








Statistics of Financial Markets

Seminar 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

- **AI**: Artificial Intelligence
- **AR**: Accuracy Ratio (of the CAP curve)
- **AUC**: Area Under the (ROC) Curve
- **CAP**: Cumulative Accuracy Profile
- **CI**: Confidence Interval
- **DM**: Deutsche Mark (the German mark)
- **EAD**: Exposure At Default
- **FN**: False Negative
- **FPR**: False Positive Rate
- **HL**: Hosmer–Lemeshow (test)
- **IRB**: Internal Ratings-Based (approach)
- **KS**: Kolmogorov–Smirnov (statistic)
- **LDA**: Linear Discriminant Analysis
- **LGD**: Loss Given Default
- **PD**: Probability of Default
- **PDO**: Points to Double the Odds
- **ROC**: Receiver Operating Characteristic (curve)
- **RWA**: Risk-Weighted Assets
- **TN**: True Negative
- **TP**: True Positive
- **TPR**: True Positive Rate
- **UCI**: University of California, Irvine (Machine Learning Repository)

Today's Question and Seminar Structure

- Question of the seminar
  - **how does a bank turn the data of an applicant into a probability of default, and how do we judge that probability?**
- Outputs to produce
  - the answers to the A exercises
  - the table or figure each B exercise asks for
  - a short argued answer to C1; the errors found in the AI answer of C2 and the corrected answer
  - the interpretation sentences each exercise asks for
- Nothing is handed in: the seminar is for practice; the solutions of the [Proposed] exercises are discussed in class
- Notebook for today: open the seminar notebook in Google Colab
  - each task names its notebook section
- Exercises, their types and their order: the Exercise Map (next slide)
- Contact: Prof. Daniel Traian Pele, danpele@ase.ro

Exercise Map

Task	Question	Type	Model
A1, A2	the PD of an applicant from a logit model; odds ratios	Solved, Proposed	A1
A3, A4	WoE and IV from a table of goods and bads	Solved, Proposed	A3
A5, A6	confusion matrix, AUC by counting pairs, Gini, KS, Brier score	Solved, Proposed	A5
A7, A8	scorecard scaling and points; prior correction and expected loss	Solved, Proposed	A7
B1, B2	IV and a WoE logit: South German Credit; IV of the Taiwan data	Solved, Proposed	B1
B3, B4	validation of the scorecard; logit against LDA on the Taiwan data	Solved, Proposed	B3
B5, B6	expected loss and IRB capital of a portfolio; fairness by sex and age	Solved, Proposed	B5
C1, C2	does gradient boosting beat the logit? what is wrong in an AI answer?	Proposed	B3, A5

- **[Solved]**: full solution in the slides and in the notebook, a model to follow; **[Proposed]**: you solve it after the model; the solution is discussed in the seminar
- A: calculations on paper; B: real data, in the notebook, each ending with an interpretation question; C: open questions, among them the audit of an AI answer

Data Used

Data set	Source	Size	Outcome
South German Credit	South German Credit (2020)	1000 credits, 1973–1975	300 bad, 700 good
Taiwan credit cards	Default of Credit Card Clients (2009)	30 000 card holders, 2005	default in October 2005 (22.1%)

- Both under CC BY 4.0, saved once in data/credit of the course repository; South German Credit is the corrected version of the UCI German credit data Grömping (2019)
- Default = 1 for a bad credit or a missed payment, 0 otherwise; bad credits are oversampled in South German Credit (population rate about 5%)
- In the notebook: `load_credit('sgc')`, `woe_table(x, y, var)`, `logit_fit(X, y)`, `auc(score, y)`; no account or key is needed

Prerequisites for Today (1/17): Default and Expected Loss

- **Default**: the borrower does not pay as agreed (Basel: more than 90 days past due, or unlikely to pay) Basel Committee (2006)
  - $Y = 1$ for a default (a "bad" payer), $Y = 0$ for a good payer
  - a **credit score** orders the applicants by risk; a **scoring model** turns their characteristics into a PD
- Three risk parameters of a loan
  - **PD** (probability of default): the probability of default within one year
  - **LGD** (loss given default): the share of the exposure lost if default occurs
  - **EAD** (exposure at default): the amount owed at default
- **Expected loss**

$$\mathrm{EL} = \mathrm{PD} \times \mathrm{LGD} \times \mathrm{EAD}$$

  - the average loss of the loan; it is covered by provisions and priced into the interest rate
  - for a portfolio: the sum of the EL of all loans; as a share of the total EAD

Prerequisites for Today (2/17): Basel IRB Capital

- **IRB** (internal ratings-based) capital for other retail loans, per unit of EAD Basel Committee (2006)

