








Statistica piețelor financiare

Seminarul 12: Modele de scoring



Daniel Traian PELE

Academia de Studii Economice din București
IDA Institute Digital Assets
Blockchain Research Center
AI4EFin Artificial Intelligence for Energy Finance
Academia Română, Institutul de Prognoză Economică
MSCA Digital Finance

Acronime folosite în acest material (1/2)

- **AI**: Artificial Intelligence (inteligență artificială)
- **AR**: Accuracy Ratio (of the CAP curve) — raportul de acuratețe (al curbei CAP)
- **AUC**: Area Under the (ROC) Curve (aria de sub curba ROC)
- **CAP**: Cumulative Accuracy Profile (profilul cumulat al acurateței)
- **CI**: Confidence Interval (interval de încredere)
- **DM**: Deutsche Mark (marca germană)
- **EAD**: Exposure At Default (expunerea la momentul nerambursării)
- **FN**: False Negative (fals negativ)
- **FPR**: False Positive Rate (rata fals pozitivelor)
- **HL**: Hosmer–Lemeshow (test) — testul Hosmer–Lemeshow
- **IRB**: Internal Ratings-Based (approach) — abordarea pe baza ratingurilor interne
- **KS**: Kolmogorov–Smirnov (statistic) — statistica Kolmogorov–Smirnov

Acronime folosite în acest material (2/2)

- **LDA**: Linear Discriminant Analysis (analiza discriminantă liniară)
- **LGD**: Loss Given Default (pierderea în caz de nerambursare)
- **PD**: Probability of Default (probabilitatea de nerambursare)
- **PDO**: Points to Double the Odds (punctele care dublează șansa bun:rău)
- **ROC**: Receiver Operating Characteristic (curve) — curba caracteristică de operare
- **RWA**: Risk-Weighted Assets (activele ponderate la risc)
- **TN**: True Negative (adevărat negativ)
- **TP**: True Positive (adevărat pozitiv)
- **TPR**: True Positive Rate — rata adevărat pozitivelor (sensibilitatea)
- **UCI**: University of California, Irvine (Machine Learning Repository) — Universitatea din California, Irvine (arhiva de date pentru învățare automată)

Întrebarea de azi și structura seminarului

- Întrebarea seminarului
  - **cum transformă o bancă datele unui solicitant într-o probabilitate de nerambursare și cum judecăm această probabilitate?**
- Rezultate de obținut
  - răspunsurile la exercițiile A
  - tabelul sau figura cerute la fiecare exercițiu B
  - un răspuns scurt și argumentat la C1; erorile găsite în răspunsul AI din C2 și răspunsul corectat
  - frazele de interpretare cerute la fiecare exercițiu
- Seminarul are rol de exercițiu și nu se notează; rezolvările exercițiilor [Propus] se discută la seminar
- Notebook-ul de azi: deschideți notebook-ul seminarului în Google Colab
  - fiecare cerință indică secțiunea din notebook
- Exercițiile, tipurile și ordinea lor: harta exercițiilor (slide-ul următor)
- Contact: prof. Daniel Traian Pele, danpele@ase.ro

Harta exercițiilor

Cerința	Întrebarea	Tipul	Model
A1, A2	PD a unui solicitant dintr-un model logit; rapoartele șanselor	Rezolvat, Propus	A1
A3, A4	WoE și IV dintr-un tabel de bun-platnici și rău-platnici	Rezolvat, Propus	A3
A5, A6	matricea de confuzie, AUC prin numărarea perechilor, Gini, KS, scorul Brier	Rezolvat, Propus	A5
A7, A8	scalarea scorecard-ului și punctele; corecția a priori și pierderea așteptată	Rezolvat, Propus	A7
B1, B2	IV și un logit WoE: South German Credit; IV pentru datele din Taiwan	Rezolvat, Propus	B1
B3, B4	validarea scorecard-ului; logit față de LDA pe datele din Taiwan	Rezolvat, Propus	B3
B5, B6	pierderea așteptată și capitalul IRB ale unui portofoliu; echitatea pe sexe și vîrste	Rezolvat, Propus	B5
C1, C2	gradient boosting comparat cu logit-ul; erorile unui răspuns AI	Propus	B3, A5

- **[Rezolvat]**: rezolvarea completă în slide-uri și în notebook, un model de urmat; **[Propus]**: îl rezolvați după model; soluția se discută la seminar
- A: calcule pe hîrtie; B: date reale, în notebook, fiecare încheiat cu o întrebare de interpretare; C: întrebări deschise, printre care verificarea unui răspuns AI

Datele folosite

Setul de date	Sursa	Mărimea	Rezultatul
South German Credit	South German Credit (2020)	1000 de credite, 1973–1975	300 rău-platnici, 700 bun-platnici
Carduri de credit, Taiwan	Default of Credit Card Clients (2009)	30 000 de deținători, 2005	nerambursare în octombrie 2005 (22,1%)

- Ambele sub licența CC BY 4.0, salvate o singură dată în data/credit din repository-ul cursului; South German Credit este versiunea corectată a datelor germane de credit din UCI Grömping (2019)
- Nerambursare = 1 pentru un credit neperformant sau o plată ratată, 0 în rest; în South German Credit, creditele neperformante sînt suprareprezentate (rata din populație este de aproximativ 5%)
- În notebook: `load_credit('sgc')`, `woe_table(x, y, var)`, `logit_fit(X, y)`, `auc(score, y)`; nu este nevoie de cont sau de cheie

