Credit Default Risk Analysis

Drivers of default using balance, income and student status

Date: 30 January 2026

Prepared by: RITA AKUMU

Contents

1. Executive Summary

Key takeaways

- Default prevalence is low (333 out of 10,000 customers) - class imbalance means accuracy alone is misleading.
- Balance is the strongest driver of default risk; risk increases sharply at higher balance bands.
- Students show a higher default rate than non-students (~4.3% vs ~2.9%).
- Baseline logistic regression achieved ROC-AUC = 0.951; at threshold 0.50: recall = 0.89, precision = 0.182.

This report analyzes credit default behavior using customer balance, income, and student status. The goal is to (1) identify which factors are most strongly associated with default, (2) quantify how default risk varies across customer segments, and (3) build a baseline predictive model that estimates the probability of default. The findings are translated into practical recommendations for screening, credit policy, and portfolio monitoring.

The dataset contains 10,000 customers and 5 variables (ID, default, student, balance, income). Default is rare (3%; 333 out of 10,000), meaning accuracy alone can be misleading (predicting “no default” for everyone would still achieve ~97% accuracy). Because of this class imbalance, the analysis emphasizes risk segmentation and uses metrics such as precision and recall in the modeling section.

Key insights show that balance is the strongest driver of default risk, with default rates increasing sharply in the highest balance bands. Students default at a higher rate than non-students (~4.3% vs ~2.9%). Income shows a generally inverse relationship with default, but the relationship appears weaker than balance in this dataset.

A baseline logistic regression model (with class weighting to address imbalance) achieved ROC-AUC = 0.951, indicating strong ability to rank customers by risk. At a 0.50 threshold, the model produced recall = 0.89 and precision = 0.182, which reflects an operational trade-off: catching most defaulters while generating many false positives. These results support using a risk score and risk bands to guide monitoring and credit policy.

2. Context and Use Cases

2.1 Why default risk matters

Default risk matters because it directly affects profitability and cash flow. When a customer defaults, the lender loses principal and interest, incurs recovery/collections costs, and may need to increase provisions for expected losses. High default rates also force stricter lending rules, which can reduce growth and customer acquisition.

Operationally, default risk drives workload and efficiency. More defaults mean more accounts to manage, more follow-ups, higher collections activity, and higher cost-to-serve. If risk is identified early, teams can focus effort on accounts that actually need intervention rather than treating everyone the same.

Strategically, measuring default risk protects the portfolio. Understanding which factors (like high balances) increase risk helps prevent concentration in risky segments and supports consistent, data-backed credit policy decisions.

2.2 How to use this analysis operationally

- Risk screening - estimate default likelihood for approve/decline decisions or conditional approvals.
- Credit limit setting - adjust exposure for higher-risk segments (e.g., high balance relative to income).
- Ongoing monitoring - flag accounts whose risk is increasing to prioritize early interventions.

3. Problem Statement and Objectives

Build and evaluate a baseline credit risk model using balance, income, and student status to predict default, then translate results into actionable recommendations for risk management.

3.1 Objectives

- Quantify overall default prevalence and class imbalance
- Identify how default risk varies by balance bands, income bands, and student status
- Compare defaulters vs non-defaulters on key predictors
- Train a baseline model and evaluate using AUC, precision, recall, and threshold trade-offs
- Recommend practical risk actions based on findings

4. Hypotheses and Analytical Questions

H1: Balance is a key driver of default

- Null (H0): Average balance is the same for defaulters and non-defaulters; balance is not associated with default.
- Alternative (H1): Defaulters have higher balances; balance is positively associated with default risk.

H2: Income is protective (higher income → lower default)

- H0: Income is not associated with default.
- H1: Higher income is associated with lower default probability.

H3: Student status affects default rate

- H0: Default rate is the same for students and non-students.
- H1: Default rate differs between students and non-students.

H4: Student effect after controlling for balance and income

- H0: After controlling for balance and income, student status has no effect on default.
- H1: Student status remains a significant predictor even after controlling for balance and income.

H5: Balance explains default better than income

- H0: Balance and income have equal predictive power.
- H1: Balance has stronger predictive power than income.