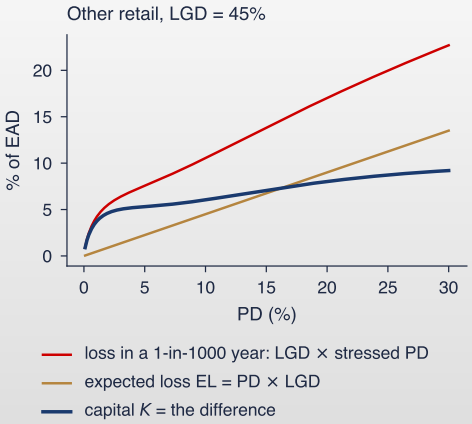
$$K = \mathrm{LGD}\Big[\Phi\Big(\frac{\Phi^{-1}(\mathrm{PD}) + \sqrt{R}\,\Phi^{-1}(0.999)}{\sqrt{1-R}}\Big) - \mathrm{PD}\Big]$$

  - Φ, Φ^{-1}: the $N(0,1)$ distribution function and its inverse
  - $\Phi(\cdot)$: the PD in a bad year that occurs once in 1000 years; minus PD: capital covers only the unexpected loss

- Correlation with the economy, from 0.16 (small PD) to 0.03 (large PD)

$$R = 0.03\,w + 0.16\,(1-w), \qquad w = \frac{1-e^{-35\,\mathrm{PD}}}{1-e^{-35}}$$

  - **RWA** (risk-weighted assets) $= 12.5\,K\,\mathrm{EAD}$; risk weight $=$ RWA/EAD



SFM_ch12_seminar

Prerequisites for Today (3/17): Odds and Log-Odds

- For a probability of default p

$$\text{odds} = \frac{p}{1-p}, \qquad \text{log-odds} = \ln \frac{p}{1-p}, \qquad \text{good:bad odds} = \frac{1-p}{p}$$

  - odds: how many defaults per non-default; $p = 0.2$: odds 0.25, log-odds -1.386, good:bad odds 4:1
  - the log-odds range over the whole real line, while p stays in $(0, 1)$

- Back from odds to probability

$$p = \frac{\text{odds}}{1 + \text{odds}}, \qquad p = \frac{1}{1 + \text{good:bad odds}}$$

  - doubling the odds is not doubling p: from odds 0.25 to 0.5, p rises from 0.2 to 0.333

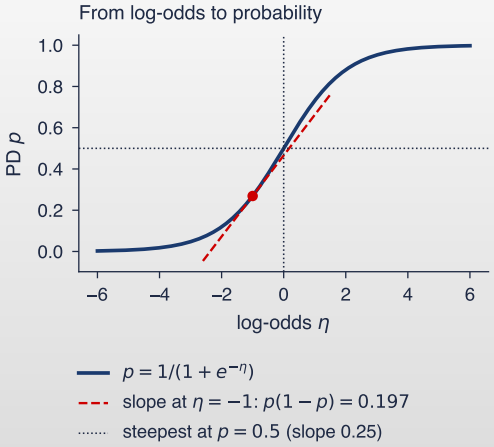
Prerequisites for Today (4/17): The Logit Model

- **Logit model** Cox (1958): the log-odds are linear in the characteristics

$$\ln \frac{p}{1-p} = \eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k, \qquad p = \frac{1}{1+e^{-\eta}}$$

  - $x_1, \ldots, x_k$: the characteristics of the applicant; η: the linear predictor; β_0: the constant (intercept)

- Reading the coefficients
  - β_j: the change in the log-odds per unit of x_j; $\beta_j > 0$: higher risk
  - e^{β_j}: the **odds ratio**, the factor that multiplies the odds; the same for every applicant
  - effect on the probability: $\Delta p \approx p(1-p)\,\beta_j$, largest near $p = 0.5$



Prerequisites for Today (5/17): Estimating the Logit

- **Maximum likelihood**: each applicant i contributes the log of the probability of what happened

$$\ell(\beta) = \sum_{i=1}^{n} \Big[y_i \ln p_i + (1 - y_i) \ln(1 - p_i) \Big], \qquad p_i = \frac{1}{1 + e^{-\eta_i}}$$

  - $y_i \in \{0, 1\}$: the observed outcome; $\hat{\beta}$ maximises ℓ (numerically); **se**: the standard errors, from the curvature of ℓ
  - $z = \hat{\beta}_j / \mathrm{se}_j$: $|z| > 1.96$ means a significant variable at 5%

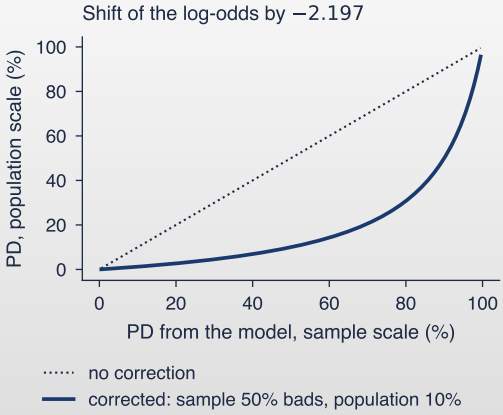
- **Training and test samples**
  - the model is estimated on the training sample (for example 70%); the test sample (30%) plays the future applicants and is used once, at the end
  - **stratified** split: the same bad rate in both samples
  - **overfitting**: a model much better on the training than on the test sample has learned noise

