

Statistics of Financial Markets

Chapter 12: Scoring models



Daniel Traian PELE

Bucharest University of Economic Studies
IDA Institute Digital Assets
Blockchain Research Center
AI4EFin Artificial Intelligence for Energy Finance
Romanian Academy, Institute for Economic Forecasting
MSCA Digital Finance

Acronyms used in this material (1/2)

- **AI**: Artificial Intelligence
- **AR**: Accuracy Ratio (of the CAP curve)
- **AUC**: Area Under the (ROC) Curve
- **BIS**: Bank for International Settlements
- **BS**: Brier Score
- **CAP**: Cumulative Accuracy Profile
- **CDS**: Credit Default Swap
- **CI**: Confidence Interval
- **DD**: Distance to Default
- **DM**: Deutsche Mark (the German mark)
- **EAD**: Exposure At Default
- **EBIT**: Earnings Before Interest and Taxes
- **EU**: European Union
- **FICO**: Fair Isaac Corporation
- **FN**: False Negative
- **FPR**: False Positive Rate
- **GBM**: Geometric Brownian Motion
- **HL**: Hosmer–Lemeshow (test)
- **IRB**: Internal Ratings-Based (approach)

Acronyms used in this material (2/2)

- **KS**: Kolmogorov–Smirnov (statistic)
- **LDA**: Linear Discriminant Analysis
- **LGD**: Loss Given Default
- **LR**: Likelihood Ratio (test)
- **MVA**: Applied Multivariate Statistical Analysis (Härdle and Simar)
- **NYU**: New York University
- **PD**: Probability of Default
- **PDO**: Points to Double the Odds
- **ROC**: Receiver Operating Characteristic (curve)
- **RWA**: Risk-Weighted Assets
- **SFE**: Statistics of Financial Markets, Exercises (Quantlet collection of the textbook)
- **TA**: Total Assets
- **TN**: True Negative
- **TP**: True Positive
- **TPR**: True Positive Rate
- **UCI**: University of California, Irvine (Machine Learning Repository)
- **US**: United States

Today's Question and Route

- **Question**: how does a bank turn the data of an applicant into a probability of default, and how do we know that the number is right?
  - Chapters 2–10 measured the risk of prices; this chapter measures the risk that a borrower does not pay back
- **Route** of the chapter
  - credit scores, PD, LGD, EAD and the Basel capital; default data and class imbalance
  - Weight of Evidence and Information Value; logistic regression and Fisher's discriminant analysis
  - the scorecard and its validation: confusion matrix, ROC, AUC, Gini, CAP, KS, Brier score, calibration
  - reject inference and fairness; the Altman Z-score and the Merton distance to default

Learning Outcomes

- Define PD, LGD, EAD and the expected loss, and explain what the Basel IRB capital adds to the expected loss
- Compute the Weight of Evidence and the Information Value of a variable from a table of counts
- Estimate a logistic regression and interpret its coefficients as changes in log-odds and as odds ratios
- Derive Fisher's linear discriminant and compare it with the logit
- Scale a scorecard and validate it out of sample with the confusion matrix, AUC, Gini, KS, the Brier score and a calibration plot

Reading and Tools

- Textbook: Franke, Härdle & Hafner (2019), Ch. 21 (probability of default) and Ch. 22 (credit risk management)
  - discriminant analysis: Härdle & Simar (2019), Ch. 14
  - credit scoring: Thomas, Crook & Edelman (2017); scorecards: Siddiqi (2006)
- Python Quantlets of this chapter: Quantlets/Ch_12
  - the conditional PD of the one-factor model is ported from the SFE Quantlet SFEdefaproba
- Lecture notebook: open in Google Colab
- Video course: MVA Multivariate Statistical Analysis

The Lending Decision

- An applicant asks for a loan; the bank must decide **accept** or **reject**, and at which interest rate
  - **default**: the borrower does not pay as agreed; Basel: the bank judges the borrower unlikely to pay, or the borrower is more than 90 days past due on a material obligation Basel Committee (2006), ¶452
- **Credit score**: a number computed from the data of an applicant that orders applicants by their risk of default
  - **PD** (probability of default): the probability of default within a horizon, usually one year
  - a good score both **ranks** (discrimination) and gives the right **level** of PD (calibration)
- Kinds of scores
  - application scores (new clients) and behavioural scores (existing clients, from their payment history)
  - consumer scores (statistical models on many small loans) and corporate scores (accounting ratios, market prices)

Early Credit Scoring: Durand and FICO

- 1941: Durand (1941) separates good and bad instalment loans of US lenders with Fisher's discriminant analysis
  - the first statistical credit-scoring study
- 1956: Bill Fair and Earl Isaac found Fair, Isaac and Company (today FICO) myFICO, history of the FICO score
  - 1989: the FICO score, computed from credit-bureau reports, between 300 and 850 points
- Statistical scoring replaced the judgement of a loan officer with a rule that is the same for everybody Hand & Henley (1997)
  - cheaper, faster and testable; the question whether the rule is fair remains



FICO headquarters, San Jose, California

Photo: Coolcaesar (2020); CC BY-SA 4.0; Wikimedia Commons

PD, LGD, EAD and the Expected Loss

- Three risk parameters of a loan (Basel II, Basel Committee (2006))
  - **PD**: probability of default within one year
  - **LGD** (loss given default): the share of the exposure lost if default occurs, after recoveries
  - **EAD** (exposure at default): the amount owed at the moment of default
- **Expected loss**: $\mathrm{EL} = \mathrm{PD} \times \mathrm{LGD} \times \mathrm{EAD}$ (with independent PD and LGD)
  - **Example**: a loan of 100 000 RON with PD $= 2\%$, LGD $= 45\%$: EL $= 0.02 \times 0.45 \times 100\,000 = 900$ RON
  - EL is a cost of doing business: it is priced into the interest rate and covered by provisions
- Capital covers the **unexpected loss**: losses above EL in a bad year

The Basel IRB Capital Formula

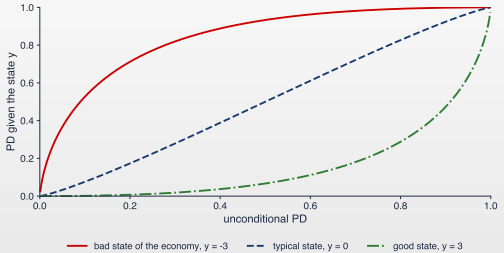
- **IRB** (internal ratings-based) approach: the bank estimates PD (and possibly LGD, EAD); the regulator gives the formula Basel Committee (2006), ¶328–330
  - $K = \mathrm{LGD}\left[\Phi\left(\frac{\Phi^{-1}(\mathrm{PD}) + \sqrt{R}\,\Phi^{-1}(0.999)}{\sqrt{1-R}}\right) - \mathrm{PD}\right]$; RWA $= 12.5\,K\,\mathrm{EAD}$
  - K: the capital per unit of EAD; Φ, Φ^{-1}: the $N(0,1)$ distribution function and its inverse; 0.999: a bad year occurs once in 1000 years
  - R: the correlation of the borrowers with one common factor; RWA: risk-weighted assets, of which the bank holds at least 8%
- Logic Gordy (2003): the term in $\Phi(\cdot)$ is the PD in a year whose common factor is exceeded only with probability 0.1%
  - so K is the credit VaR 0.1% of the loss minus EL: the unexpected loss (Chapter 10)