5. Data Overview and Preparation

5.1 Dataset overview

The dataset was obtained from the official Kaggle Website. It has 5 columns and 10,000 rows. The columns are the identifier which Is a unique number, whether the person is a defaulter or not, this is classified as either yes or no, whether one is a student or not, also classified as yes or no, and the income of the person and the balance of the account.

The dataset has no missing values or duplicates, and I will assume that the income and balance are in the United States Dollar (USD)

5.2 Data preparation

In order to perform proper analysis, I encoded row 2 and 3, (defaulter and student) which were answered in yes/ no into 1 and 0

The income and balances had no outliers, and all were around the same range.

6. Exploratory Data Analysis

There are no missing values or duplicates in any of the columns.

The data ranges are valid and there are no outliers that could distort the averages.

Balance distribution: The median balance is 823.64 and the mean is 835.37, suggesting balances are not heavily skewed. The middle 50% of customers have balances between 481.73 (P25) and 1,166.31 (P75). Only 10% of customers exceed 1,471.63 (P90), and only 1% exceed 2,008.47 (P99). These percentiles are used to define balance risk bands and evaluate how default rates change across balance levels.

Income distribution: The median balance is 34,552.64 and the mean is 33,516.98, this suggests that the income is slightly negatively skewed. The middle 50% of customers have incomes between 21,340.46(P25) and 43,807 (P75). Only 10% of customers exceed 50,766.15(P90) and only 1% exceed 61,656.74 (P99)

Balance shows a slight right skew (skewness = 0.25), suggesting a small number of customers have higher balances than the rest. Income is approximately symmetric (skewness = 0.07), indicating incomes are more evenly distributed without strong tail effects.

The standard deviation of balance is 483.71 versus a mean of 835.37 (CV ≈ 58%), indicating relatively high dispersion in customer balances. Income has a standard

deviation of 13,336.64 versus a mean of 33,516.98 (CV ≈ 40%), suggesting income is also dispersed but more stable relative to its mean than balance.

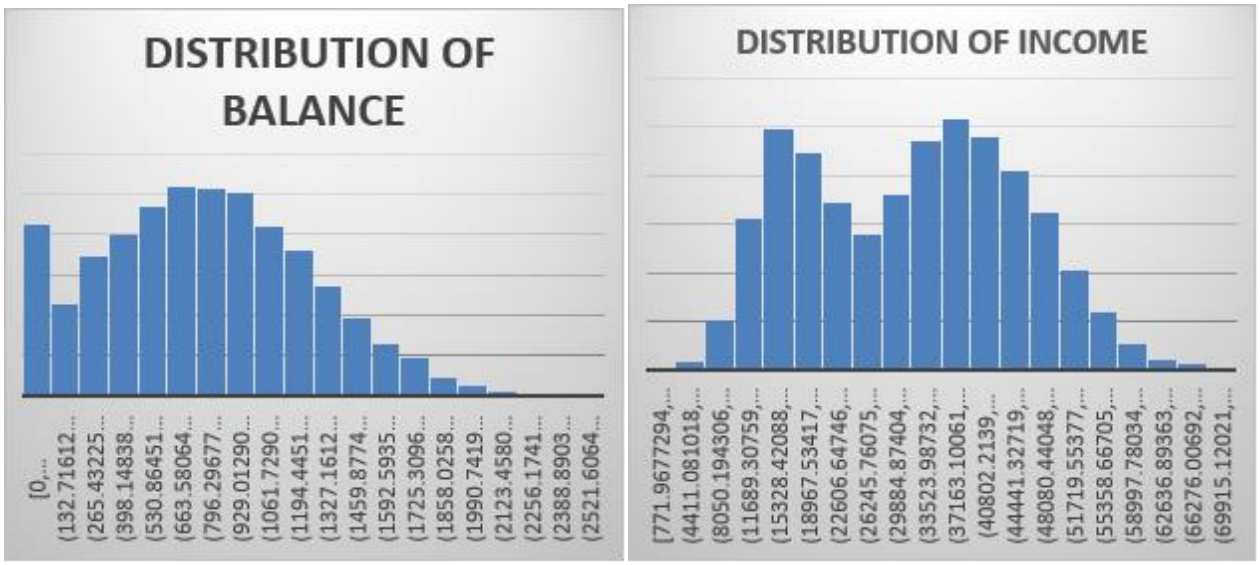

Figure: Distribution of balance and income (as provided in the source report).

