Lesson 2.2 Exercises: Credit Scoring — From FICO to ML

Module 2: The Access Problem

Prof. Dr. Joerg Osterrieder

Digital Finance — BSc Course

Exercise 1: WoE and IV Calculation

Scenario: A credit analyst is evaluating the feature "Years at Current Employer" for inclusion in a scorecard. The feature has been binned into three groups (all values synthetic):

Bin	# Goods	# Bads	Distr. Goods	Distr. Bads
$<$ 2 years	300	150	0.30	0.50
2–5 years	400	100	0.40	0.333
$>$ 5 years	300	50	0.30	0.167
Total	**1,000**	**300**	**1.00**	**1.00**

Tasks:

a. Calculate the WoE for each bin.

b. Calculate the IV for this feature.

c. Based on Siddiqi's IV thresholds, classify the predictive power of this feature.

d. Explain intuitively why the bin "$<$ 2 years" has a negative WoE.

Difficulty: Intermediate — requires WoE/IV formulae and interpretation.

Exercise 2: Expected Loss Across a Loan Portfolio

Scenario: A bank has a small consumer loan portfolio with three risk grades (synthetic data):

Grade	# Loans	Avg. EAD	PD	LGD
A (Low risk)	5,000	\$15,000	1.0%	30%
B (Medium risk)	3,000	\$12,000	4.0%	40%
C (High risk)	1,000	\$8,000	10.0%	55%

Tasks:

a. Calculate the Expected Loss (EL) for each risk grade.

b. Calculate the total portfolio Expected Loss.

c. Express the total EL as a percentage of total portfolio exposure.

d. If the bank prices loans to cover EL plus a 2% profit margin, what minimum interest rate spread over funding cost should Grade C borrowers pay?

Difficulty: Intermediate — requires EL formula and portfolio-level thinking.

Exercise 3: Comparing Two Credit Models

Scenario: A FinTech lender tests two credit scoring models on a holdout dataset of 10,000 applications (3% default rate). Results (synthetic):

Model	AUC	KS Statistic	Top-Decile Default Rate
Logistic Regression	0.74	0.38	9.2%
Gradient Boosting (XGBoost)	0.81	0.48	14.5%

Tasks:

a. Calculate the Gini coefficient for each model.

b. Calculate the lift in the top decile for each model (population default rate = 3%).

c. The lender currently declines the top-risk 20% of applicants. Estimate how many additional defaults would be caught by switching from the logistic model to the XGBoost model.

d. The regulator requires human-readable decline reasons. Propose a strategy that uses the XGBoost model for decisioning while satisfying this requirement.

Difficulty: Advanced — requires metric computation and strategic thinking.

Exercise 4: Calibration Assessment

Scenario: A credit model assigns predicted PDs to five score bands. After 12 months, the actual default rates are observed (synthetic data):

Score Band	# Borrowers	Predicted PD	Observed Default Rate
700–750	2,000	1.0%	1.2%
650–699	3,000	3.0%	3.1%
600–649	2,500	6.0%	7.8%
550–599	1,500	10.0%	14.5%
< 550	500	18.0%	25.0%

Tasks:

a. For each score band, calculate the ratio of observed to predicted default rate.

b. Is this model under-predicting or over-predicting default risk? For which score bands is the miscalibration most severe?

c. The bank priced all loans based on the predicted PDs. Calculate the total unexpected loss (difference between actual defaults and predicted defaults) across the portfolio.

d. Propose two actions the bank should take to address this miscalibration.

Difficulty: Advanced — requires calibration analysis and business judgement.

Exercise 5: Population Stability Index

Scenario: A scorecard was developed on a 2022 application sample. The 2024 application population shows a different score distribution (synthetic data):

Score Range	2022 (Expected)	2024 (Actual)
< 550	10%	18%
550–599	15%	20%
600–649	25%	25%
650–699	30%	22%
≥ 700	20%	15%

Tasks:

a. Calculate the PSI for this scorecard.

b. Interpret the PSI value: what action is recommended?

c. The shift occurred after a recession. Explain why a recession would cause the score distribution to shift leftward (toward lower scores).

d. Should the bank rebuild the model, recalibrate it, or do nothing? Justify your recommendation.

Difficulty: Advanced — requires PSI computation and contextual analysis.

Exercise 6: Alternative Data for Thin-File Borrowers

Scenario: A FinTech lender in Southeast Asia is considering three alternative data sources to score thin-file borrowers (those with no traditional credit bureau data):

Criterion	Source A: Mobile Airtime	Source B: E-commerce History	Source C: Social Media
Gini (thin-file)	0.38	0.44	0.30
Population coverage	85%	40%	70%
Stability (PSI over 6 months)	0.12	0.08	0.22
Privacy risk	Medium	Low	High

Tasks:

a. Rank the three sources by discriminatory power (Gini). Which is strongest?

b. Rank by stability (lower PSI = more stable). Which is most stable?

c. The lender can only afford to integrate two sources. Which two should they choose? Justify your recommendation considering accuracy, stability, coverage, and privacy.

d. A regulator proposes banning social media data. What would be the impact on the lender's thin-file scoring capability?