The BIS tower in Basel

Photo: Taxiarchos228 (2014); CC BY-SA 4.0; Wikimedia Commons

PD in Good and Bad Years: the One-Factor Model



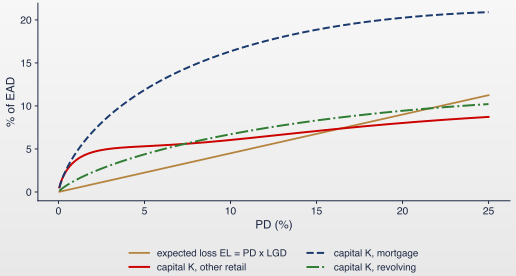
- Borrower i defaults if $\sqrt{\rho}\,Y + \sqrt{1-\rho}\,\varepsilon_i < \Phi^{-1}(\mathrm{PD})$; Y: the state of the economy; given $Y = y$: $\mathrm{PD}(y) = \Phi\big((\Phi^{-1}(\mathrm{PD}) - \sqrt{\rho}\,y)/\sqrt{1-\rho}\big)$, $\rho = 0.2$
- ε_i: the borrower's own shock; Y and ε_i independent $N(0,1)$; ρ: the weight of the common factor
- Interpretation: a loan with PD $= 2\%$ defaults with probability 21.3% in a bad year ($y = -3$), 1.1% in a typical year and almost never in a good year; IRB uses $y = \Phi^{-1}(0.001)$

SFM_ch12_validation_irb

Worked Example: IRB Capital of a Retail Loan

- Other retail exposures: $R = 0.03\,w + 0.16\,(1 - w)$, $w = (1 - e^{-35\,\mathrm{PD}})/(1 - e^{-35})$ Basel Committee (2006), ¶330
  - $w \in [0, 1]$ grows with PD, so R falls from 0.16 (very safe loans) to 0.03 (very risky loans)
  - PD = 2%: $R = 0.095$; $\Phi^{-1}(0.02) = -2.054$, $\Phi^{-1}(0.999) = 3.090$
- **Step by step**
  - argument: $(-2.054 + \sqrt{0.095} \times 3.090)/\sqrt{1 - 0.095} = -1.160$; $\Phi(-1.160) = 12.31\%$: the PD in a bad year
  - $K = 0.45 \times (12.31\% - 2\%) = 4.64\%$ of EAD: 4 639 RON for a loan of 100 000 RON
  - RWA $= 12.5 \times 4\,639 = 57\,986$ RON: a risk weight of 58%
- Interpretation: EL (900 RON) is priced in; the capital (4 639 RON) absorbs a year in which defaults reach 12.31% instead of 2%

Expected Loss and IRB Capital as Functions of PD



- LGD $= 45\%$; capital K for other retail (R from 0.16 to 0.03), residential mortgages ($R = 0.15$) and revolving retail exposures ($R = 0.04$)
- Interpretation: EL grows linearly with PD; K grows fast at low PD and then flattens, because a very risky loan is mostly an expected (priced) loss

SFM_ch12_validation_irb

The South German Credit Data

- 1 000 consumer credits of a large regional bank in southern Germany, 1973–1975 South German Credit (2020); Grömping (2019)
  - outcome: the contract was complied with (good) or not (bad, **default** $= 1$): 300 bad and 700 good credits
  - 20 variables: checking account, duration, credit history, purpose, amount (DM, Deutsche Mark), savings, employment, age and others
- A **stratified sample**: bad credits are heavily oversampled; the bad rate of the bank was about 5% Grömping (2019)
  - all borrowers had passed the bank's checks: rejected applicants are missing (reject inference, Section 8)
  - the cost of accepting a bad credit is set to five times the cost of rejecting a good one
- Licence CC BY 4.0; saved once in the course repository (data/credit)

Coding Errors in the German Credit Data

- The data were used for decades in the UCI version "Statlog (German Credit)", with code labels that were wrong Grömping (2019)
  - in that version, "no checking account" was the safest group: 394 credits, 11.7% bad
  - in the corrected data, those 394 credits have a balance of at least 200 DM or a salary account; "no checking account" is the riskiest group (49.3% bad)
- The numbers were right, the labels were wrong: every model was fine, every interpretation of this variable was reversed
  - checks: compare the frequency tables with the original source (here the book of Fahrmeir and Hamerle), ask whether the signs make economic sense
- We use the corrected South German Credit data throughout

Bad Rate by Checking Account



- Share of bad credits in each group; dashed: the sample average of 30%
- Interpretation: the risk falls from 49.3% (no account) to 11.7% (balance of at least 200 DM): one variable already separates the applicants well

SFM_ch12_woe_iv

Class Imbalance and Oversampling

- Defaults are rare: a few percent of the loans of a bank; a sample with few bads tells little about them
  - banks therefore oversample bads (here 30% instead of about 5%)
- **Accuracy paradox**: the rule "everybody is good" is right for 70% of this sample and for 77.9% of the Taiwan data
  - so accuracy alone is a poor measure; we need measures that look at the bads and the goods separately (Section 7)
- **Taiwan credit card data** Default of Credit Card Clients (2009); Yeh & Lien (2009): 30 000 card holders, default on the October 2005 payment, 22.1% defaults
  - inputs: credit limit, age, the repayment status of April–September 2005, bills and payments; we use them for calibration and fairness

Training and Test Samples

- **Stratified split**: 700 credits for estimation (training), 300 for testing, with 30% bads in both
  - every choice (bins, WoE, variables, coefficients) is made on the training sample only
  - the test sample is used once, at the end: it plays the role of the future applicants
- **Leakage**: information from the test sample that enters the model, e.g. bins chosen on all data
  - it makes the test result too optimistic; with 300 test credits one split is also noisy, so we add cross-validation (Section 7)

Weight of Evidence

- Split a variable into bins $i = 1, \dots, m$; g_i, b_i: goods and bads in bin i; G, B: their totals
  - $\mathrm{WoE}_i = \ln \dfrac{g_i / G}{b_i / B}$: the log of the share of goods over the share of bads in the bin
- Meaning: $\mathrm{WoE}_i = \ln \dfrac{g_i}{b_i} - \ln \dfrac{G}{B}$: the log-odds of being good in the bin minus the log-odds in the whole sample
  - $\mathrm{WoE} > 0$: safer than average; $\mathrm{WoE} < 0$: riskier; $\mathrm{WoE} = 0$: no information
  - a bin without goods or without bads would give $\pm\infty$: add 0.5 to both counts
- **Information Value**: $\mathrm{IV} = \sum_i \left(\dfrac{g_i}{G} - \dfrac{b_i}{B} \right) \mathrm{WoE}_i \geq 0$
  - a symmetric Kullback–Leibler divergence between the distributions of goods and bads across the bins
  - rule of thumb Siddiqi (2006): below 0.02 useless, 0.02–0.1 weak, 0.1–0.3 medium, above 0.3 strong

Worked Example: WoE and IV of the Checking Account

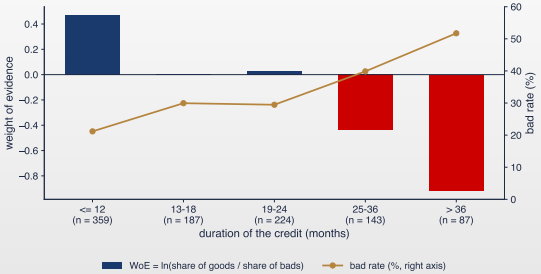
Checking account	goods	bads	g_i/G	b_i/B	WoE	IV term
no checking account	139	135	0.199	0.450	-0.818	0.206
balance $<$ 0 DM	164	105	0.234	0.350	-0.401	0.046
0 to 200 DM	49	14	0.070	0.047	0.405	0.009
$\geq$ 200 DM or salary	348	46	0.497	0.153	1.176	0.404
total	700	300	1	1		0.666

- **Step by step**, last row: $g/G = 348/700 = 0.497$, $b/B = 46/300 = 0.153$
  - WoE $= \ln(0.497/0.153) = 1.176$; IV term $= (0.497 - 0.153) \times 1.176 = 0.404$
- Interpretation: IV $= 0.666 > 0.3$: a strong variable; the WoE falls monotonically from the safest to the riskiest group