6.1 Overall default prevalence

The dataset shows an overall default rate of 3% (333 out of 10,000 customers).

This indicates that the rate of default is low but the risk impact of the default is high. Because defaults are relatively rare, predicting no default can be misleading as it would achieve 97% accuracy. This analysis focuses on risk segmentation and drivers of default, and uses precision/recall and threshold-based evaluation to support credit decisions.

Default status	Count	Share of portfolio
No default	9,667	97%
Default	333	3%
Total	10,000	100%

6.2 Default rate by student status

Students show a higher default rate (~4.3%) compared to non-students (~2.9%). This suggests students are more likely to default in this sample

6.3 Default rate by income bands

Default probability generally decreases with income. Minor deviations appear in upper percentiles due to smaller sample sizes and higher financial risk-taking among high-income individuals. The top 1% shows the lowest default rate, confirming the overall inverse relationship between income and default.

6.4 Default rate by balance bands

Balance reflects outstanding debt/exposure, so higher balances increase repayment burden and are associated with higher default risk. In this dataset, default rates rise sharply across higher balance bands, indicating balance is a strong driver of default and useful for risk segmentation.

6.5 Balance and income by default group

Customers who defaulted had a substantially higher average balance (1,747.82) compared to non-defaulters (803.94), indicating balance is strongly associated with default risk. Average income was slightly lower among defaulters (32,089) than non-defaulters (33,566), suggesting income has a weaker relationship with default in this sample. This suggests default risk is more closely linked to exposure/outstanding debt than to income alone, and supports using balance bands (and ideally balance-to-income) for risk segmentation.

Default status	Avg. balance	Avg. income
No default	803.94	33,566.17
Default	1,747.82	32,089.15