Difficulty: Advanced — requires multi-criteria decision-making.

Exercise 7: Fairness Audit of a Credit Model

Scenario: A bank's credit model is audited for fairness. The approval rates by demographic group are (synthetic data):

Group	Applicants	Approval Rate	Default Rate (approved)
Group A (majority)	8,000	72%	3.5%
Group B (minority)	2,000	48%	3.8%

Tasks:

a) Calculate the **disparate impact ratio** (Group B approval rate / Group A approval rate). Does it fall below the US "four-fifths rule" threshold of 0.80?

b) The default rates among approved borrowers are similar (3.5% vs. 3.8%). What does this suggest about the rejected Group B applicants?

c) A data scientist proposes adding zip code to improve model accuracy. The model's AUC increases from 0.76 to 0.79, but the approval gap widens. Should the bank add zip code? Argue both sides.

d) Propose one model-level and one policy-level intervention to reduce the approval gap without significantly increasing default losses.

Difficulty: Advanced — requires fairness analysis and policy reasoning.

Exercise 8: Building a FinTech Credit Scoring System

Scenario: You are hired as the lead data scientist for a FinTech lender launching in a developing market. The population is 60% thin-file (no bureau data). You have access to mobile money transaction histories and utility payment records. Your target portfolio: 100,000 micro-loans of $500 each, maximum acceptable default rate: 8%.

Tasks:

a) Design the feature engineering pipeline: list 6 features you would derive from mobile money and utility data, and explain the economic rationale for each.

b) You build a gradient boosting model with AUC = 0.78 on a holdout set. The calibration plot shows the model under-predicts PD for borrowers in the 5%–10% PD range. What is the business risk?

c) The regulator requires adverse action notices. You cannot show gradient boosting internals. Describe your explainability strategy.

d) After 6 months, PSI = 0.30. A new mobile money provider entered the market, changing transaction patterns. What do you do? Outline a 3-step remediation plan.

e) Calculate the maximum acceptable default rate in terms of total losses: if LGD = 50%, what is the portfolio's maximum Expected Loss at your 8% PD threshold?

Difficulty: Integrative — combines feature engineering, model evaluation, explainability, monitoring, and risk management.

Answer Key

Exercise 1:

(a) WoE($<$2yr) $= \ln(0.30/0.50) = \ln(0.60) = -0.511$. WoE(2–5yr) $= \ln(0.40/0.333) = \ln(1.20) = 0.182$. WoE($>$5yr) $= \ln(0.30/0.167) = \ln(1.80) = 0.588$.

(b) IV $= (0.30 - 0.50)(-0.511) + (0.40 - 0.333)(0.182) + (0.30 - 0.167)(0.588) = 0.102 + 0.012 + 0.078 =$ **0.192**.

(c) IV $= 0.192$ falls in the 0.10–0.30 range = **medium predictor**. Include in the scorecard.

(d) Negative WoE means this bin has proportionally more bads than goods. Short employment tenure is associated with higher default risk, which is intuitive—less job stability correlates with repayment difficulty.

Exercise 2:

(a) EL(A) $= 5{,}000 \times \$15{,}000 \times 0.01 \times 0.30 =$ **\$225,000**. EL(B) $= 3{,}000 \times \$12{,}000 \times 0.04 \times 0.40 =$ **\$576,000**. EL(C) $= 1{,}000 \times \$8{,}000 \times 0.10 \times 0.55 =$ **\$440,000**.

(b) Total EL $= \$225{,}000 + \$576{,}000 + \$440{,}000 =$ **\$1,241,000**.

(c) Total exposure $= 5{,}000 \times \$15\text{K} + 3{,}000 \times \$12\text{K} + 1{,}000 \times \$8\text{K} = \$75\text{M} + \$36\text{M} + \$8\text{M} = \$119\text{M}$. EL% $= \$1.241\text{M} / \$119\text{M} =$ **1.04%**.

(d) Grade C: PD $\times$ LGD $= 10\% \times 55\% = 5.5\%$. Add 2% margin $\Rightarrow$ minimum spread = **7.5%** over funding cost.

Answer Key (continued)

Exercise 3:

(a) Gini(LR) $= 2 \times 0.74 - 1 =$ **0.48**. Gini(XGB) $= 2 \times 0.81 - 1 =$ **0.62**.

(b) Lift(LR) $= 9.2\%/3\% =$ **3.07**$\times$. Lift(XGB) $= 14.5\%/3\% =$ **4.83**$\times$.

(c) Top 20% = 2,000 applicants. Total defaults = 300. If LR captures 55% of defaults in top 20% = 165; XGB captures 70% = 210. Additional defaults caught $\approx$ **45**. (Estimates based on AUC-implied capture rates.)