Noțiuni necesare azi (1/17): nerambursarea și pierderea așteptată

- **Nerambursarea** (default): debitorul nu plătește conform contractului (Basel: peste 90 de zile de întîrziere sau probabilitate mică de plată) Comitetul de la Basel (2006)
  - $Y = 1$ pentru nerambursare (rău-platnic), $Y = 0$ pentru bun-platnic
  - un **scor de credit** ordonează solicitanții după risc; un **model de scoring** transformă caracteristicile lor într-o PD
- Trei parametri de risc ai unui credit
  - **PD** (probability of default): probabilitatea de nerambursare într-un an
  - **LGD** (loss given default): partea din expunere pierdută în caz de nerambursare
  - **EAD** (exposure at default): suma datorată la nerambursare
- **Pierderea așteptată**

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

  - pierderea medie a creditului; este acoperită prin provizioane și inclusă în dobîndă
  - pentru un portofoliu: suma EL a tuturor creditelor; ca pondere din EAD totală

Noțiuni necesare azi (2/17): capitalul IRB din Basel

- Capitalul **IRB** (internal ratings-based, bazat pe ratinguri interne) pentru alte credite de retail, pe unitatea de EAD Comitetul de la Basel (2006)

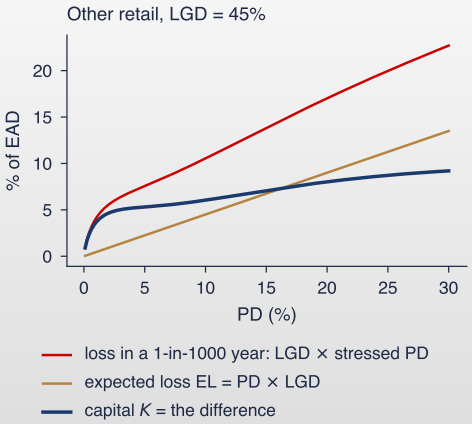
$$K = \mathrm{LGD}\left[\Phi\left(\frac{\Phi^{-1}(\mathrm{PD}) + \sqrt{R}\,\Phi^{-1}(0{,}999)}{\sqrt{1-R}}\right) - \mathrm{PD}\right]$$

  - Φ, Φ^{-1}: funcția de repartiție $N(0,1)$ și inversa ei
  - $\Phi(\cdot)$: PD într-un an nefavorabil care apare o dată la 1000 de ani; minus PD: capitalul acoperă doar pierderea neașteptată

- Corelația cu economia, de la 0,16 (PD mică) la 0,03 (PD mare)

$$R = 0{,}03\,w + 0{,}16\,(1-w), \qquad w = \frac{1 - e^{-35\,\mathrm{PD}}}{1 - e^{-35}}$$

  - **RWA** (risk-weighted assets, activele ponderate la risc) $= 12{,}5\,K\,\mathrm{EAD}$; ponderea de risc = RWA/EAD



SFM_ch12_seminar

Noțiuni necesare azi (3/17): șansa și logaritmul șansei

- Pentru o probabilitate de nerambursare p

$$\text{șansa} = \frac{p}{1-p}, \qquad \text{logaritmul șansei} = \ln \frac{p}{1-p}, \qquad \text{șansa bun:rău} = \frac{1-p}{p}$$

  - șansa: cîte nerambursări la o rambursare; $p = 0{,}2$: șansa 0,25, logaritmul șansei $-1{,}386$, șansa bun:rău 4:1
  - logaritmul șansei ia valori pe toată axa reală, în timp ce p rămîne în $(0, 1)$

- Înapoi, de la șansă la probabilitate

$$p = \frac{\text{șansa}}{1 + \text{șansa}}, \qquad p = \frac{1}{1 + \text{șansa bun:rău}}$$

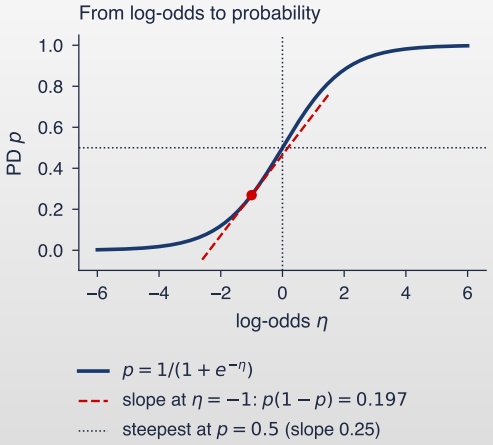
  - dublarea șansei nu înseamnă dublarea lui p: de la șansa 0,25 la 0,5, p crește de la 0,2 la 0,333

Noțiuni necesare azi (4/17): modelul logit

- **Modelul logit** Cox (1958): logaritmul șansei este liniar în caracteristici

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

  - $x_1, \ldots, x_k$: caracteristicile solicitantului; η: predictorul liniar; β_0: termenul liber
- Interpretarea coeficienților
  - β_j: modificarea logaritmului șansei la o unitate de x_j; $\beta_j > 0$: risc mai mare
  - e^{β_j}: **raportul șanselor** (odds ratio), factorul care înmulțește șansa; același pentru orice solicitant
  - efectul asupra probabilității: $\Delta p \approx p(1-p)\,\beta_j$, maxim în jurul lui $p = 0{,}5$



Noțiuni necesare azi (5/17): estimarea modelului logit

- **Verosimilitatea maximă**: fiecare solicitant i contribuie cu logaritmul probabilității a ceea ce s-a întîmplat

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

  - $y_i \in \{0, 1\}$: rezultatul observat; $\hat{\beta}$ maximizează ℓ (numeric); **se**: erorile standard, din curbura lui ℓ
  - $z = \hat{\beta}_j / \mathrm{se}_j$: $|z| > 1{,}96$ înseamnă o variabilă semnificativă la 5%