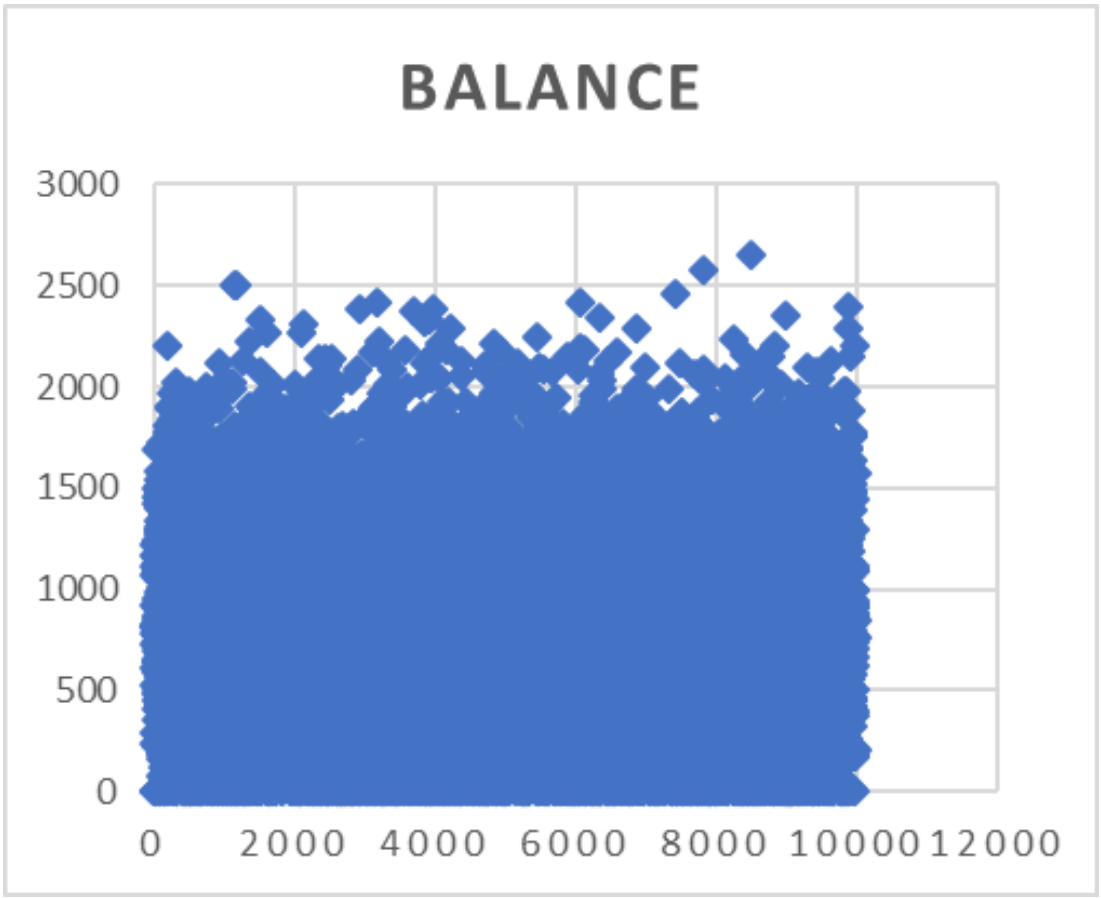

Figure: Balance scatter plot (as provided in the source report).

7. Modeling and Evaluation

A baseline logistic regression model was trained to predict default using student status, balance, and income, with class imbalance handled via class weighting. On the test set,

the model achieved ROC-AUC = 0.951, indicating strong ability to rank customers by default risk. Using a 0.50 decision threshold, the model achieved recall = 0.89, correctly identifying 89 out of 100 defaulters, but with precision = 0.182 due to 400 false positives (non-defaulters flagged as risky). This indicates the model is effective for risk detection and monitoring, while threshold tuning is required to balance missed defaults against the operational cost of reviewing flagged accounts.

The model outputs a probability of default for each customer. A decision threshold was tested to convert probabilities into default/non-default predictions. Lower thresholds increase recall (capturing more true defaulters) but produce more false positives, while higher thresholds increase precision (fewer false positives) but miss more defaulters. In this dataset, increasing the threshold from 0.40 to 0.60 raises precision from 0.152 to 0.222 while reducing recall from 0.90 to 0.87, illustrating the operational trade-off between catching risky accounts and minimizing unnecessary reviews.

7.1 Threshold trade-offs

Threshold	Precision	Recall	TP	FP	FN	TN
0.05	0.066	1.00	100	1,426	0	1,474
0.10	0.080	1.00	100	1,144	0	1,756
0.15	0.094	0.99	99	950	1	1,950
0.20	0.104	0.97	97	838	3	2,062
0.25	0.117	0.96	96	723	4	2,177
0.30	0.129	0.94	94	632	6	2,268
0.40	0.152	0.90	90	501	10	2,399
0.50	0.182	0.89	89	400	11	2,500
0.60	0.222	0.87	87	305	13	2,595

Model performance note

- ROC-AUC of 0.951 indicates strong discrimination (ranking) between defaulters and non-defaulters.
- Precision is low at common thresholds due to class imbalance; operationally, this creates false positives and requires triage via risk bands.

8. Findings and Implications

Insight: Default is rare overall (~3%).

- Evidence: 333 defaults out of 10,000 customers.
- Implication: Use recall/precision (not accuracy) if modeling; focus on identifying high-risk segments.

Insight: Default risk increases sharply with balance.

- Evidence: Default rate rises from near 0% in low bands to ~21% in band 5 and ~76% in band 6.
- Implication: Balance bands are a strong basis for monitoring/credit limits.

Insight: Students have a higher default rate than non-students.

- Evidence: ~4.3% vs ~2.9%.
- Implication: Student status may be a useful segmentation variable (verify within balance bands).

Insight: Balance differentiates defaulters more than income.

- Evidence: Avg balance: 1,747 (defaulters) vs 804 (non); income difference is small.
- Implication: Prioritize balance (and balance-to-income) in risk rules.

9. Recommendations

- Credit limit policy: Set tighter limits or require review for customers entering high balance bands (5-6).
- Monitoring: Create an “early warning list” for customers whose balance moves from band 4 → 5.
- Collections prioritization: Prioritize outreach to band 5-6 accounts (highest expected risk).
- Better affordability metric: Track balance-to-income ratio to refine risk beyond balance alone.

10. Limitations and Next Steps

10.1 Limitations

- Only 3 predictors (no credit history, tenure, delinquencies, loan amount).
- Default is imbalanced (3%) → accuracy can be misleading.
- Correlation/association ≠ causation.

10.2 Next steps

- Add more predictors

- Try alternative models (tree-based)
- Calibrate probabilities
- Cost-sensitive thresholding