Binning: Practical Rules

- Why bins: they capture non-linear effects, limit the influence of outliers and give every applicant a small table of points
  - numeric variables: a few intervals (here duration ≤ 12, 13–18, 19–24, 25–36, > 36 months); categorical: the categories
- Rules Siddiqi (2006); Thomas, Crook & Edelman (2017)
  - at least about 5% of the sample in every bin; merge small categories
  - WoE should change in a way that makes economic sense (often monotonically)
  - bins and WoE are estimated on the training sample and applied unchanged to new applicants
- A WoE variable enters the logit with one coefficient: the scorecard stays additive and easy to read

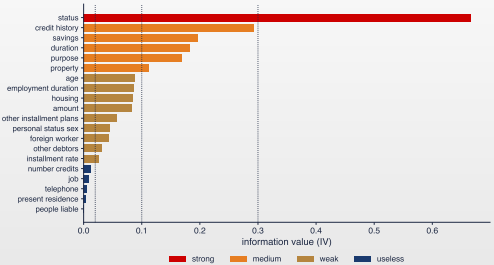
WoE of the Credit Duration



- Full sample; bars: WoE of each duration bin; line: bad rate (right axis); IV = 0.182 (medium)
- Interpretation: credits up to one year are safer (WoE = 0.47, 21% bad); above three years the bad rate reaches 52% (WoE = −0.92); 13–24 months are average

SFM_ch12_woe_iv

Information Value of the 20 Variables



- Full sample; dotted: the thresholds 0.02, 0.1 and 0.3; on the training sample, 6 variables have IV ≥ 0.1 and enter the scorecard
- Interpretation: the checking account (IV $= 0.666$) and the credit history (0.293) dominate; telephone, residence and dependants carry no information
- IV looks at one variable at a time: two strong but correlated variables do not add their information

SFM_ch12_woe_iv

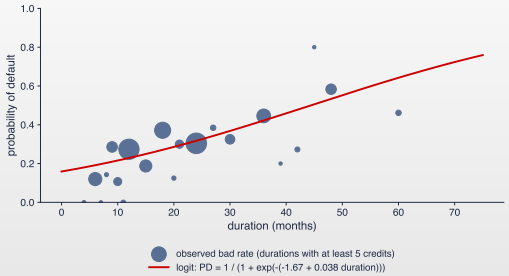
Odds and Log-Odds

- A linear model $P(Y = 1) = \beta_0 + x^\top \beta$ can give probabilities below 0 or above 1
  - we model a transformation of p that can take any real value
- **Odds** of default: $\dfrac{p}{1-p} \in (0, \infty)$; **log-odds** (logit): $\ln \dfrac{p}{1-p} \in (-\infty, \infty)$
  - $p = 0.20$: odds $= 0.2/0.8 = 0.25$ (one default for four repaid loans); log-odds $= \ln 0.25 = -1.386$
  - $p = 0.5$: odds 1, log-odds 0; the logit is symmetric: $\mathrm{logit}(1-p) = -\mathrm{logit}(p)$
- Bankers often speak of the **good:bad odds** $(1-p)/p$: here 4:1

The Logit Model

- $\ln \frac{p_i}{1-p_i} = \beta_0 + x_i^\top \beta = \eta_i, \quad p_i = P(Y_i = 1 \mid x_i) = \frac{1}{1+e^{-\eta_i}}$ Cox (1958)
  - $x_i = (x_{i1}, \ldots, x_{ik})^\top$: vector of the k characteristics of applicant i; $\beta = (\beta_1, \ldots, \beta_k)^\top$: vector of their coefficients, the same for all applicants
  - β_0: the constant (intercept); $\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$: the linear predictor, the log-odds of applicant i
- **Maximum likelihood** (Chapter 6): $\ell(\beta) = \sum_i [y_i \ln p_i + (1-y_i)\ln(1-p_i)]$
  - first-order condition $\sum_i (y_i - p_i)x_i = 0$: no closed form, solved by Newton–Raphson
  - standard errors from the inverse of $\sum_i p_i(1-p_i)x_i x_i^\top$; Wald test $z = \hat{\beta}_j / \mathrm{se}(\hat{\beta}_j)$
- The logit is also used for corporate bankruptcy, starting with Ohlson (1980); the probit uses $\Phi(\eta)$ instead of the logistic function
  - logit and probit give almost the same PDs; the logit has the odds interpretation

A Logistic Curve: PD and Duration



- Logit with duration only: $\hat{\beta}_0 = -1.666$, $\hat{\beta}_1 = 0.0375$ per month (se 0.0057); dots: observed bad rates
- Interpretation: each extra year multiplies the odds of default by $e^{12 \times 0.0375} = 1.57$; the PD rises from 22.9% at 12 months to 53.4% at 48 months

SFM_ch12_logit_lda

Interpreting the Coefficients

- β_j: the change in the **log-odds** when x_j grows by one unit, the other variables fixed
  - e^{β_j}: the **odds ratio**: the odds are multiplied by e^{β_j}; $100(e^{\beta_j} - 1)\%$ is the change in the odds
  - a 95% confidence interval: $\exp(\hat{\beta}_j \pm 1.96\,\mathrm{se})$
- The effect on the **probability** depends on where the applicant is
  - $\partial p / \partial x_j = p(1-p)\beta_j$: largest at $p = 0.5$, small for very safe or very risky applicants
- A dummy variable (0/1): e^{β_j} compares the odds of the group with those of the reference group
  - the interpretation needs a model that contains the relevant variables: an omitted variable changes the coefficients

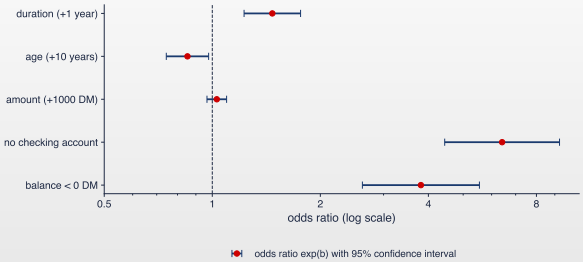
A Small Logit on the Full Sample

Variable	$\hat{\beta}$	se	z	p	odds ratio (95% CI)
constant	−2.101	0.311	−6.75	< 0.001	
duration (years)	0.385	0.092	4.17	< 0.001	1.47 (1.23–1.76)
age (decades)	−0.159	0.069	−2.30	0.022	0.85 (0.74–0.98)
amount (1000 DM)	0.029	0.032	0.90	0.368	1.03 (0.97–1.10)
no checking account	1.859	0.188	9.89	< 0.001	6.42 (4.44–9.28)
balance $<$ 0 DM	1.338	0.192	6.99	< 0.001	3.81 (2.62–5.55)

- $n = 1\,000$; log-likelihood -526.1 against -610.9 with the constant only: LR $= 169.6$ on 5 degrees of freedom
- Reference group of the account dummies: a non-negative balance (0–200 DM, or at least 200 DM or a salary account)

SFM_ch12_logit_lda

Odds Ratios with Confidence Intervals



- Each dot: $e^{\hat{\beta}}$; bars: 95% intervals; log scale; vertical line: no effect (1)
- Interpretation: without a checking account the odds of default are 6.42 times those of the reference group; a balance below zero multiplies them by 3.81
- One more year of duration raises the odds by 47%; ten more years of age lower them by about 15%
- The amount is not significant once the duration is in the model (the two are correlated)