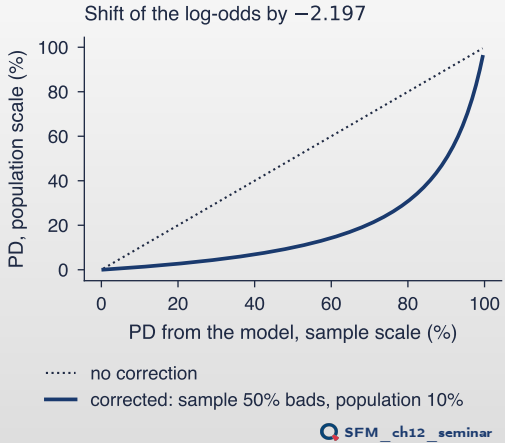
- **Eșantionul de estimare și eșantionul de test**
  - modelul se estimează pe eșantionul de estimare (de exemplu 70%); eșantionul de test (30%) joacă rolul viitorilor solicitanți și se folosește o singură dată, la final
  - împărțire **stratificată**: aceeași rată a rău-platnicilor în ambele eșantioane
  - **supraajustarea** (overfitting): un model mult mai bun pe eșantionul de estimare decît pe cel de test a învățat zgomot

Noțiuni necesare azi (6/17): suprareprezentarea și corecția a priori

- **Suprareprezentarea**: eșantionul are rata rău-platnicilor π_s, iar populația o rată mai mică π
  - modelul logit estimat pe eșantion dă PD la scara eșantionului, prea mari pentru bancă
- **Corecția a priori**: adunăm aceeași constantă la fiecare logaritm al șansei

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

  - apoi $p_{\text{pop}} = 1/(1+e^{-\eta_{\text{pop}}})$
  - ordinea solicitanților nu se schimbă, nivelul PD se schimbă
  - graficul folosește rate ilustrative (50% și 10%)



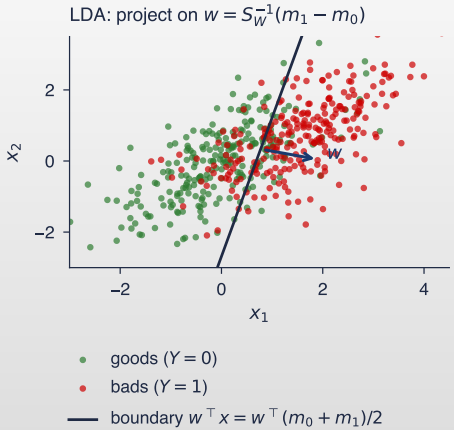
Noțiuni necesare azi (7/17): analiza discriminantă liniară

- **LDA** (linear discriminant analysis, analiza discriminantă liniară) Fisher (1936): direcția care separă cel mai bine cele două grupuri

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

  - m_1, m_0: vectorii medii ai rău-platnicilor și bun-platnicilor; S_1, S_0: matricele lor de covarianță; n_1, n_0: mărimile lor
  - S_W: covarianța comună (pooled); scorul este $w^\top x$

- De la scor la PD
  - direcția lui Fisher nu cere nicio distribuție; dacă ambele grupuri sînt Normale, cu aceeași covarianță, logaritmul șansei este liniar în x, ca în logit
  - departe de normalitate, LDA poate ordona bine, dar poate da PD la un nivel greșit



Noțiuni necesare azi (8/17): weight of evidence și information value

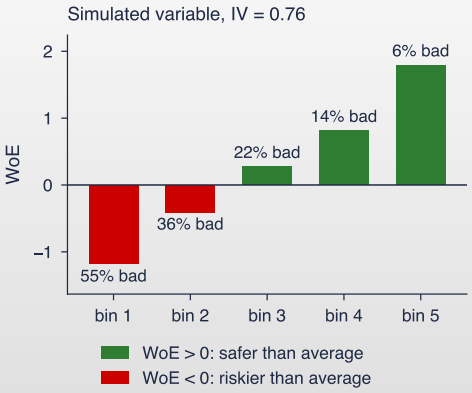
- Împărțim o variabilă în intervale (clase); g_i, b_i: bun-platnicii și rău-platnicii din intervalul i; G, B: totalurile
  - **Weight of Evidence**:

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

  - g_i/G: ponderea tuturor bun-platnicilor care se află în intervalul i; WoE pozitiv: mai sigur decît media, negativ: mai riscant
  - un interval fără bun-platnici sau fără rău-platnici: adăugăm 0,5 la ambele frecvențe
- **Information Value**: puterea variabilei

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

  - regula practică Siddiqi (2006): sub 0,02 inutilă, 0,02–0,1 slabă, 0,1–0,3 medie, peste 0,3 puternică
  - un **logit WoE**: fiecare variabilă este înlocuită cu WoE al intervalului ei, calculat doar pe eșantionul de estimare



SFM_ch12_seminar

Noțiuni necesare azi (9/17): praguri, matricea de confuzie și costuri

- Un **prag** c: respingem solicitantul dacă PD $> c$
  - **TP** (true positive): rău-platnic respins; **FN** (false negative): rău-platnic acceptat
  - **FP** (false positive): bun-platnic respins; **TN** (true negative): bun-platnic acceptat
- Ratele **matricei de confuzie**

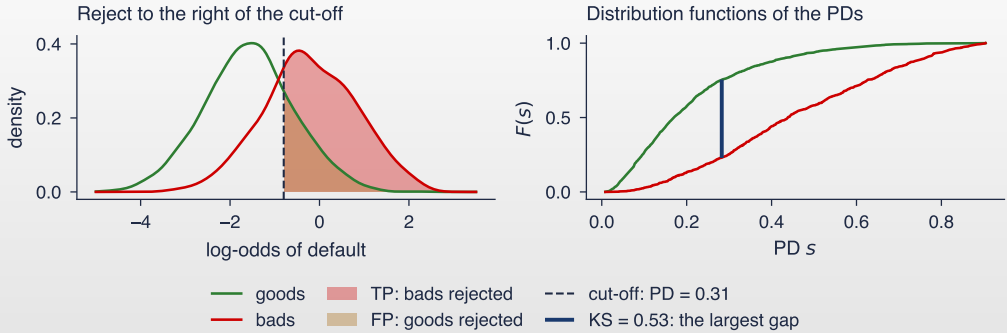
$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \qquad \mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}, \qquad \text{acuratețea} = \frac{\mathrm{TP}+\mathrm{TN}}{n}$$