Prerequisites for Today (6/17): Oversampling and the Prior Correction

- **Oversampling**: the sample has a bad rate π_s, the population a smaller rate π
  - the logit estimated on the sample gives PDs on the scale of the sample, too high for the bank
- **Prior correction**: add the same constant to every log-odds

$$\eta_{\textbf{pop}} = \eta_{\textbf{sample}} + \ln \frac{\pi/(1-\pi)}{\pi_s/(1-\pi_s)}$$

  - then $p_{\textbf{pop}} = 1/(1+e^{-\eta_{\textbf{pop}}})$
  - the order of the applicants does not change, the level of the PDs does
  - the chart uses illustrative rates (50% and 10%)



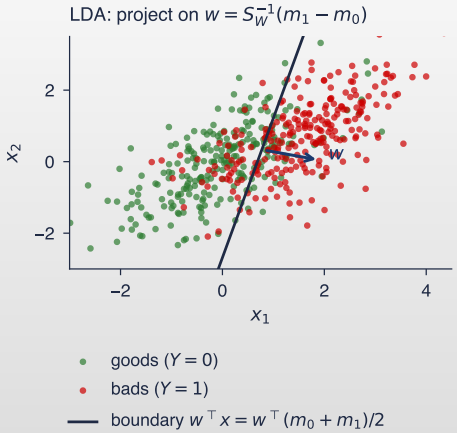
Prerequisites for Today (7/17): Linear Discriminant Analysis

- **LDA** (linear discriminant analysis) Fisher (1936): the direction that separates the two groups best

$$w = S_W^{-1}(m_1 - m_0), \qquad S_W = \frac{(n_0 - 1)S_0 + (n_1 - 1)S_1}{n_0 + n_1 - 2}$$

  - m_1, m_0: the mean vectors of bads and goods; S_1, S_0: their covariance matrices; n_1, n_0: their sizes
  - S_W: the common (pooled) covariance; the score is $w^\top x$

- From score to PD
  - Fisher's direction needs no distribution; if both groups are Normal with the same covariance, the log-odds are linear in x, as in the logit
  - far from Normality, the LDA may rank well but give PDs at the wrong level



Prerequisites for Today (8/17): Weight of Evidence and Information Value

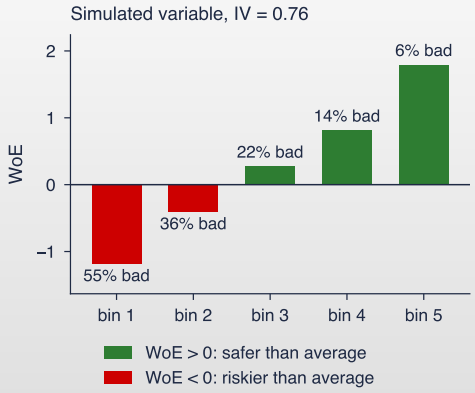
- Split a variable into bins; g_i, b_i: the goods and bads in bin i; G, B: their totals
  - **Weight of Evidence**:

$$\mathrm{WoE}_i = \ln \frac{g_i/G}{b_i/B}$$

  - g_i/G: the share of all goods that fall in bin i; positive WoE: safer than average, negative: riskier
  - a bin without goods or without bads: add 0.5 to both counts
- **Information Value**: the strength of the variable

$$\mathrm{IV} = \sum_i \Big(\frac{g_i}{G} - \frac{b_i}{B}\Big)\mathrm{WoE}_i \geq 0$$

  - rule of thumb Siddiqi (2006): below 0.02 useless, 0.02–0.1 weak, 0.1–0.3 medium, above 0.3 strong
  - a **WoE logit**: every variable is replaced by the WoE of its bin, computed on the training sample only



Prerequisites for Today (9/17): Cut-Offs, the Confusion Matrix and Costs

- A **cut-off** c: reject the applicant if PD $> c$
  - **TP** (true positive): bad rejected; **FN** (false negative): bad accepted
  - **FP** (false positive): good rejected; **TN** (true negative): good accepted
- Rates of the **confusion matrix**

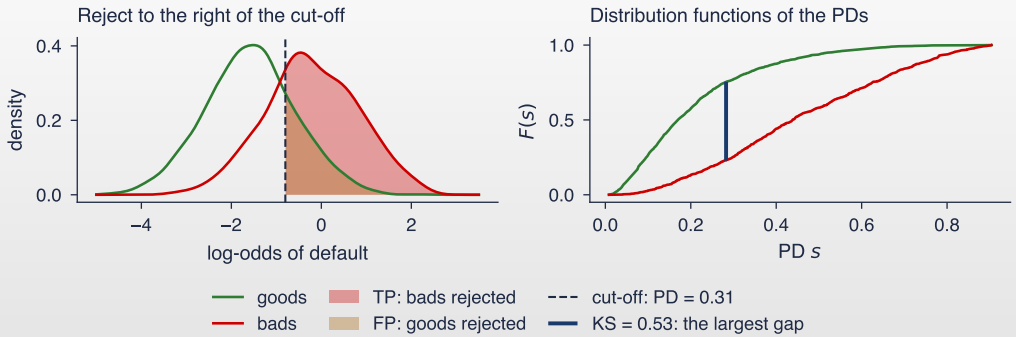
$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \qquad \mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}, \qquad \text{accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{n}$$