SFM_ch12_logit_lda

Worked Example: the PD of an Applicant

- Applicant: 36 months, 30 years old, 4000 DM, no checking account; coefficients of the previous table, rounded
  - $\eta = -2.101 + 0.385 \times 3 + (-0.159) \times 3 + 0.029 \times 4 + 1.859$
  - $\eta = -2.101 + 1.155 + (-0.477) + 0.116 + 1.859 = 0.552$
- PD $= 1/(1 + e^{-0.552}) = 63.5\%$; odds $e^{0.552} = 1.74$
  - the same applicant with a non-negative balance: $\eta = -1.307$, PD $= 21.3\%$
  - one more year of duration: $\Delta p \approx p(1-p)\beta \approx 8.9$ percentage points
- Caution: this PD refers to a sample with 30% bads, not to the bank's population (prior correction, two slides ahead)

The WoE Logit Scorecard

WoE variable	IV (training)	$\hat{\beta}$	se
constant		−0.850	0.095
checking account	0.591	−0.829	0.127
credit history	0.215	−0.675	0.206
purpose	0.205	−0.937	0.217
savings	0.166	−0.726	0.246
duration	0.159	−1.082	0.239
years with the employer	0.122	−0.873	0.265

- Training sample (700 credits); each variable replaced by the WoE of its bin; logit for default
  - all coefficients are negative: a higher WoE (safer bin) lowers the log-odds of default
  - a coefficient of -1 would mean that the variable adds exactly its own WoE; values between -0.7 and -1.1 reflect the overlap between variables
- Test sample: AUC $= 0.801$ (Section 7); in-sample log-likelihood -347.2

SFM_ch12_scorecard

Correcting for Oversampling

- If bads and goods are sampled with different rates, only the constant of the logit changes
  - $\ln \text{odds}_{\text{pop}} = \ln \text{odds}_{\text{sample}} + \ln \dfrac{\pi/(1-\pi)}{\pi_s/(1-\pi_s)}$; π: population bad rate, π_s: sample rate
  - the slopes are unchanged: the ranking of the applicants does not depend on the sampling
- **Step by step**: $\pi_s = 0.30$, $\pi = 0.05$: shift $= \ln(0.05/0.95) - \ln(0.30/0.70) = -2.944 - (-0.847) = -2.097$
  - a sample PD of 30% becomes 5%; a sample PD of 50% becomes 10.9%
  - the applicant of the worked example: $0.552 - 2.097 = -1.545$, PD $= 17.6\%$ in the bank's population instead of 63.5%
- Interpretation: AUC and Gini are unchanged by the correction; EL and capital are not

Fisher's Discriminant Function (1936)

- Fisher (1936): find the linear combination $w^\top x$ that separates two groups best
  - he used measurements of iris flowers; Durand applied it to loans five years later
- Fisher criterion: $J(w) = \dfrac{[w^\top (m_1 - m_0)]^2}{w^\top S_W w}$
  - m_0, m_1: mean vectors of goods and bads; S_W: the pooled within-group covariance matrix
  - maximum: $w \propto S_W^{-1}(m_1 - m_0)$ Härdle & Simar (2019), Ch. 14
- No distribution is needed for this criterion: it is a projection that maximises the distance between the groups relative to their spread



R.A. Fisher, 1913

Photo: unknown author (1913); public domain; Wikimedia Commons



Iris versicolor

Photo: Ermell (2020); CC BY-SA 4.0; Wikimedia Commons

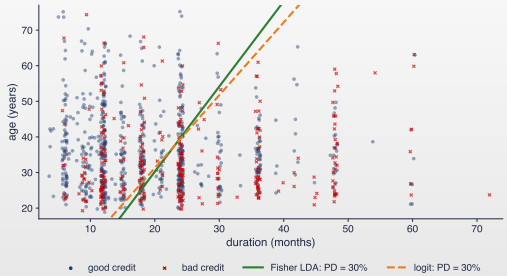
LDA as a Probability Model

- Assume $x \mid Y = k \sim N(m_k, \Sigma)$, the same Σ in both groups, and prior $\pi_1 = P(Y = 1)$
  - Bayes: $\ln \dfrac{P(Y=1 \mid x)}{P(Y=0 \mid x)} = \underbrace{\ln \dfrac{\pi_1}{1-\pi_1} - \tfrac{1}{2}(m_1 + m_0)^\top w}_{c} + w^\top x, \quad w = \Sigma^{-1}(m_1 - m_0)$
  - the log-odds are **linear in** x, exactly as in the logit; the prior enters only the constant
- The difference is in the estimation
  - LDA: plug in the sample means and the pooled covariance (a closed form)
  - logit: maximise the conditional likelihood of Y given x; no assumption on the distribution of x
- Classification rule: assign to "bad" if $c + w^\top x > \ln(C_{FP}/C_{FN})$ (costs, Section 7)

Worked Example: Fisher's Direction with Two Variables

- $x =$ (duration in months, age in years), full sample: 700 goods, 300 bads
  - $m_{\mathbf{0}} = (19.21, 36.22)$, $m_{\mathbf{1}} = (24.86, 33.96)$, $m_{\mathbf{1}} - m_{\mathbf{0}} = (5.65, -2.26)$
  - $S_W = \begin{pmatrix} 138.8 & -2.46 \\ -2.46 & 127.9 \end{pmatrix}$, $\det S_W = 17756$
- **Step by step**: $S_W^{-\mathbf{1}} = \frac{1}{\det S_W} \begin{pmatrix} 127.9 & -(-2.46) \\ -(-2.46) & 138.8 \end{pmatrix}$, $w = S_W^{-\mathbf{1}}(m_{\mathbf{1}} - m_{\mathbf{0}}) = (0.0404, -0.0169)$
  - logit with the same two variables: slopes $(0.0375, -0.0183)$
  - ratio duration/age: LDA -2.39, logit -2.05: one month of duration weighs as much as about two years of age, in the other direction

Fisher's Boundary and the Logit Boundary



- Duration and age of every credit (a small jitter added); lines: the points with PD = 30% (the sample rate) for the LDA and for the logit
- Interpretation: both boundaries are lines with almost the same slope; the two clouds overlap heavily, so two variables are far from enough

SFM_ch12_logit_lda

Comparison of Logit and LDA

	Logit	LDA (Fisher)
model	$P(Y \mid x)$ only	$P(x \mid Y)$ Normal, common Σ
estimation	maximum likelihood, iterative	means and covariance, closed form
if x is Normal	consistent, less efficient	more efficient
dummies, WoE, skewed x	fine	ranking fine, PDs may be off
test AUC (WoE inputs)	0.801	0.801

- Test sample: the correlation of the two log-odds scores is 0.999; DeLong test of equal AUCs: $z = 0.18$, p $= 0.86$ DeLong et al. (1988)
- Interpretation: for ranking, the two methods are practically identical; the logit is preferred for PDs because it does not need Normal inputs

SFM_ch12_logit_lda

Scaling a Scorecard

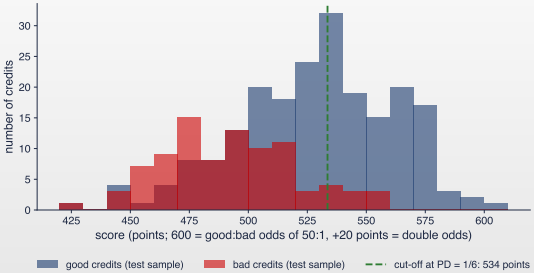
- **Scorecard**: a table that gives points to every attribute; the score of an applicant is the sum of the points Siddiqi (2006)
  - higher score = lower risk; the score is a linear function of the good:bad log-odds
- $\text{Score} = \text{Offset} + \text{Factor} \times \ln \frac{1-p}{p}$, two conventions fix the scale
  - **PDO** (points to double the odds): +PDO points $\Leftrightarrow$ the good:bad odds double: $\text{Factor} = \text{PDO}/\ln 2$
  - a base score at given odds: $\text{Offset} = \text{Base} - \text{Factor} \ln(\text{odds}_0)$
- Here: PDO = 20, 600 points at odds 50:1
  - $\text{Factor} = 20/\ln 2 = 28.85$; $\text{Offset} = 600 - 28.85 \times \ln 50 = 600 - 28.85 \times 3.912 = 487.12$