(d) Use XGBoost for the accept/decline decision. Generate adverse action reasons using SHAP values on the XGBoost model, mapping the top 4 contributing features to human-readable reason codes (e.g., "High credit utilisation", "Short employment tenure").

Exercise 4:

(a) Ratios: 700–750: $1.2/1.0 = 1.20$. 650–699: $3.1/3.0 = 1.03$. 600–649: $7.8/6.0 = 1.30$. 550–599: $14.5/10.0 = 1.45$. $<$550: $25.0/18.0 = 1.39$.

(b) The model **under-predicts** defaults across all bands, with the worst miscalibration in the 550–599 band (1.45$\times$). Higher-risk borrowers' PDs are most under-estimated.

(c) Predicted defaults: $2{,}000 \times 1\% + 3{,}000 \times 3\% + 2{,}500 \times 6\% + 1{,}500 \times 10\% + 500 \times 18\% = 20+90+150+150+90 = 500$. Actual defaults: $2{,}000 \times 1.2\% + 3{,}000 \times 3.1\% + 2{,}500 \times 7.8\% + 1{,}500 \times 14.5\% + 500 \times 25\% = 24+93+195+217.5+125 = 654.5$. Unexpected loss $\approx$ **155 additional defaults**.

(d) (1) Recalibrate PDs by applying a scaling factor to predicted PDs. (2) Re-price loans in the lower score bands to reflect higher actual risk.

Answer Key (continued)

Exercise 5:

(a) PSI $= (0.18 - 0.10)\ln(0.18/0.10) + (0.20 - 0.15)\ln(0.20/0.15) + (0.25 - 0.25)\ln(1) + (0.22 - 0.30)\ln(0.22/0.30) + (0.15 - 0.20)\ln(0.15/0.20) = 0.08 \times 0.588 + 0.05 \times 0.288 + 0 + (-0.08)(-0.310) + (-0.05)(-0.288) = 0.047 + 0.014 + 0 + 0.025 + 0.014 = \mathbf{0.100}$.

(b) PSI $= 0.10$ is at the boundary of "no shift" and "moderate shift." Investigate but no immediate rebuild required.

(c) A recession increases unemployment and financial stress, pushing more applicants into lower score bands. Features like DTI, utilisation, and delinquency worsen, shifting the score distribution left.

(d) Recommended: **recalibrate** the model (adjust the PD mapping) rather than full rebuild, since PSI $= 0.10$ is moderate. Monitor monthly. If PSI exceeds 0.25 or AUC degrades by >5 points, trigger a full rebuild.

Exercise 7:

(a) Disparate impact ratio $= 48\%/72\% = \mathbf{0.667}$. This is below 0.80, indicating potential adverse impact.

(b) Similar default rates among approved borrowers suggest that rejected Group B applicants may include many creditworthy individuals. The model is likely **over-rejecting** Group B.

(c) *For:* Higher AUC means better risk differentiation and lower losses. *Against:* Zip code is a proxy for race; the approval gap widens, deepening disparate impact and creating legal liability. The 3-point AUC gain may not justify the fairness cost.

(d) *Model-level:* Remove zip code and other proxy features; use fairness-constrained optimisation (e.g., equalised odds). *Policy-level:* Apply a lower score cut-off for Group B (calibrated so that the marginal approval default rate matches Group A), or offer a "second look" manual review for borderline Group B applicants.

Answer Key (continued)

Exercise 8:

(a) Six features from mobile money and utility data: (1) *Average monthly mobile money inflow*—proxy for income stability. (2) *Transaction frequency (last 90 days)*—regular activity suggests financial engagement. (3) *Ratio of outflows to inflows*—spending discipline indicator. (4) *Utility payment consistency (on-time %)*—direct repayment behaviour proxy. (5) *Maximum consecutive months with utility payments*—stability signal. (6) *Number of unique mobile money counterparties*—network breadth as a social capital indicator.

(b) Under-predicting PD in the 5%–10% range means the bank will approve borrowers whose true risk is higher than estimated. Loans will be **under-priced**, and actual losses will exceed provisions. This erodes capital.

(c) Use SHAP (SHapley Additive exPlanations) on the gradient boosting model to compute per-applicant feature contributions. Map the top 3–4 SHAP drivers to standardised reason codes (e.g., "Irregular income pattern," "Low utility payment consistency").

(d) 3-step remediation: (1) **Investigate:** Identify which features shifted (likely transaction frequency and counterparty count changed with the new provider). (2) **Retrain:** Rebuild the model on recent 6-month data including new-provider transactions. (3) **Monitor:** Set automated PSI alerts at the feature level, not just the score level.

(e) Max EL = PD $\times$ LGD $\times$ EAD = 8% $\times$ 50% $\times$ (100,000 $\times$ \$500) = 0.04 $\times$ \$50M = **\$2,000,000** (4% of total exposure).