  - TPR: the share of bads caught; FPR: the share of goods rejected; n: the number of applicants
- **Cost per applicant**, with C_{FN}, C_{FP} the costs of one error of each kind

$$\frac{C_{FN}\,\mathrm{FN} + C_{FP}\,\mathrm{FP}}{n}$$

  - **Bayes cut-off**: reject when the expected cost of accepting, $p\,C_{FN}$, exceeds that of rejecting, $(1-p)\,C_{FP}$, i.e. if $p > C_{FP}/(C_{FP} + C_{FN})$
  - example: $C_{FN} = 5$, $C_{FP} = 1$ gives the cut-off $1/6$

Prerequisites for Today (10/17): Score Distributions, Cut-Off and KS



- Simulated scores of goods and bads: moving the cut-off to the left catches more bads (TP) but also rejects more goods (FP)
- **KS** (Kolmogorov–Smirnov) statistic: the largest vertical gap between the two distribution functions of the PDs

$$\mathrm{KS} = \max_s \left| F_{\mathbf{bad}}(s) - F_{\mathbf{good}}(s) \right|, \qquad F(s) = \text{share of PDs} \leq s$$

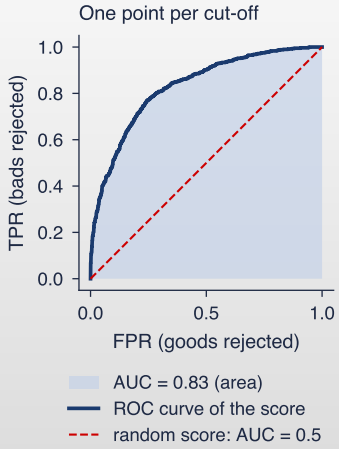
SFM_ch12_seminar

Prerequisites for Today (11/17): The ROC Curve and the AUC

- **ROC** (receiver operating characteristic) curve: (FPR, TPR) for every cut-off; **AUC**: the area under it Hanley & McNeil (1982)

$$\mathrm{AUC} = \frac{1}{n_{\mathbf{1}}\, n_{\mathbf{0}}} \sum_{i:\, y_i=\mathbf{1}} \sum_{j:\, y_j=\mathbf{0}} \left(1\{p_i > p_j\} + \tfrac{\mathbf{1}}{\mathbf{2}} 1\{p_i = p_j\} \right)$$

  - the share of (bad, good) pairs in the right order; $n_{\mathbf{1}}$, $n_{\mathbf{0}}$: the numbers of bads and goods; $1\{\cdot\}$: 1 if true, 0 otherwise
  - P(PD of a random bad $>$ PD of a random good): 0.5: no discrimination, 1: perfect



Prerequisites for Today (12/17): Gini, CAP and the DeLong Test

- **Gini coefficient**: the AUC rescaled to $[-1, 1]$

$$\mathrm{Gini} = 2\,\mathrm{AUC} - 1$$

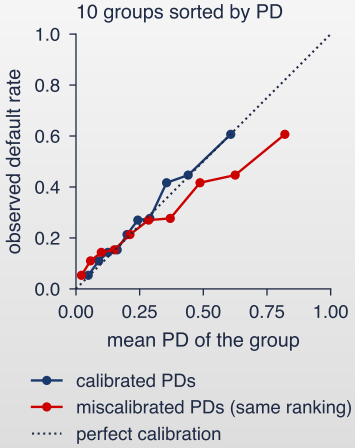
  - 0: a random score; 1: a perfect ranking
- **CAP** (cumulative accuracy profile) curve
  - sort the applicants from the riskiest; for each share x of applicants, plot the share of all bads among them

$$\mathrm{AR} = \frac{\text{area between the CAP of the model and the diagonal}}{\text{area between the perfect CAP and the diagonal}}$$

  - the **accuracy ratio** AR equals the Gini coefficient (up to ties)
- **DeLong** DeLong et al. (1988): the uncertainty of an AUC estimated on a test sample
  - a nonparametric standard error $\mathrm{SE}(\widehat{\mathrm{AUC}})$ from the pairs (bad, good); 95% interval: $\widehat{\mathrm{AUC}} \pm 1.96\,\mathrm{SE}$
  - two models on the same applicants: $z = (\widehat{\mathrm{AUC}}_{\mathbf{1}} - \widehat{\mathrm{AUC}}_{\mathbf{2}})/\mathrm{SE}(\text{difference})$, approximately $N(0,1)$; $|z| > 1.96$: different at 5%

Prerequisites for Today (13/17): Calibration

- **Calibration**: the mean PD of a group should equal its observed default rate
  - sort the applicants by PD, cut them into G groups of equal size and compare, in each group, the mean PD with the share of defaults
  - points on the diagonal: calibrated PDs
- Ranking and calibration are different properties
  - the red PDs order the applicants exactly like the blue ones (same AUC), but their level is wrong
  - a bank needs both: the ranking for the decision, the level for EL and capital