  - TPR: ponderea rău-platnicilor identificați; FPR: ponderea bun-platnicilor respinși; n: numărul solicitanților
- **Costul pe solicitant**, cu C_{FN}, C_{FP} costurile unei erori de fiecare tip

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

  - **pragul Bayes**: respingem cînd costul așteptat al acceptării, $p\,C_{FN}$, depășește costul respingerii, $(1-p)\,C_{FP}$, adică dacă $p > C_{FP}/(C_{FP}+C_{FN})$
  - exemplu: $C_{FN} = 5$, $C_{FP} = 1$ dau pragul $1/6$

Noțiuni necesare azi (10/17): distribuțiile scorurilor, pragul și KS



- Scoruri simulate pentru bun-platnici și rău-platnici: mutarea pragului spre stînga identifică mai mulți rău-platnici (TP), dar respinge și mai mulți bun-platnici (FP)
- Statistica **KS** (Kolmogorov–Smirnov): cea mai mare distanță pe verticală dintre cele două funcții de repartiție ale PD

$$\mathrm{KS} = \max_s \left| F_{\textbf{rău}}(s) - F_{\textbf{bun}}(s) \right|, \qquad F(s) = \text{ponderea PD} \leq s$$

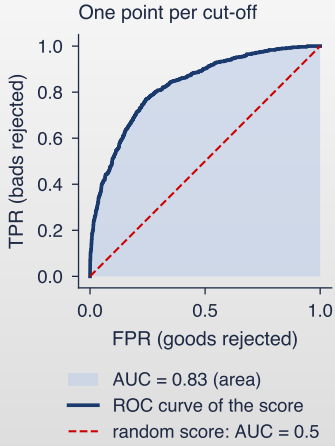
SFM_ch12_seminar

Noțiuni necesare azi (11/17): curba ROC și AUC

- Curba **ROC** (receiver operating characteristic): (FPR, TPR) pentru fiecare prag; **AUC**: aria de sub ea Hanley & McNeil (1982)

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

  - ponderea perechilor (rău, bun) în ordinea corectă; $n_{\mathbf{1}}$, $n_{\mathbf{0}}$: numerele de rău-platnici și bun-platnici; $1\{\cdot\}$: 1 dacă este adevărat, 0 altfel
  - P(PD unui rău-platnic aleator > PD unui bun-platnic aleator): 0,5: nicio discriminare, 1: perfectă



Noțiuni necesare azi (12/17): Gini, CAP și testul DeLong

- **Coeficientul Gini**: AUC rescalat la $[-1, 1]$

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

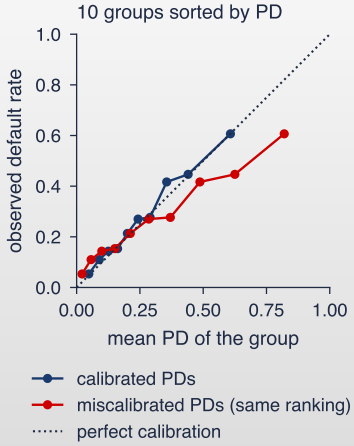
  - 0: un scor aleator; 1: o ordonare perfectă
- Curba **CAP** (cumulative accuracy profile)
  - ordonăm solicitanții de la cel mai riscant; pentru fiecare ponderе x de solicitanți, reprezentăm ponderea tuturor rău-platnicilor aflați printre ei

$$\text{AR} = \frac{\text{aria dintre CAP-ul modelului și diagonală}}{\text{aria dintre CAP-ul perfect și diagonală}}$$

  - **accuracy ratio** AR este egal cu coeficientul Gini (cu excepția egalităților)
- **DeLong** DeLong et al. (1988): incertitudinea unui AUC estimat pe un eșantion de test
  - o eroare standard neparametrică $\text{SE}(\widehat{\text{AUC}})$ calculată din perechile (rău, bun); intervalul de 95%: $\widehat{\text{AUC}} \pm 1{,}96\,\text{SE}$
  - două modele pe aceiași solicitanți: $z = (\widehat{\text{AUC}}_1 - \widehat{\text{AUC}}_2)/\text{SE}(\text{diferenței})$, aproximativ $N(0,1)$; $|z| > 1{,}96$: diferite la 5%

Noțiuni necesare azi (13/17): calibrarea

- **Calibrarea**: PD medie a unui grup trebuie să fie egală cu rata lui observată de nerambursare
  - ordonăm solicitanții după PD, îi împărțim în G grupuri de mărimi egale și comparăm, în fiecare grup, PD medie cu ponderea nerambursărilor
  - puncte pe diagonală: PD calibrate
- Ordonarea și calibrarea sînt proprietăți diferite
  - PD roșii ordonează solicitanții exact ca cele albastre (același AUC), dar nivelul lor este greșit
  - o bancă are nevoie de ambele: ordonarea pentru decizie, nivelul pentru EL și capital



Noțiuni necesare azi (14/17): scorul Brier și testul Hosmer–Lemeshow

- **Scorul Brier** Brier (1950): eroarea pătratică medie a PD; o valoare mai mică este mai bună