Worked Example: Scores and Points

- **PD to score**: PD $= 5\%$, odds 19:1: Score $= 487.12 + 28.85 \times \ln 19 = 487.12 + 28.85 \times 2.944 = 572.1$
  - PD $= 30\%$ (odds 7:3): $487.12 + 28.85 \times 0.847 = 511.6$
- **Score to PD**: 560 points: ln odds $= (560 - 487.12)/28.85 = 2.526$, odds 12.5:1, PD $= 7.4\%$
- **Points of an attribute** ($k = 6$ variables; logit for default with constant $\beta_{\mathbf{0}}$ and slope β_j)
  - $\text{points}_{ij} = -(\beta_j \,\text{WoE}_{ij} + \beta_{\mathbf{0}}/k)\,\text{Factor} + \text{Offset}/k$; the minus sign turns default log-odds into good:bad log-odds
  - checking account (training WoE): no account 66 points, balance $<$ 0 77, 0–200 DM 97, at least 200 DM or salary 110
- Prior correction in points: for a population with 5% bads the default log-odds fall by 2.097, so the good:bad log-odds rise by 2.097 and every score rises by $28.85 \times 2.097 = 60.5$ points

SFM_ch12_scorecard

Scores of the Test Sample



- 300 test credits; scores from 422 to 607 points; mean: goods 529, bads 492; dashed: the cut-off at PD $= 1/6$ (Section 7), 534 points
- Interpretation: the two distributions are shifted but overlap: every cut-off rejects some goods and accepts some bads; the validation measures quantify this overlap

SFM_ch12_scorecard

The Confusion Matrix

	predicted bad (reject)	predicted good (accept)
actual bad	TP	FN
actual good	FP	TN

- A cut-off c: reject if PD $> c$
- TP (true positive): a bad rejected; FN (false negative): a bad accepted
- FP (false positive): a good rejected; TN (true negative): a good accepted

- Rates
  - **TPR** (true positive rate, sensitivity) $= \mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$: share of bads caught
  - **FPR** (false positive rate) $= \mathrm{FP}/(\mathrm{FP}+\mathrm{TN}) = 1-$ specificity: share of goods rejected
  - precision $= \mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$; accuracy $= (\mathrm{TP}+\mathrm{TN})/n$
- Every rate depends on c; the measures that follow summarise all cut-offs at once

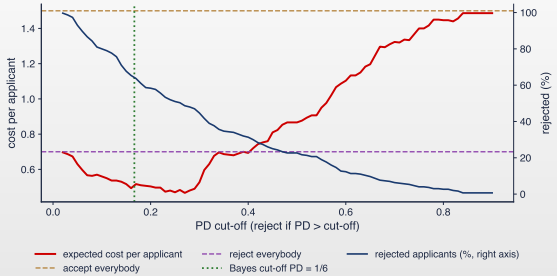
Two Cut-Offs on the Test Sample

cut-off	TP	FN	FP	TN	accuracy	TPR	FPR	cost
PD > 0.5	43	47	25	185	76.0%	47.8%	11.9%	0.867
PD $> 1/6$	82	8	112	98	60.0%	91.1%	53.3%	0.507

- Cost per applicant $= (5\,\mathrm{FN} + 1\,\mathrm{FP})/n$: accepting a bad costs five times as much as rejecting a good Grömping (2019)
  - the Bayes cut-off minimises the expected cost: reject if $p\,C_{FN} > (1-p)\,C_{FP}$, i.e. $p > C_{FP}/(C_{FP}+C_{FN}) = 1/6$
- Interpretation: the cut-off of 1/6 has a lower accuracy (60.0% against 76.0%) but a much lower cost (0.507 against 0.867): it catches 91.1% of the bads

SFM_ch12_validation_irb

Expected Cost and the Choice of the Cut-Off



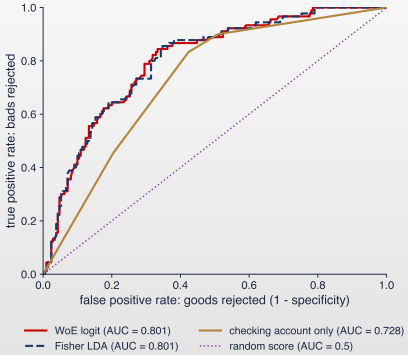
- Test sample; red: cost per applicant; dashed: accept everybody (1.50) or reject everybody (0.70); blue: share rejected (right axis)
- Interpretation: the cost is flat between PD 0.1 and 0.3; the minimum on these 300 credits is at 0.27, close to the Bayes cut-off of 1/6, which is chosen before seeing the test data

SFM_ch12_validation_irb

ROC Curve and AUC

- **ROC** (receiver operating characteristic): the points (FPR(c), TPR(c)) for all cut-offs c
  - from (0, 0) (reject nobody) to (1, 1) (reject everybody); the diagonal is a random score
- **AUC** (area under the curve) $= P(s_{\text{bad}} > s_{\text{good}}) + \frac{1}{2} P(s_{\text{bad}} = s_{\text{good}})$ Hanley & McNeil (1982)
  - s: the risk score (here the PD); compute it by counting pairs: the share of (bad, good) pairs ordered correctly
  - the Mann–Whitney statistic divided by $n_1 n_0$; 0.5: no discrimination, 1: perfect
- **Gini** $= 2\,\mathrm{AUC} - 1$; AUC does not depend on the bad rate or on a monotone rescaling of the score
  - standard error and tests of AUC: DeLong et al. (1988)

ROC Curves on the Test Sample



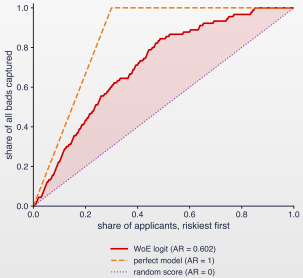
- 300 test credits, 18 900 (bad, good) pairs
- WoE logit: AUC $= 0.801$, 95% CI $[0.749, 0.854]$ (DeLong, se 0.027); Gini $= 0.603$
- Checking account alone: AUC $= 0.728$; DeLong test against the scorecard: $z = 2.86$, $p = 0.004$
- Interpretation: a random bad has a higher PD than a random good in about 80% of the pairs; the other five variables add a significant improvement over the checking account

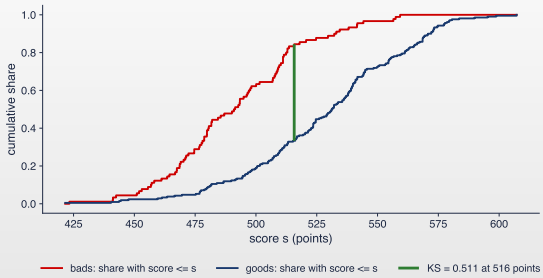
SFM_ch12_validation_irb

CAP Curve, Accuracy Ratio and KS

- **CAP** (cumulative accuracy profile): sort the applicants from the riskiest; plot the share of all bads among the first x%
  - perfect model: all bads first, the curve reaches 1 at $x =$ bad rate; random: the diagonal
  - **AR** (accuracy ratio) $= \dfrac{\text{area between model CAP and diagonal}}{\text{area between perfect CAP and diagonal}}$; for a continuous score AR = Gini
- **KS** (Kolmogorov–Smirnov): $\max_s |F_{\text{bad}}(s) - F_{\text{good}}(s)|$, the largest gap between the score distributions of bads and goods
  - F_{bad}, F_{good}: the empirical distribution functions of the scores in the two groups; KS $\in [0, 1]$
  - read on the ROC curve: KS $= \max_c (\mathrm{TPR} - \mathrm{FPR})$; the score where it occurs is a natural first cut-off
- All three measure **discrimination** (ranking), not calibration