Prerequisites for Today (14/17): The Brier Score and the Hosmer–Lemeshow Test

- **Brier score** Brier (1950): the mean squared error of the PDs; lower is better

$$\mathrm{BS} = \frac{1}{n}\sum_{i=1}^{n}(p_i - y_i)^2$$

  - p_i: the PD of applicant i; $y_i \in \{0, 1\}$: the outcome; it rewards both ranking and calibration
  - benchmark: a constant PD equal to the bad rate π gives $\mathrm{BS} = \pi(1 - \pi)$
  - a confident wrong PD costs the most: a good payer with PD 0.95 adds $(0.95 - 0)^2 = 0.9025$

- **Hosmer–Lemeshow** test Hosmer & Lemeshow (1980) with G groups sorted by PD

$$\mathrm{HL} = \sum_{g=1}^{G}\frac{(O_g - E_g)^2}{E_g\,(1 - \bar{p}_g)} \approx \chi^2(G - 2)$$

  - O_g: the observed defaults; E_g: the expected defaults (the sum of the PDs); $\bar{p}_g$: the mean PD of group g
  - H_0: the PDs are calibrated; a large HL (small p-value): miscalibration

Prerequisites for Today (15/17): Scorecards

- **Scorecard** Siddiqi (2006): a score linear in the log of the good:bad odds

$$\text{Score} = \text{Offset} + \text{Factor} \times \ln(\text{good:bad odds})$$

  - **PDO** (points to double the odds): adding PDO points doubles the good:bad odds

$$\text{Factor} = \frac{\text{PDO}}{\ln 2}, \qquad \text{Offset} = \text{Base} - \text{Factor} \ln(\text{odds}_{\mathbf{0}})$$

  - Base: the score given at the good:bad odds $\text{odds}_{\mathbf{0}}$; from a score back to the PD: $\ln(\text{odds}) = (\text{Score} - \text{Offset})/\text{Factor}$

- **Points of an attribute** (WoE logit with k variables)

$$\text{points}_{ij} = -\left(\beta_j \, \text{WoE}_{ij} + \frac{\beta_{\mathbf{0}}}{k}\right) \text{Factor} + \frac{\text{Offset}}{k}$$

  - i: the attribute (bin) of variable j; the minus sign: the logit models the odds of default, the score the good:bad odds; the points of all variables add up to the score

Prerequisites for Today (16/17): Fairness of a Credit Decision

- **Fairness** at a cut-off Hardt, Price & Srebro (2016): compare groups A (for example men and women)
  - **demographic parity**: equal approval rates, $P(\text{approved} \mid A = a)$ the same for every group a
  - **equal opportunity**: equal approval rates of the good payers, $P(\text{approved} \mid Y = 0, A = a)$
  - **calibration within groups**: the same PD means the same default rate, $P(Y = 1 \mid \mathrm{PD} = p, A = a) = p$
- Remarks
  - when the default rates differ between groups, the criteria cannot all hold at once
  - leaving sex and age out of the model does not guarantee equal treatment: other variables can be correlated with them
  - approving the 75% with the lowest PD means a cut-off at the 75% quantile of the PDs

Prerequisites for Today (17/17): Flexible Models and Gradient Boosting

- **Gradient boosting**: a sum of many small decision trees, each fitted to the errors of the previous ones
  - a **decision tree** splits the applicants by simple rules (for example duration > 24 months) and gives each final group its own PD
  - captures non-linear effects and interactions without WoE bins
  - benchmarks report small but real gains over the logit Lessmann et al. (2015)
- Costs of flexibility
  - overfitting: a large gap between training and test AUC
  - no simple points per attribute: harder to explain a refusal to the applicant and to the supervisor
  - **cross-validation**: split the training sample into K parts, estimate on $K - 1$, evaluate on the remaining one, and average

A1: The PD of an Applicant [Solved]

Task

- Logit for default: constant -2.101; duration (years) 0.385; age (decades) -0.159; amount (1000 DM) 0.029; no checking account 1.859; balance $<$ 0 DM 1.338.
- Applicant: 2 years, 30 years old, 3000 DM, no checking account.
- 1. Compute the log-odds η and the PD.
- 2. Compute the odds of default and interpret $e^{\mathbf{1.859}}$.
- 3. Compute the PD with one more year of duration and compare it with $p(1-p)\beta$.
- Report: four numbers and one sentence.

Solution

- 1. $\eta = -2.101 + 0.385 \times 2 - 0.159 \times 3 + 0.029 \times 3 + 1.859 = 0.138$; PD $= 1/(1+e^{-\mathbf{0.138}}) = 53.4\%$
- 2. odds $= e^{\mathbf{0.138}} = 1.148$; $e^{\mathbf{1.859}} = 6.42$: without a checking account the odds are 6.42 times those of the reference group
- 3. $\eta + 0.385$: PD $= 62.8\%$ (+9.3 points); $p(1-p)\beta = 9.6$ points
- The odds ratio is constant; the change in PD depends on where the applicant is.