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

  - p_i: PD a solicitantului i; $y_i \in \{0, 1\}$: rezultatul; recompensează atît ordonarea, cît și calibrarea
  - reper: o PD constantă egală cu rata rău-platnicilor π dă $\mathrm{BS} = \pi(1 - \pi)$
  - o PD greșită, dar foarte sigură, costă cel mai mult: un bun-platnic cu PD 0,95 adaugă $(0{,}95 - 0)^2 = 0{,}9025$

- Testul **Hosmer–Lemeshow** Hosmer & Lemeshow (1980) cu G grupuri ordonate după 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: nerambursările observate; E_g: nerambursările așteptate (suma PD); $\bar{p}_g$: PD medie a grupului g
  - H_0: PD sînt calibrate; un HL mare (p-value mic): calibrare greșită

Noțiuni necesare azi (15/17): scorecard-urile

- **Scorecard** Siddiqi (2006): un scor liniar în logaritmul șansei bun:rău

$$\text{Scorul} = \text{Offset} + \text{Factor} \times \ln(\text{șansa bun:rău})$$

  - **PDO** (points to double the odds): adăugarea a PDO puncte dublează șansa bun:rău

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

  - Baza: scorul acordat la șansa bun:rău $\text{șansa}_{\mathbf{0}}$; de la scor înapoi la PD: $\ln(\text{șansa}) = (\text{Scorul} - \text{Offset})/\text{Factor}$

- **Punctele unui atribut** (logit WoE cu k variabile)

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

  - i: atributul (intervalul) variabilei j; semnul minus: logit-ul modelează șansa de nerambursare, scorul șansa bun:rău; punctele tuturor variabilelor se adună în scor

Noțiuni necesare azi (16/17): echitatea unei decizii de creditare

- **Echitatea** (fairness) la un prag Hardt, Price & Srebro (2016): comparăm grupuri A (de exemplu bărbați și femei)
  - **paritatea demografică**: rate egale de aprobare, $P(\text{aprobat} \mid A = a)$ aceeași pentru orice grup a
  - **egalitatea de șanse**: rate egale de aprobare a bun-platnicilor, $P(\text{aprobat} \mid Y = 0, A = a)$
  - **calibrarea în fiecare grup**: aceeași PD înseamnă aceeași rată de nerambursare, $P(Y = 1 \mid \mathrm{PD} = p, A = a) = p$
- Observații
  - cînd ratele de nerambursare diferă între grupuri, criteriile nu pot fi îndeplinite toate simultan
  - excluderea sexului și a vîrstei din model nu garantează un tratament egal: alte variabile pot fi corelate cu ele
  - aprobarea celor 75% cu cea mai mică PD înseamnă un prag egal cu cuantila de 75% a PD

Noțiuni necesare azi (17/17): modele flexibile și gradient boosting

- **Gradient boosting**: o sumă de mulți arbori de decizie mici, fiecare estimat pe erorile celor anteriori
  - un **arbore de decizie** împarte solicitanții după reguli simple (de exemplu durata > 24 de luni) și dă fiecărui grup final propria PD
  - surprinde efecte neliniare și interacțiuni fără intervale WoE
  - comparațiile publicate raportează cîștiguri mici, dar reale, față de logit Lessmann et al. (2015)
- Costurile flexibilității
  - supraajustarea: o diferență mare între AUC pe eșantionul de estimare și pe cel de test
  - fără puncte simple pe atribut: mai greu de explicat o respingere solicitantului și autorității de supraveghere
  - **validarea încrucișată** (cross-validation): împărțim eșantionul de estimare în K părți, estimăm pe $K - 1$, evaluăm pe partea rămasă și facem media

A1: PD a unui solicitant [Rezolvat]

Cerință

- Logit pentru nerambursare: termenul liber $-2{,}101$; durata (ani) $0{,}385$; vîrsta (decenii) $-0{,}159$; suma (1000 DM) $0{,}029$; fără cont curent $1{,}859$; sold < 0 DM $1{,}338$.
- Solicitantul: 2 ani, 30 de ani, 3000 DM, fără cont curent.
- 1. Calculați logaritmul șansei η și PD.
- 2. Calculați șansa de nerambursare și interpretați $e^{\mathbf{1{,}859}}$.
- 3. Calculați PD cu un an în plus de durată și comparați-o cu $p(1-p)\beta$.
- Raportați: patru valori și o frază.

Rezolvare

- 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. șansa $= e^{\mathbf{0{,}138}} = 1{,}148$; $e^{\mathbf{1{,}859}} = 6{,}42$: fără cont curent, șansa este de 6,42 ori cea a grupului de referință
- 3. $\eta + 0{,}385$: PD $= 62{,}8\%$ (+9,3 puncte procentuale); $p(1-p)\beta = 9{,}6$ puncte procentuale
- Raportul șanselor este constant; modificarea PD depinde de poziția solicitantului.

Înapoi Lista exercițiilor

A2: un al doilea solicitant [Propus]

Rezolvarea se discută la seminar

Cerință

- Același logit ca în A1. Solicitantul: 1 an, 45 de ani, 2000 DM, sold sub 0 DM. Model: A1.
- 1. Calculați η, șansa și PD.
- 2. Calculați PD cu un an în plus de durată.
- 3. Comparați modificarea PD cu cea din A1 și explicați diferența.
- Raportați: patru valori și o frază.

Înapoi Lista exercițiilor

A3: WoE și IV pentru contul curent [Rezolvat]

Cerință

- South German Credit, eșantionul complet: bun-platnici/rău-platnici după contul curent: fără cont 139/135; sub 0 DM 164/105; 0–200 DM 49/14; cel puțin 200 DM sau salariu 348/46.
- 1. Calculați g_i/G și b_i/B pentru fiecare grup.
- 2. Calculați WoE pentru fiecare grup.
- 3. Calculați IV și clasificați variabila.
- Raportați: un tabel și o frază.