CAP and KS on the Test Sample





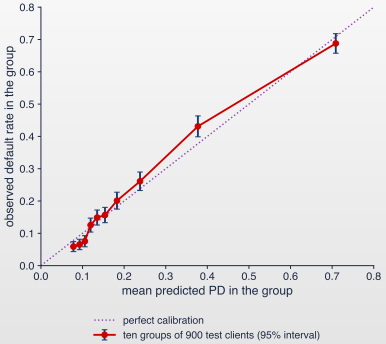
- AR = 0.602 (Gini = 0.603: the same up to rounding); KS = 0.511 at 516 points
- Interpretation: rejecting the riskiest 30% catches 60% of the bads (54 of 90); at 516 points the scorecard separates bads and goods best

SFM_ch12_validation_irb

Calibration and the Brier Score

- **Calibration**: among applicants with PD $\approx p$, a share of about p defaults
  - reliability diagram: sort by PD, form groups (e.g. ten), plot the mean PD against the observed default rate
  - Hosmer–Lemeshow Hosmer & Lemeshow (1980): $\sum_g \frac{(O_g - n_g \bar{p}_g)^2}{n_g \bar{p}_g (1 - \bar{p}_g)} \approx \chi^2(G-2)$
  - G groups; n_g: size, O_g: observed defaults, $\bar{p}_g$: mean PD of group g; a large value: poor calibration
- **Brier score** Brier (1950): $\mathrm{BS} = \frac{1}{n}\sum_i (p_i - y_i)^2$, the mean squared error of the PDs; lower is better
  - reference: a constant PD equal to the training bad rate; skill $= 1 - \mathrm{BS}/\mathrm{BS}_{\mathrm{ref}}$
  - scorecard test sample: BS $= 0.162$ against 0.210 (skill 23%); HL with 10 groups $= 6.520$, p $= 0.589$
- Brier rewards both ranking and calibration; AUC sees only the ranking

Calibration on the Taiwan Data



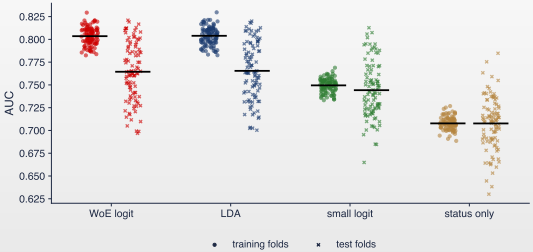
- Logit on 30 000 card holders (70% estimation); test sample of 9 000; ten groups of 900 by PD; bars: 95% intervals
- Test AUC $= 0.777$ (estimation 0.774); Brier $= 0.136$ against 0.172 (skill 21%)
- Hosmer–Lemeshow $= 40.660$, $p < 0.001$: rejected, although the points lie close to the diagonal
- Interpretation: with 9000 clients the test detects small deviations (the safest group: PD 7.8%, observed 5.9%); judge the size of the gaps, not only the p-value

SFM_ch12_validation_irb

Out-of-Sample and Cross-Validation

- **In-sample** performance is optimistic: the model has adapted to the noise of the training data (overfitting)
  - one test sample of 300 credits gives AUC $= 0.801$ with a standard error of 0.027: one split is noisy
- k**-fold cross-validation**: split into $k = 5$ folds; estimate on four, test on the fifth; rotate; repeat 20 times
  - the whole procedure is repeated inside every fold: bins, WoE, IV selection and coefficients
  - report the mean and the spread of the 100 test AUCs, and the gap between training and test AUC
- **Out-of-time** validation: estimate on older applications, test on later ones (two slides ahead)

Training and Test AUC in Repeated Cross-Validation



- 20 repetitions of 5 folds; dots: training folds, crosses: test folds; black lines: means
- WoE logit: training 0.804, test 0.764 (sd 0.032); LDA: 0.804, 0.765; small logit: 0.749, 0.744; checking account: 0.708
- Interpretation: the scorecard loses 0.039 AUC out of sample, the five-variable logit almost nothing; the single split above was lucky (0.801 against 0.764 on average)

SFM_ch12_validation_irb

Out-of-Time Validation and Drift

- A scorecard is used for years on applicants who arrive **after** the estimation period
  - the economy, the products and the applicants change: the relation between x and default drifts
  - out-of-time test: estimate on, say, 2019–2021, test on 2022–2023; compare AUC and calibration across years
- Monitoring: the distribution of the scores and of each variable is compared with the training period
  - a shift with stable ranking calls for recalibration (the constant); a loss of ranking calls for a new model
- Data limitation: neither data set of this chapter has application dates, so an out-of-time test is not possible here; random splits are the most we can do
- Further reading, an out-of-time design for crypto "zombie" assets: Będowska-Sójka, Wójcik & Pele (2026)

Case Study: Benchmarking Credit Scoring Methods

- Lessmann et al. (2015): 41 classifiers, six accuracy measures, eight retail credit data sets (among them the German credit data of UCI)
  - an update of the benchmark of Baesens et al. (2003), ten years later
- Findings
  - several classifiers predict significantly better than logistic regression, the industry standard; heterogeneous ensembles perform best
  - more accurate scorecards can bring sizeable financial returns
  - the common accuracy measures give similar signals; random forests are recommended as the benchmark for any new method
- Beating the logit alone is no longer evidence of progress; on our 1000 credits, logit and LDA differ by less than one standard error of AUC; Chapter 13 adds trees and neural networks

Reject Inference

- The outcome is known only for **accepted** applicants; the model is used on **all** applicants ("through the door")
  - a sample selected by the old score: the riskiest applicants are missing, and so are their outcomes
  - the South German Credit borrowers had all passed the bank's checks Grömping (2019)
- **Reject inference**: methods that try to correct for the missing outcomes Banasik, Crook & Thomas (2003); Thomas, Crook & Edelman (2017)
  - reweighting: give accepted applicants who resemble the rejected ones a larger weight
  - augmentation and parcelling: impute outcomes for the rejected from a model
  - the most reliable source: a small random sample of applicants accepted regardless of the score

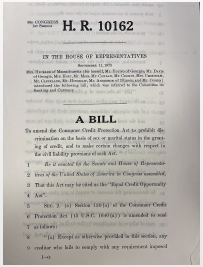
A Reject-Inference Experiment on the Taiwan Data

- An "old score" (late payment in September and the limit) accepts the best 70% of the training clients (15 085); their default rate is 13.2%, that of the rejected 44.9%
  - among the accepted nobody was late in September: the strongest variable cannot be estimated and drops out
- Test sample (all applicants): AUC of the model estimated on the accepted only 0.736, on all applicants 0.777, old score 0.743
  - rejected test clients: observed default rate 46.1%; mean PD of the accepted-only model 29.6%, of the full model 44.1%
- Interpretation: trained on the accepted only, the new model ranks worse than the old score and underestimates the risk of the rejected by about 36%

SFM_ch12_reject_fairness

Fairness in Credit Scoring

- US law forbids discrimination in credit by sex, marital status, race, religion, national origin or age Equal Credit Opportunity Act (FTC)
  - EU: credit scoring of persons is a high-risk use of AI (documentation, human oversight) European Commission, AI Act
- Leaving the protected variable out is not enough: other variables can stand in for it
  - **demographic parity**: equal approval rates across groups
  - **equal opportunity**: equal approval rates of the good payers Hardt, Price & Srebro (2016)
  - **calibration within groups**: the same PD means the same risk in every group
- When default rates differ across groups, these criteria cannot all hold at once: the choice is a matter of policy for banks and regulators