Back Exercise list

A2: A Second Applicant [Proposed]

Solution discussed in class

Task

- The same logit as in A1. Applicant: 1 year, 45 years old, 2000 DM, balance below 0 DM. Model: A1.
- 1. Compute η, the odds and the PD.
- 2. Compute the PD with one more year of duration.
- 3. Compare the change in PD with that of A1 and explain the difference.
- Report: four numbers and one sentence.

A3: WoE and IV of the Checking Account [Solved]

Task

- South German Credit, full sample: goods and bads by checking account: no account 139/135; below 0 DM 164/105; 0–200 DM 49/14; at least 200 DM or salary 348/46.
- 1. Compute g_i/G and b_i/B for each group.
- 2. Compute the WoE of each group.
- 3. Compute the IV and classify the variable.
- Report: a table and one sentence.

Solution

	g	b	g/G	b/B	WoE	IV
no account	139	135	0.199	0.450	−0.818	0.206
$<$ 0 DM	164	105	0.234	0.350	−0.401	0.046
0–200 DM	49	14	0.070	0.047	0.405	0.009
$\geq$ 200 DM / salary	348	46	0.497	0.153	1.176	0.404
total	700	300	1	1		0.666

- e.g. $\ln(0.497/0.153) = 1.176$; IV $= 0.666 > 0.3$: strong.
- WoE falls monotonically from the safest to the riskiest group.

SFM_ch12_seminar

Back Exercise list

A4: WoE and IV of the Savings [Proposed]

Solution discussed in class

Task

- Goods/bads by savings: none or unknown 386/217; below 100 DM 69/34; 100–500 DM 52/11; 500–1000 DM 42/6; at least 1000 DM 151/32. Model: A3.
- 1. Compute the WoE of each group.
- 2. Compute the IV and classify the variable.
- 3. Propose a merger of groups with similar WoE.
- Report: a table and two sentences.

Back Exercise list

A5: Confusion Matrix and AUC by Counting [Solved]

Task

- Ten applicants, PD and outcome ($Y = 1$: default): 0.05 (0), 0.10 (0), 0.12 (0), 0.20 (0), 0.25 (1), 0.30 (0), 0.40 (0), 0.55 (1), 0.60 (0), 0.80 (1).
- 1. Build the confusion matrix at the cut-off 0.5 and compute accuracy, TPR and FPR.
- 2. Compute the AUC by counting the (bad, good) pairs in the right order, and the Gini.
- 3. Compute the KS statistic and the Brier score.
- Report: seven numbers and one sentence.

Solution

- 1. TP 2, FN 1, FP 1, TN 6: accuracy 80.0%, TPR 66.7%, FPR 14.3%
- 2. goods below each bad: $4 + 6 + 7 = 17$ of 21 pairs: AUC $= 0.810$, Gini $= 0.619$
- 3. at PD ≤ 0.20: 4/7 of goods, 0/3 of bads: KS $= 0.571$; Brier $= 1.4819/10 = 0.148$
- The bad with PD 0.25 is ranked below three goods and the bad with 0.55 below one: 4 of 21 pairs are in the wrong order.

Back Exercise list

A6: A Second Sample [Proposed]

Solution discussed in class

Task

- Eight applicants: 0.08 (0), 0.15 (0), 0.22 (1), 0.35 (0), 0.42 (0), 0.50 (1), 0.65 (1), 0.90 (0). Model: A5.
- 1. Build the confusion matrix at the cut-off 0.4 and compute accuracy, TPR and FPR.
- 2. Compute the AUC by counting pairs, and the Gini.
- 3. Compute the KS statistic and the Brier score.
- Report: seven numbers and one sentence.

Back Exercise list

A7: Scaling a Scorecard [Solved]

Task

- PDO $= 20$; 600 points at good:bad odds 50:1. A WoE logit with $k = 6$ variables has constant $\beta_0 = -0.850$ and, for the checking account, $\beta = -0.829$; WoE: no account -0.826, at least 200 DM 1.043.
- 1. Compute the Factor and the Offset.
- 2. Compute the score of a PD of 5% and the PD of a score of 560.
- 3. Compute the points of the two checking-account groups.
- Report: six numbers and one sentence.

Solution

- 1. Factor $= 20/\ln 2 = 28.85$; Offset $= 600 - 28.85 \ln 50 = 487.12$
- 2. $487.12 + 28.85 \times \ln 19 = 572.08$; 560: ln odds $= 2.526$, odds 12.50:1, PD $= 7.4\%$
- 3. Offset$/k = 81.19$, $\beta_0/k = -0.1417$: no account $-(-0.829 \times (-0.826) + (-0.1417)) \times 28.85 + 81.19 = 65.5$; at least 200 DM: 110.2
- A well-funded account is worth 44.7 points more than no account: 2.2 doublings of the good:bad odds.

Back Exercise list

A8: Another Scale, Prior Correction and Expected Loss [Proposed]

Solution discussed in class

Task

- A bank uses PDO = 40 and 500 points at odds 20:1. A model built on a sample with 30% bads gives an applicant PD = 30%; the population bad rate is 5%; EAD = 10 000 DM, LGD = 45%. Model: A7.
- 1. Compute the Factor, the Offset, the score of a PD of 10% and the PD of a score of 520.
- 2. Correct the PD of 30% to the population.
- 3. Compute the expected loss with both PDs.
- Report: seven numbers and one sentence.

