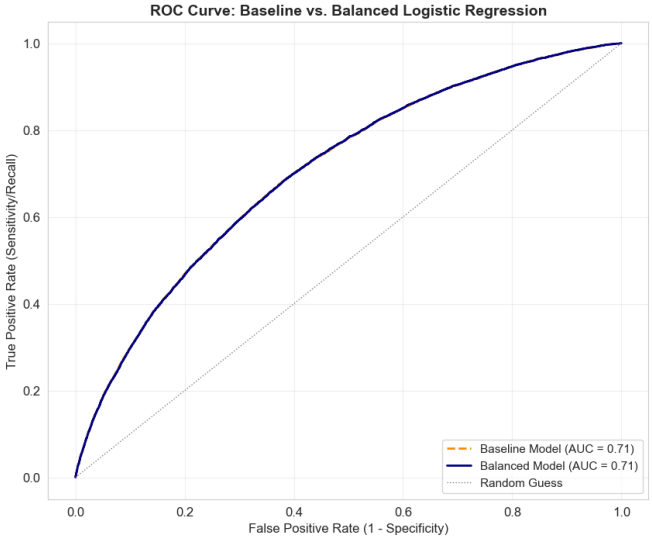


Fig. 1. ROC Curve: Baseline vs. Balanced Logistic Regression.

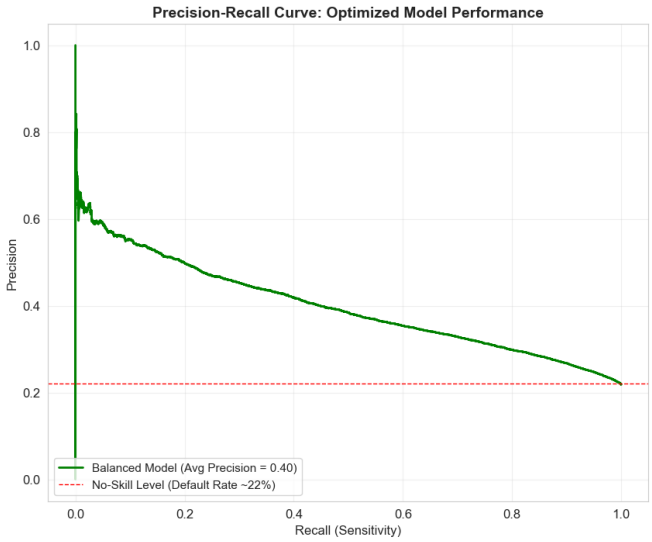


Fig. 2. Precision–Recall Curve: Optimized Model Performance.

In the imbalanced setting, the Precision–Recall view is more decision-relevant than ROC alone. The optimized model achieves an average precision around 0.40, clearly above the no-skill prevalence baseline of about 0.22, confirming meaningful lift in minority-class identification.

The confusion matrix indicates 6,936 true-risk detections against 3,501 missed risky cases, corresponding to recall near 66%. This pattern aligns with the project objective: increase risky-loan detection while accepting higher false positives (13,610) as a deliberate risk-control trade-off.

### B. Business Interpretation of Trade-offs

Although optimized accuracy is lower, the model increases risky-loan detection by 58 percentage points. This shift is operationally desirable in risk contexts where false negatives (missed bad loans) are significantly more expensive than false positives.

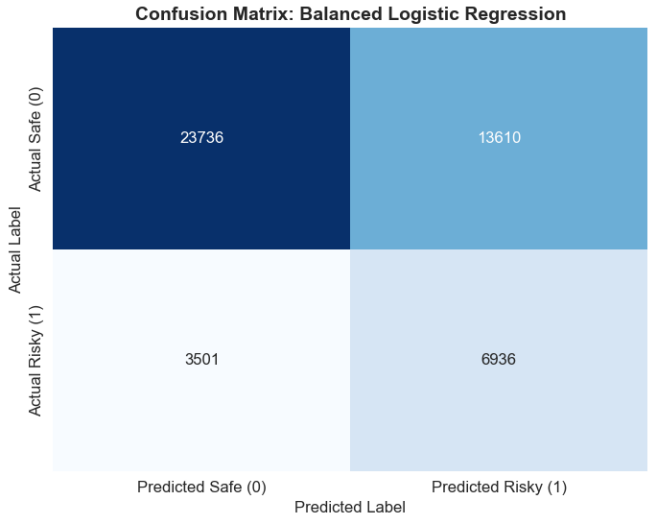


Fig. 3. Confusion Matrix: Balanced Logistic Regression.

### C. Model Selection Rationale

Balanced Random Forest underperformed in this balanced setting (project benchmark F1 lower than Logistic Regression), while Logistic Regression remained more stable and interpretable. The final model therefore balances performance, transparency, and policy usability.