H. R. 10162

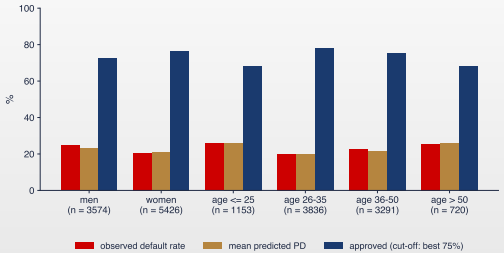
IN THE HOUSE OF REPRESENTATIVES

A BILL

H.R. 10162 (1973), the bill that introduced the Equal Credit Opportunity Act

Photo: H.R. 10162 (1973); public domain; Wikimedia Commons

Approval and Default by Group: Taiwan



- Test sample; sex and age are not inputs of the model; cut-off: approve the 75% with the lowest PD
- Women: default rate 20.5%, approved 76.6%; men: 24.7%, 72.5%; good payers approved: women 85.1%, men 83.2%
- Interpretation: the model is calibrated in every group (mean PD close to the observed rate), so groups with more defaults (men, the youngest and the oldest) are approved less often

SFM_ch12_reject_fairness

Case Study: Machine Learning and Inequality in Credit Access

- Fuster et al. (2022): US mortgages; default predicted with a traditional logit and with machine-learning models
  - question: who gains and who loses when the lender switches to the more accurate model?
- Findings
  - the flexible models predict default better
  - Black and Hispanic borrowers are less likely to gain from the new technology
  - in an equilibrium model, the disparity in interest rates between and within groups increases, mainly because of the greater flexibility
- Lesson: accuracy and fairness must be measured together; a better AUC is not a sufficient argument

The Altman Z-Score (1968)

- Altman (1968): 66 US manufacturing firms, 33 that filed for bankruptcy in 1946–1965 and 33 matched survivors
  - Fisher's discriminant analysis on five accounting ratios
- $Z = 1.2X_1 + 1.4X_2 + 3.3X_3 + 0.6X_4 + 1.0X_5$
  - X_1 working capital / total assets (TA); X_2 retained earnings / TA; X_3 EBIT (earnings before interest and taxes) / TA
  - X_4 market value of equity / book value of debt; X_5 sales / TA
  - zones: $Z < 1.81$ distress, 1.81–2.99 grey, $Z > 2.99$ safe
- About 95% of the sample correctly classified one year before bankruptcy; the accuracy falls quickly for longer horizons



NYU Stern School of Business, Altman's institution

Photo: Ajay Suresh (2019); CC BY 2.0; Wikimedia Commons

Worked Example: a Z-Score

- A hypothetical firm: $X_1 = 0.15$, $X_2 = 0.20$, $X_3 = 0.08$, $X_4 = 0.90$, $X_5 = 1.10$
  - $Z = 1.2 \times 0.15 + 1.4 \times 0.20 + 3.3 \times 0.08 + 0.6 \times 0.90 + 1.0 \times 1.10$
  - $Z = 0.180 + 0.280 + 0.264 + 0.540 + 1.100 = 2.364$: the grey zone
- Which ratio would move the firm to the safe zone?
  - an EBIT/TA of 0.27 instead of 0.08 adds $3.3 \times 0.19 \approx 0.63$: $Z \approx 2.99$
- Limits
  - the weights come from 66 US manufacturers of 1946–1965; other sectors and countries need re-estimation
  - Z is a discriminant score, not a PD; later models give PDs: logit Ohlson (1980), hazard models Shumway (2001)

Merton's Distance to Default

- Merton (1974): the value V_t of the firm's assets follows a GBM (geometric Brownian motion, Chapter 4) with drift μ and volatility σ_V
  - debt with face value D due at T; default if $V_T < D$; equity is a call option on the assets
- $\ln V_T \sim N\big(\ln V_0 + (\mu - \sigma_V^2/2)T,\ \sigma_V^2 T\big)$, so
  - $\mathrm{DD} = \dfrac{\ln(V_0/D) + (\mu - \sigma_V^2/2)T}{\sigma_V\sqrt{T}}, \quad \mathrm{PD} = \Phi(-\mathrm{DD})$
  - DD: the number of standard deviations between the expected asset value and the debt
- V and σ_V are not observed: they are solved from the equity price and the equity volatility (option pricing)



Robert C. Merton, 2010

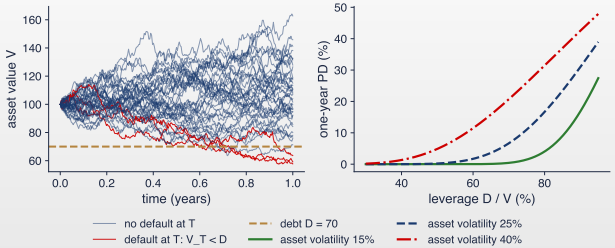
Photo: Massachusetts Institute of Technology (2010); CC BY-SA 4.0; Wikimedia Commons

Worked Example: Distance to Default

- $V_0 = 100$, $D = 70$, $\sigma_V = 25\%$, $\mu = 5\%$, $T = 1$ year
  - $\ln(100/70) = 0.3567$; $\mu - \sigma_V^2/2 = 0.05 - 0.03125 = 0.01875$
  - $\mathrm{DD} = 0.3754/0.25 = 1.502$; $\mathrm{PD} = \Phi(-1.502) = 6.66\%$
- From the market: equity 40, equity volatility 50%, debt 70, risk-free rate 3%
  - solving $E = V\Phi(d_1) - De^{-rT}\Phi(d_2)$ and $\sigma_E E = \Phi(d_1)\sigma_V V$: $V = 107.9$, $\sigma_V = 18.6\%$, $\mathrm{DD} = 2.39$, $\mathrm{PD} = 0.84\%$
  - E, σ_E: value and volatility of equity; r: the risk-free rate; d_1, d_2: the Black–Scholes terms of Chapter 4, with V in place of the share price
- Interpretation: a market-based PD reacts every day to prices; an accounting score changes once a year

SFM_ch12_altman_merton

Merton: Asset Paths and PD



- Left: 40 simulated asset paths ($V_0 = 100$, $\mu = 5\%$, $\sigma_V = 25\%$); red: the paths that end below the debt of 70
- Right: one-year PD against leverage D/V for three asset volatilities
- Interpretation: PD is almost zero at low leverage and rises steeply between 70% and 90%; the higher the volatility, the earlier the rise

SFM_ch12_altman_merton

Case Study: Predictive Accuracy of the Merton Model

- Bharath & Shumway (2008): US firms; the Merton DD against a "naive" DD with the same functional form but without solving the model
  - design: hazard models of default and out-of-sample forecasts; comparison with CDS (credit default swap) spreads and bond yields
- Findings
  - the naive DD performs slightly better than the Merton DD out of sample
  - other variables (e.g. past returns) add information: DD is not a sufficient statistic for PD
  - its functional form (leverage scaled by volatility) is what makes it useful
- Lesson: theory gives a good variable, statistics decides how to use it; the same holds for the Z-score

An Open Research Question

- **Do flexible machine-learning scores improve default prediction enough to justify their cost in transparency and fairness?**
  - today: WoE logit and LDA reach a test AUC of about 0.764; benchmarks find significant gains for ensembles Lessmann et al. (2015)
  - the gains may be unequal across groups Fuster et al. (2022); support vector machines for default: Chen, Härdle & Moro (2011)
- Why it is open: gains depend on the data set, the measure and the population; the outcomes of rejected applicants are never seen
- AI tools can speed up such a study, but its results must still be checked Wang et al. (2023)

Possible Uses of AI Tools

- **Literature**: a list of benchmark studies of credit scoring and of studies on fairness in lending, with their data sets
- **Code**: a first draft of a cross-validated comparison of the WoE logit with gradient boosting on the Taiwan data
- **Robustness**: other cut-offs, other measures (Brier, cost), group-wise AUC and calibration
- Example prompt
  - `Write Python code that compares a WoE logistic regression with gradient boosting on the Taiwan credit card default data by repeated stratified 5-fold cross-validation, recomputing the WoE bins inside each fold, and reports AUC, Brier score and the approval rates of men and women at a 75 percent approval cut-off.`