Rezolvare

	g	b	g/G	b/B	WoE	IV
fără cont	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 / salariu	348	46	0,497	0,153	1,176	0,404
total	700	300	1	1		0,666

- de exemplu $\ln(0{,}497/0{,}153) = 1{,}176$; IV $= 0{,}666 > 0{,}3$: puternică.
- WoE scade monoton de la grupul cel mai sigur la cel mai riscant.

SFM_ch12_seminar

înapoi Lista exercițiilor

A4: WoE și IV pentru economii [Propus]

Rezolvarea se discută la seminar

Cerință

- Bun-platnici/rău-platnici după economii: fără sau necunoscute 386/217; sub 100 DM 69/34; 100–500 DM 52/11; 500–1000 DM 42/6; cel puțin 1000 DM 151/32. Model: A3.
- 1. Calculați WoE pentru fiecare grup.
- 2. Calculați IV și clasificați variabila.
- 3. Propuneți o unire a grupurilor cu WoE apropiat.
- Raportați: un tabel și două fraze.

Înapoi Lista exercițiilor

A5: matricea de confuzie și AUC prin numărare [Rezolvat]

Cerință

- Zece solicitanți, PD și rezultatul ($Y = 1$: nerambursare): 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. Construiți matricea de confuzie la pragul 0,5 și calculați acuratețea, TPR și FPR.
- 2. Calculați AUC numărînd perechile (rău, bun) în ordinea corectă, precum și Gini.
- 3. Calculați statistica KS și scorul Brier.
- Raportați: șapte valori și o frază.

Rezolvare

- 1. TP 2, FN 1, FP 1, TN 6: acuratețe 80,0%, TPR 66,7%, FPR 14,3%
- 2. bun-platnici sub fiecare rău-platnic: $4 + 6 + 7 = 17$ din 21 de perechi: AUC $= 0{,}810$, Gini $= 0{,}619$
- 3. la PD $\leq 0{,}20$: 4/7 din bun-platnici, 0/3 din rău-platnici: KS $= 0{,}571$; Brier $= 1{,}4819/10 = 0{,}148$
- Rău-platnicul cu PD 0,25 este ordonat sub trei bun-platnici, iar cel cu 0,55 sub unul: 4 din 21 de perechi sînt în ordinea greșită.

Înapoi Lista exercițiilor

A6: un al doilea eșantion [Propus]

Rezolvarea se discută la seminar

Cerință

- Opt solicitanți: 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. Construiți matricea de confuzie la pragul 0,4 și calculați acuratețea, TPR și FPR.
- 2. Calculați AUC prin numărarea perechilor, precum și Gini.
- 3. Calculați statistica KS și scorul Brier.
- Raportați: șapte valori și o frază.

Înapoi Lista exercițiilor

A7: scalarea unui scorecard [Rezolvat]

Cerință

- PDO $= 20$; 600 de puncte la șansa bun:rău 50:1. Un logit WoE cu $k = 6$ variabile are termenul liber $\beta_0 = -0{,}850$ și, pentru contul curent, $\beta = -0{,}829$; WoE: fără cont $-0{,}826$, cel puțin 200 DM 1,043.
- 1. Calculați Factor și Offset.
- 2. Calculați scorul unei PD de 5% și PD a unui scor de 560.
- 3. Calculați punctele celor două grupuri de cont curent.
- Raportați: șase valori și o frază.

Rezolvare

- 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 șansa $= 2{,}526$, șansa 12,50:1, PD $= 7{,}4\%$
- 3. Offset$/k = 81{,}19$, $\beta_0/k = -0{,}1417$: fără cont $-(-0{,}829\times(-0{,}826)+(-0{,}1417))\times 28{,}85+81{,}19 = 65{,}5$; cel puțin 200 DM: 110,2
- Un cont bine alimentat valorează cu 44,7 puncte mai mult decît lipsa contului: 2,2 dublări ale șansei bun:rău.

înapoi Lista exercițiilor

A8: altă scală, corecția a priori și pierderea așteptată [Propus]

Rezolvarea se discută la seminar

Cerință

- O bancă folosește PDO $= 40$ și 500 de puncte la șansa 20:1. Un model estimat pe un eșantion cu 30% rău-platnici dă unui solicitant PD $= 30\%$; rata din populație este 5%; EAD $= 10\,000$ DM, LGD $= 45\%$. Model: A7.
- 1. Calculați Factor, Offset, scorul unei PD de 10% și PD a unui scor de 520.
- 2. Corectați PD de 30% pentru populație.
- 3. Calculați pierderea așteptată cu ambele PD.
- Raportați: șapte valori și o frază.

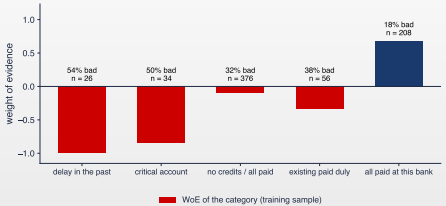
înapoi Lista exercițiilor

B1: un scorecard WoE pentru South German Credit [Rezolvat]

- **Întrebarea**: ce variabile aduc informație despre nerambursare și cît de bine ordonează un logit WoE solicitanții noi?
- **Date**: South German Credit; o împărțire stratificată: 700 de credite pentru estimare, 300 pentru test (30% rău-platnici în ambele)
- **Cerințe**
  1. Calculați tabelele WoE și IV pentru cele 20 de variabile pe eșantionul de estimare (durata, suma și vîrsta în cinci intervale fixe).
  2. Păstrați variabilele cu IV de cel puțin 0,1 și estimați un logit pentru nerambursare pe valorile lor WoE.
  3. Calculați AUC pe eșantionul de estimare și pe cel de test, cu un interval DeLong de 95% pentru AUC de test.
  4. Reprezentați grafic WoE pentru grupurile de istoric de credit.
  5. Interpretare: are sens economic ordinea grupurilor de istoric de credit?
- **Raportați**: lista variabilelor selectate cu IV-ul lor, coeficienții, două valori AUC cu intervalul, graficul și două fraze
- Notebook: secțiunea B1