## VII. Model Interpretation and Risk Drivers

### A. Coefficient Interpretation Framework

In Logistic Regression:

- Positive coefficients increase default likelihood.
- Negative coefficients decrease default likelihood.

Larger absolute values indicate stronger contribution.

### B. Top Risk Features

The strongest positive-risk contributors include grade levels E/F/D/G/C, followed by `purpose_small_business` and selected state indicators (e.g., Mississippi). A negative coefficient for Wyoming indicates a relative protective signal. Figure 4 visualizes the ranked feature effects used for business interpretation, where the magnitude and sign of coefficients directly indicate risk direction and relative influence.

The chart confirms that higher grade-risk buckets and small-business purpose are dominant positive-risk signals, while Wyoming appears as a protective factor with a negative coefficient.

### C. Policy-Level Implications

These signals support immediate underwriting strategy:

1) Re-price or cap high-risk grade segments (D–G).
2) Introduce enhanced documentation/cash-flow checks for small-business purpose loans.

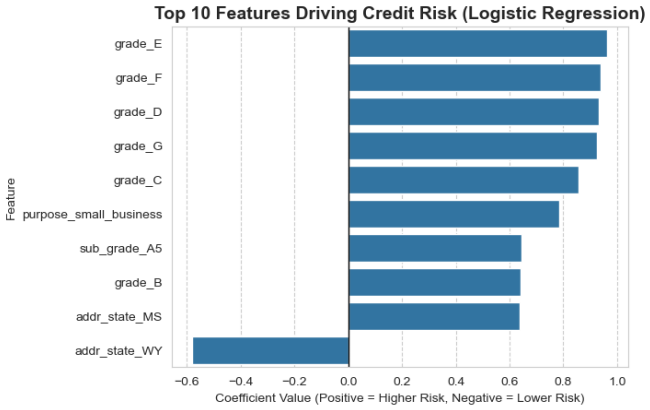


Fig. 4. Top 10 Features Driving Credit Risk (Logistic Regression).

TABLE IV
ILLUSTRATIVE HIGH-IMPACT FEATURES

| Rank | Feature | Coefficient |
|---|---|---|
| 1 | Grade E | +0.96 |
| 2 | Grade F | +0.94 |
| 3 | Grade D | +0.93 |
| 4 | Grade G | +0.92 |
| 5 | Grade C | +0.86 |
| 6 | Purpose: Small Business | +0.78 |
| 7 | State: Mississippi | +0.64 |
| 8 | State: Wyoming | -0.58 |

3) Monitor regional exposure and dynamically adjust policy thresholds.

## VIII. Power BI Dashboard for Stakeholder Enablement

### A. Architecture and Data Consistency

The dashboard is connected to loan_data_cleaned.csv, ensuring consistency with the model training dataset and leakage-safe preprocessing rules.

### B. Page 1: Executive Portfolio Summary

The executive page includes KPI cards and macro trend monitoring:

- Total Loans: 239K
- Total Funded Amount: $3B
- Overall Default Rate: 21.84%
- Average Interest Rate: 12.28%

It also includes trend lines, geographic heat maps, grade-default gradients, and DTI distribution comparison.

### C. Page 2: Risk Driver and Segment Analysis

Analyst-oriented controls include slicers for grade, term, verification status, and home ownership. Segment matrix analysis highlights Grade C as a major estimated-loss contributor due to the interaction between volume and risk rate.

### D. Operational Value

The dashboard transforms analysis into decision support by answering three questions in one workflow: what changed, why it changed, and where intervention should be prioritized.

## IX. Limitations and Future Recommendations

### A. Primary Limitation: Missing Raw FICO

Documented FICO fields were unavailable in the delivered data. As a result, the model uses grade proxies, which likely imposes a performance ceiling and limits within-grade granularity.

### B. Recommended Implementation Roadmap

1) Integrate live bureau data (including raw FICO) at application time.
2) Retrain model with expanded risk signals and recalibrate decision thresholds.
3) Apply threshold tuning by business cost function (default loss vs rejection cost).
4) Deploy drift monitoring and scheduled retraining governance.
5) Establish targeted policy controls for small business applications.

## X. Conclusion

This project demonstrates a complete applied data science lifecycle for lending risk management: data dictionary verification, rigorous cleaning, leakage prevention, class-imbalance-aware modeling, interpretation, and BI deployment. The optimized classifier substantially improves risky-loan detection (8% to 66% recall) and provides transparent drivers for policy action. Combined with dashboard-based monitoring, the solution provides ID/X Partners with a practical baseline for production risk screening and a clear path for further gains through richer credit bureau integration.