What to Check

- The coding of the outcome: default $= 1$; a reversed coding flips every sign and every curve
- The WoE sign convention: ln(goods/bads) here; some sources use ln(bads/goods)
- No leakage: bins, WoE and variable selection inside each training fold
- AUC is not accuracy, and Gini $= 2\,\text{AUC} - 1$, not AUC $- 0.5$
- PDs from an oversampled data set need the prior correction before any EL or capital number
- References: every cited paper and data set must exist; check the DOI

Project Idea

- **Question**: on the Taiwan credit card data, how much AUC does gradient boosting add over the WoE logit, and what happens to the approval rates by sex and age?
  - data: Default of Credit Card Clients (2009) (CC BY 4.0), saved in data/credit
- Steps
  - WoE bins and IV for every variable; a WoE logit; gradient boosting with default settings
  - repeated stratified cross-validation: AUC with DeLong intervals, Brier score, calibration
  - fairness at one cut-off: demographic parity and equal opportunity for sex and age groups
- Deliverable: one table, one chart, and a paragraph on what the data can and cannot show
- Declare any AI use, and list the errors of the AI that you corrected

Key Takeaways

- EL = PD × LGD × EAD is priced in; IRB capital covers the unexpected loss
- WoE turns every variable into log-odds evidence; IV ranks the variables
- Logit: e^{β} is an odds ratio; LDA (Fisher) gives the same linear log-odds under Normality and ranks equally well here
- A scorecard: Score = Offset + Factor ln(good:bad odds), PDO = 20 points per doubling
- Validate discrimination (AUC, Gini, KS) and calibration (Brier, reliability diagram) out of sample; choose the cut-off from costs
- Rejected applicants and protected groups need explicit checks; Altman and Merton add accounting and market information

Key Formulas

Quantity	Formula
EL	$\mathrm{PD} \times \mathrm{LGD} \times \mathrm{EAD}$
WoE, IV	$\ln \frac{g_i/G}{b_i/B}$, $\sum_i (g_i/G - b_i/B)\,\mathrm{WoE}_i$
Logit	$\ln \frac{p}{1-p} = \beta_0 + x^\top \beta$, $p = 1/(1+e^{-\eta})$
LDA	$w = S_W^{-1}(m_1 - m_0)$
Prior correction	$+\ln \frac{\pi/(1-\pi)}{\pi_s/(1-\pi_s)}$ to the constant
Scorecard	$\mathrm{Offset} + \frac{\mathrm{PDO}}{\ln 2} \ln \frac{1-p}{p}$
AUC, Gini	$P(s_{\mathrm{bad}} > s_{\mathrm{good}})$, $2\,\mathrm{AUC} - 1$
KS, Brier	$\max_s \lvert F_{\mathrm{bad}}(s) - F_{\mathrm{good}}(s)\rvert$, $\frac{1}{n}\sum (p_i - y_i)^2$
Bayes cut-off	$p > C_{FP}/(C_{FP} + C_{FN})$
Merton	$\mathrm{DD} = \frac{\ln(V/D) + (\mu - \sigma^2/2)T}{\sigma\sqrt{T}}$, $\mathrm{PD} = \Phi(-\mathrm{DD})$

Self-Assessment

- **Question**: a model has an accuracy of 70% on the South German Credit test sample. Is it good?
  - **Answer**: not necessarily: "everybody is good" already reaches 70%; look at AUC, TPR, FPR and the cost
- **Question**: the coefficient of a dummy in a logit is 0.69. What does it mean?
  - **Answer**: the odds of default are multiplied by $e^{0.69} \approx 2$ for the group, relative to the reference group
- **Question**: does correcting the PDs for oversampling change the AUC?
  - **Answer**: no: it shifts all log-odds by the same constant, so the ranking is unchanged
- Next: Chapter 13, machine learning

References (1/2)

- Altman, E.I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance*, 23(4), 589–609.
- Banasik, J., Crook, J., Thomas, L. (2003). Sample selection bias in credit scoring models. *Journal of the Operational Research Society*, 54(8), 822–832.
- Basel Committee on Banking Supervision (2005). An explanatory note on the Basel II IRB risk weight functions. Bank for International Settlements.
- Basel Committee on Banking Supervision (2006). International convergence of capital measurement and capital standards: a revised framework (comprehensive version). Bank for International Settlements.
- Będowska-Sójka, B., Wójcik, P., Pele, D.T. (2026). Early warning systems for cryptocurrency markets: predicting 'zombie' assets using machine learning. *North American Journal of Economics and Finance*, 102543.
- Bharath, S.T., Shumway, T. (2008). Forecasting default with the Merton distance to default model. *Review of Financial Studies*, 21(3), 1339–1369.
- Brier, G.W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1–3.
- Chen, S., Härdle, W.K., Moro, R.A. (2011). Modeling default risk with support vector machines. *Quantitative Finance*, 11(1), 135–154.
- Cox, D.R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society, Series B*, 20(2), 215–232.
- Default of Credit Card Clients [data set] (2009). UCI Machine Learning Repository, CC BY 4.0.
- DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44(3), 837–845.
- Durand, D. (1941). *Risk Elements in Consumer Instalment Financing*, technical edition. National Bureau of Economic Research.
- European Commission (2024). AI Act: regulatory framework on artificial intelligence. Shaping Europe's digital future.
- Federal Trade Commission. Equal Credit Opportunity Act, 15 U.S.C. §§ 1691–1691f.
- Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2), 179–188.
- Franke, J., Härdle, W.K., Hafner, C.M. (2019). *Statistics of Financial Markets: An Introduction*, 5th ed. Springer.
- Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., Walther, A. (2022). Predictably unequal? The effects of machine learning on credit markets. *Journal of Finance*, 77(1), 5–47.

References (2/2)

- Gordy, M.B. (2003). A risk-factor model foundation for ratings-based bank capital rules. *Journal of Financial Intermediation*, 12(3), 199–232.
- Grömping, U. (2019). South German credit data: correcting a widely used data set. Reports in Mathematics, Physics and Chemistry 4/2019, Beuth University of Applied Sciences Berlin.
- Hand, D.J., Henley, W.E. (1997). Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society, Series A*, 160(3), 523–541.
- Hanley, J.A., McNeil, B.J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1), 29–36.
- Härdle, W.K., Simar, L. (2019). *Applied Multivariate Statistical Analysis*, 5th ed. Springer.
- Hardt, M., Price, E., Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems* 29; arXiv:1610.02413.
- Hosmer, D.W., Lemeshow, S. (1980). Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics – Theory and Methods*, 9(10), 1043–1069.
- Lessmann, S., Baesens, B., Seow, H.-V., Thomas, L.C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: an update of research. *European Journal of Operational Research*, 247(1), 124–136.
- Merton, R.C. (1974). On the pricing of corporate debt: the risk structure of interest rates. *Journal of Finance*, 29(2), 449–470.
- myFICO. The history of the FICO score. Fair Isaac Corporation.
- Ohlson, J.A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, 18(1), 109–131.
- Shumway, T. (2001). Forecasting bankruptcy more accurately: a simple hazard model. *Journal of Business*, 74(1), 101–124.
- Siddiqi, N. (2006). *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*. Wiley.
- South German Credit [data set] (2020). UCI Machine Learning Repository, CC BY 4.0.
- Thomas, L.C., Crook, J., Edelman, D. (2017). *Credit Scoring and Its Applications*, 2nd ed. SIAM.
- Wang, H., Fu, T., Du, Y., Gao, W., et al. (2023). Scientific discovery in the age of artificial intelligence. *Nature*, 620, 47–60.
- Yeh, I.-C., Lien, C.-H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2), 2473–2480.