◂ Înapoi ▸ Lista exercițiilor

B1: rezolvare [Rezolvat]



	termen liber	contul curent	istoricul de credit	destinația creditului	economiile	durata	vechimea la angajator
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 variabile cu IV $\geq 0{,}1$; AUC: estimare 0,789, test 0,801, CI 95% $[0{,}749; 0{,}854]$
- Interpretare: da: o întîrziere în trecut (54% neperformante) și un cont critic (50%) sînt cele mai riscante, toate creditele la această bancă rambursate la timp (18%) cele mai sigure; în etichetele UCI necorectate, acest grup, cel mai sigur, se numea „cont critic"

SFM_ch12_seminar

B2: Information Value pentru datele din Taiwan [Propus]

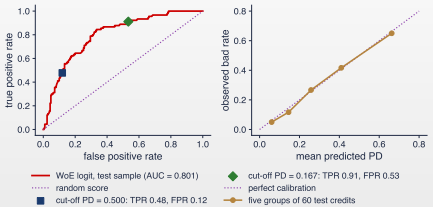
- **Întrebarea**: ce fel de informație prezice nerambursarea unui deținător de card: comportamentul de plată sau datele personale? Model: B1.
- **Date**: cardurile de credit din Taiwan, 30 000 de clienți; aceeași împărțire stratificată 70/30
- **Cerințe**
  1. Calculați IV pentru `PAY_0` și `PAY_2` (starea plăților în septembrie și august; categoriile de la -2 la 3 sau mai mult), pentru `LIMIT_BAL`, `PAY_AMT1`, `BILL_AMT1` și `AGE` (quintile) și pentru `SEX`, `EDUCATION` și `MARRIAGE` (categorii) pe eșantionul de estimare.
  2. Calculați rata de nerambursare pentru fiecare categorie `PAY_0`.
  3. Ordonați variabilele după IV și clasificați-le cu regula practică.
  4. Interpretare: de ce este starea plății din ultima lună mult mai puternică decît datele personale?
- **Raportați**: un tabel cu nouă valori IV, șase rate de nerambursare și două fraze
- Notebook: secțiunea B2

B3: validarea scorecard-ului [Rezolvat]

- **Întrebarea**: cît de bun este scorecard-ul din B1 pe eșantionul de test și ce prag ar trebui să folosească banca?
- **Date**: cele 300 de credite de test din B1 cu PD-urile lor; costuri: 5 pentru un rău-platnic acceptat, 1 pentru un bun-platnic respins
- **Cerințe**
  1. Construiți matricea de confuzie la pragurile 0,5 și 1/6, cu acuratețea, TPR, FPR și costul pe solicitant.
  2. Calculați AUC, Gini, raportul de acuratețe al curbei CAP și statistica KS.
  3. Calculați scorul Brier, comparați-l cu o PD constantă de 30% și aplicați testul Hosmer–Lemeshow cu cinci grupuri.
  4. Reprezentați curba ROC cu cele două praguri și graficul de calibrare.
  5. Interpretare: ce prag i-ați recomanda băncii?
- **Raportați**: un tabel cu două matrice de confuzie, șase statistici, graficul și două fraze
- Notebook: secțiunea B3

◂ Înapoi ▸ Lista exercițiilor

B3: rezolvare [Rezolvat]



prag	TP	FN	FP	TN	acuratețe	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 față de 0,210; HL cu cinci grupuri = 0,589 (p = 0,899): nicio dovadă de calibrare greșită
- Interpretare: pragul de 1/6: respinge mai mulți solicitanți (64,7%), dar identifică 91,1% din rău-platnici, iar costul scade de la 0,867 la 0,507; acuratețea este un criteriu greșit cînd erorile au costuri diferite

SFM_ch12_seminar

B4: logit sau LDA pe datele din Taiwan? [Propus]

- **Întrebarea**: diferă logit-ul și LDA a lui Fisher prin ordonare sau prin calibrare? Model: B3.
- **Date**: Taiwan, împărțire 70/30; 13 variabile standardizate: logaritmul limitei, vîrsta, indicatorii stării plății din septembrie, numărul lunilor cu întîrziere, gradul de utilizare, logaritmul plăților
- **Cerințe**
  1. Estimați un logit și o LDA pe eșantionul de estimare cu aceleași variabile.
  2. Calculați AUC de test pentru ambele și testul DeLong de egalitate a AUC.
  3. Calculați scorul Brier, statistica Hosmer–Lemeshow (zece grupuri) și PD medie pentru ambele, față de rata de nerambursare.
  4. Calculați acuratețea regulii „nimeni nu intră în nerambursare" și a celor două modele la pragul 0,5.
  5. Interpretare: este LDA mai slabă la ordonare sau la calibrare?
- **Raportați**: un tabel și două fraze
- Notebook: secțiunea B4

◂ Înapoi ▸ Lista exercițiilor

B5: cerințele de capital calculate din PD [Rezolvat]