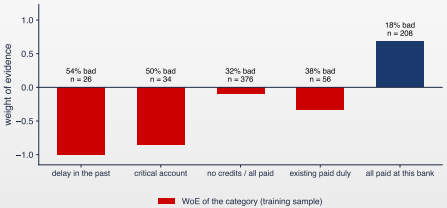
Back Exercise list

B1: A WoE Scorecard for South German Credit [Solved]

- **Question**: which variables carry information about default, and how well does a WoE logit rank new applicants?
- **Inputs**: South German Credit; a stratified split: 700 credits for estimation, 300 for testing (30% bads in both)
- **Tasks**
  1. Compute the WoE tables and the IV of the 20 variables on the training sample (duration, amount and age in five fixed bins).
  2. Keep the variables with IV of at least 0.1 and estimate a logit for default on their WoE values.
  3. Compute the AUC on the training and on the test sample, with a 95% DeLong interval for the test AUC.
  4. Draw the WoE of the credit-history groups.
  5. Interpretation: does the order of the credit-history groups make economic sense?
- **Report**: the list of selected variables with their IV, the coefficients, two AUCs with the interval, the chart and two sentences
- Notebook: section B1

Back Exercise list

B1: Solution [Solved]



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

- 6 variables with IV ≥ 0.1; AUC: training 0.789, test 0.801, 95% CI $[0.749, 0.854]$
- Interpretation: yes: a delay in the past (54% bad) and a critical account (50%) are the riskiest, all credits at this bank paid back duly (18%) the safest; in the uncorrected UCI labels this safest group was called "critical account"

SFM_ch12_seminar

B2: Information Value of the Taiwan Data [Proposed]

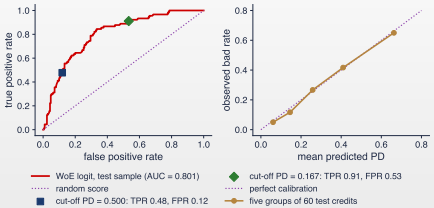
- **Question**: which kind of information predicts the default of a card holder: payment behaviour or personal data? Model: B1.
- **Inputs**: Taiwan credit cards, 30 000 clients; the same stratified 70/30 split
- **Tasks**
  1. Compute the IV of `PAY_0` and `PAY_2` (repayment status in September and August; categories -2 to 3 or more), of `LIMIT_BAL`, `PAY_AMT1`, `BILL_AMT1` and `AGE` (quintiles) and of `SEX`, `EDUCATION` and `MARRIAGE` (categories) on the training sample.
  2. Compute the default rate of each `PAY_0` category.
  3. Rank the variables by IV and classify them with the rule of thumb.
  4. Interpretation: why is the last month's repayment status so much stronger than the personal data?
- **Report**: a table of nine IVs, six default rates and two sentences
- Notebook: section B2

Back Exercise list

B3: Validating the Scorecard [Solved]

- **Question**: how good is the B1 scorecard on the test sample, and which cut-off should the bank use?
- **Inputs**: the 300 test credits of B1 with their PDs; costs: 5 for an accepted bad, 1 for a rejected good
- **Tasks**
  1. Build the confusion matrix at the cut-offs 0.5 and 1/6, with accuracy, TPR, FPR and the cost per applicant.
  2. Compute the AUC, the Gini, the accuracy ratio of the CAP curve and the KS statistic.
  3. Compute the Brier score, compare it with a constant PD of 30%, and run the Hosmer–Lemeshow test with five groups.
  4. Draw the ROC curve with the two cut-offs and the calibration plot.
  5. Interpretation: which cut-off would you recommend to the bank?
- **Report**: a table of two confusion matrices, six statistics, the chart and two sentences
- Notebook: section B3

B3: Solution [Solved]



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

- AUC = 0.801, Gini = 0.603, AR = 0.602, KS = 0.511; Brier = 0.162 against 0.210; HL with five groups = 0.589 (p = 0.899): no evidence of miscalibration
- Interpretation: the cut-off of 1/6: it rejects more applicants (64.7%) but catches 91.1% of the bads, and the cost falls from 0.867 to 0.507; accuracy is the wrong criterion when errors have different costs

SFM_ch12_seminar

B4: Logit or LDA on the Taiwan Data? [Proposed]

- **Question**: do the logit and Fisher's LDA differ in ranking or in calibration? Model: B3.
- **Inputs**: Taiwan, 70/30 split; 13 standardised inputs: log limit, age, indicators of the September repayment status, number of late months, utilisation, log payments
- **Tasks**
  1. Estimate a logit and an LDA on the training sample with the same inputs.
  2. Compute the test AUC of both and the DeLong test of equal AUCs.
  3. Compute the Brier score, the Hosmer–Lemeshow statistic (ten groups) and the mean PD of both, against the default rate.
  4. Compute the accuracy of the rule "nobody defaults" and of both models at the cut-off 0.5.
  5. Interpretation: is the LDA worse at ranking or at calibration?
- **Report**: one table and two sentences
- Notebook: section B4