- **Întrebarea**: cîtă pierdere așteptată și cît capital Basel are portofoliul de test din B1 și cît contează corecția a priori?
- **Date**: cele 300 de credite de test din B1; EAD = suma creditului; LGD = 45%; rata din populație 5%; alte expuneri de retail
- **Cerințe**
  1. Corectați PD a fiecărui credit de la rata de 30% din eșantion la rata de 5% din populație.
  2. Calculați pierderea așteptată a portofoliului, în DM și ca pondere din EAD, cu și fără corecție.
  3. Calculați capitalul IRB K și activele ponderate la risc ale portofoliului cu PD corectate.
  4. Interpretare: de ce schimbă corecția pierderea așteptată, dar nu și AUC?
- **Raportați**: opt valori și două fraze
- Notebook: secțiunea B5

◂ Înapoi ▸ Lista exercițiilor

B5: rezolvare [Rezolvat]

- Corecția logaritmului șansei: $\ln(0{,}05/0{,}95) - \ln(0{,}30/0{,}70) = -2{,}097$
  - PD medie: 30,7% pe scala eșantionului, 7,3% după corecție (mediana 4,2%, maximul 54,4%)
- EAD 957 819 DM; EL 41 744 DM = 4,4% din EAD (fără corecție: 16,3%)
  - capitalul IRB K: 54 542 DM = 5,7% din EAD; RWA 681 771 DM, o pondere de risc de 71,2%
- PD medie corectată este peste 5%, deoarece corecția se aplică logaritmului șansei, nu PD
- Interpretare: corecția adaugă aceeași constantă la fiecare logaritm al șansei: ordinea solicitanților, deci AUC, nu se schimbă, în timp ce fiecare PD scade, iar EL coboară de la 16,3% la 4,4% din EAD

SFM_ch12_seminar

B6: echitatea pe datele din Taiwan [Propus]

- **Întrebarea**: tratează logit-ul din B4 în mod egal bărbații și femeile, precum și clienții tineri și pe cei mai în vîrstă? Model: B5 și noțiunea necesară echitatea unei decizii de creditare (16/17).
- **Date**: eșantionul de test din Taiwan (9000 de clienți); sexul și vîrsta nu sînt variabile ale modelului; aprobăm cei 75% cu PD cea mai mică
- **Cerințe**
  1. Calculați rata de aprobare și rata de nerambursare pentru bărbați și femei, precum și pentru clienții de pînă la 25 de ani și cei mai în vîrstă.
  2. Calculați rata de aprobare a bun-platnicilor (egalitatea de șanse) în fiecare grup.
  3. Calculați rata de nerambursare printre clienții aprobați din fiecare grup.
  4. Interpretare: de care criteriu de echitate se apropie cel mai mult modelul?
- **Raportați**: un tabel și două fraze
- Notebook: secțiunea B6

◂ Înapoi ▸ Lista exercițiilor

C1: gradient boosting comparat cu logit-ul [Propus]

- **Întrebarea**: comparațiile raportează cîștiguri pentru modelele flexibile Lessmann et al. (2015); cît de mare este cîștigul pe datele din Taiwan și merită pierderea unui scorecard simplu?
- **Date**: variabilele și împărțirea din B4; modele: B3, B4
- **Cerințe**
  1. Estimați gradient boosting (scikit-learn, setările implicite) pe eșantionul de estimare.
  2. Comparați AUC de test cu cel al logit-ului prin testul DeLong și comparați scorurile Brier.
  3. Comparați AUC de estimare și de test pentru gradient boosting.
  4. Interpretare: este cîștigul suficient de mare pentru a înlocui scorecard-ul?
- **Raportați**: un tabel și un plan de proiect
- Notebook: secțiunea C1

înapoi Lista exercițiilor

C2: verificarea unui răspuns generat de AI [Propus]

- **Un student a cerut unui asistent AI să rezume scorecard-ul din B1. Răspunsul:**
- `(a) AUC este 0,80, deci scorecard-ul clasifică corect 0,80 din solicitanți.`
- `(b) Coeficientul Gini este AUC - 0,5 = 0,30.`
- `(c) Un solicitant cu PD de 30% din acest model are o probabilitate de 30% de nerambursare la bancă.`
- `(d) Cu PDO = 20, o PD care scade la jumătate, de la 20% la 10%, adaugă exact 20 de puncte.`
- `(e) KS este cea mai mare distanță verticală dintre distribuțiile scorurilor rău-platnicilor și bun-platnicilor.`
- `(f) LDA poate fi folosită doar cînd fiecare variabilă are distribuţia Normală.`
- **Cerințe**
  - **1. Pentru fiecare afirmație, precizați dacă este corectă; dacă nu este, formulați afirmația corectă și, acolo unde se poate, dați valoarea corectă din notebook (secțiunea C2).**
  - **2. Raportați: o listă de șase verdicte, fiecare cu un rînd de justificare.**

Înapoi Lista exercițiilor

Idei de reținut

- Un logit modelează logaritmul șansei; e^{β} este un raport al șanselor; efectul asupra PD este $p(1-p)\beta$
- WoE și IV transformă un tabel de bun-platnici și rău-platnici într-o dovadă și o ordonare a variabilelor
- AUC numără perechile ordonate corect; Gini $= 2\,\text{AUC} - 1$; acuratețea înșală cînd clasele sînt dezechilibrate
- Alegem pragul din costuri; corectăm PD din eșantioane suprareprezentate înaintea oricărui calcul de EL sau capital
- Un răspuns AI este doar o primă versiune: verificați ce înseamnă AUC, formula Gini, corecția a priori și scala PDO

După seminar

- Cursul 12 dezvoltă fiecare temă de azi: parametrii Basel, WoE și IV, logit și LDA, scorecard-urile, validarea, reject inference, echitatea, Altman și Merton
- C1 poate deveni un proiect de echipă: comparație prin validare încrucișată, costuri și echitate pe grupuri
- Lectură: Franke, Härdle & Hafner (2019), cap. 21–22; Härdle & Simar (2019), cap. 14; Siddiqi (2006)

Bibliografie

- **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.**