Back Exercise list

B5: Capital Requirements from PDs [Solved]

- **Question**: how much expected loss and Basel capital does the test portfolio of B1 carry, and how much does the prior correction matter?
- **Inputs**: the 300 test credits of B1; EAD = credit amount; LGD = 45%; population bad rate 5%; other retail exposures
- **Tasks**
  1. Correct the PD of every credit from the sample rate of 30% to the population rate of 5%.
  2. Compute the expected loss of the portfolio, in DM and as a share of EAD, with and without the correction.
  3. Compute the IRB capital K and the risk-weighted assets of the portfolio with the corrected PDs.
  4. Interpretation: why does the correction change the expected loss but not the AUC?
- **Report**: eight numbers and two sentences
- Notebook: section B5

Back Exercise list

B5: Solution [Solved]

- Shift of the log-odds: $\ln(0.05/0.95) - \ln(0.30/0.70) = -2.097$
  - mean PD: 30.7% in the sample scale, 7.3% after the correction (median 4.2%, maximum 54.4%)
- EAD 957 819 DM; EL 41 744 DM = 4.4% of EAD (without correction: 16.3%)
  - IRB capital K: 54 542 DM = 5.7% of EAD; RWA 681 771 DM, a risk weight of 71.2%
- The mean corrected PD is above 5% because the correction is applied to the log-odds, not to the PD
- Interpretation: the correction adds the same constant to every log-odds: the order of the applicants, hence AUC, is unchanged, while every PD shrinks and EL falls from 16.3% to 4.4% of EAD

SFM_ch12_seminar

B6: Fairness on the Taiwan Data [Proposed]

- **Question**: does the logit of B4 treat men and women, and young and older clients, alike? Model: B5 and the prerequisite Fairness of a Credit Decision (16/17).
- **Inputs**: Taiwan test sample (9000 clients); sex and age are not model inputs; approve the 75% with the lowest PD
- **Tasks**
  1. Compute the approval rate and the default rate of men and women, and of clients up to 25 years and older.
  2. Compute the approval rate of the good payers (equal opportunity) in each group.
  3. Compute the default rate among the approved clients of each group.
  4. Interpretation: which fairness criterion does the model come closest to satisfying?
- **Report**: one table and two sentences
- Notebook: section B6

Back Exercise list

C1: Gradient Boosting Compared with the Logit [Proposed]

- **Question**: benchmarks report gains for flexible models Lessmann et al. (2015); how large is the gain on the Taiwan data, and is it worth the loss of a simple scorecard?
- **Inputs**: the inputs and the split of B4; models: B3, B4
- **Tasks**
  1. Fit gradient boosting (scikit-learn, default settings) on the training sample.
  2. Compare its test AUC with the logit by the DeLong test, and compare the Brier scores.
  3. Compare the training and the test AUC of gradient boosting.
  4. Interpretation: is the gain large enough to replace the scorecard?
- **Report**: a table and a plan for a project
- Notebook: section C1

C2: Audit an AI Answer [Proposed]

- A student asked an AI assistant to summarise the B1 scorecard. The answer:
- (a) The AUC is 0.80, so the scorecard classifies 0.80 of the applicants correctly.
- (b) The Gini coefficient is AUC - 0.5 = 0.30.
- (c) An applicant with a PD of 30% from this model has a 30% chance of default at the bank.
- (d) With PDO = 20, a PD that halves from 20% to 10% adds exactly 20 points.
- (e) KS is the largest vertical gap between the score distributions of bads and goods.
- (f) LDA may only be used when every variable is Normally distributed.
- Tasks
  - 1. For each statement, say whether it is correct; if not, give the correct statement and, where possible, the correct number from the notebook (section C2).
  - 2. Report: a list of six verdicts with one line of justification each.

Points to Remember

- A logit models the log-odds; e^{β} is an odds ratio; the effect on PD is $p(1-p)\beta$
- WoE and IV turn a table of goods and bads into evidence and a ranking of variables
- AUC counts correctly ordered pairs; Gini $= 2\,\text{AUC} - 1$; accuracy misleads when classes are imbalanced
- Choose the cut-off from the costs; correct oversampled PDs before any EL or capital
- An AI answer is a draft: check what AUC means, the Gini formula, the prior and the PDO scale

After the Seminar

- Lecture 12 develops each topic of today: the Basel parameters, WoE and IV, logit and LDA, scorecards, validation, reject inference, fairness, Altman and Merton
- C1 can grow into a team project: cross-validated benchmark, costs and fairness by group
- Reading: Franke, Härdle & Hafner (2019), Ch. 21–22; Härdle & Simar (2019), Ch. 14; Siddiqi (2006)

References

- **Basel Committee on Banking Supervision (2006).** International convergence of capital measurement and capital standards: a revised framework (comprehensive version). **Bank for International Settlements**.
- **Brier, G.W. (1950).** Verification of forecasts expressed in terms of probability. ***Monthly Weather Review***, **78(1), 1–3**.
- **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**.
- **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**.
- **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**.
- **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**.
- **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**.