Oravecz Beatrix

Szelekciós torzítás és csökkentése az adósminősítési modelleknél

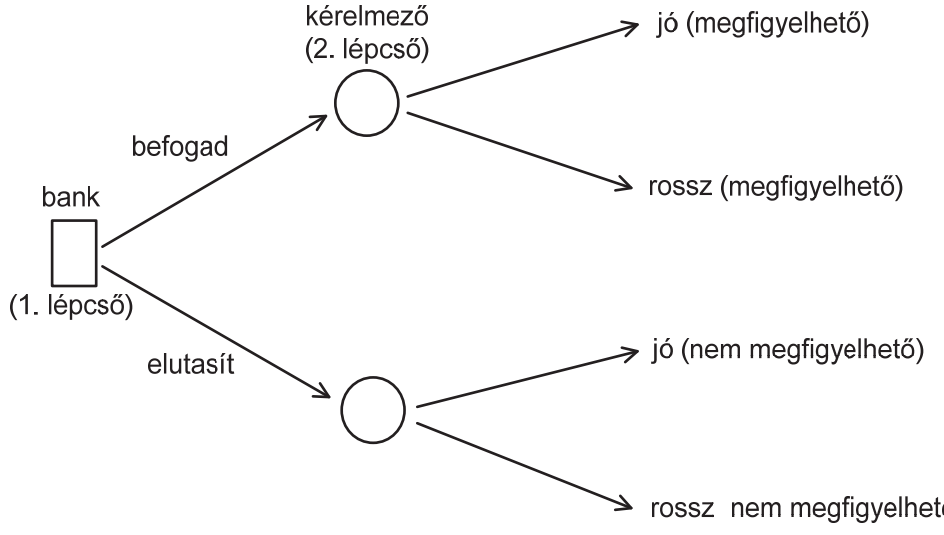
Az adósminősítéshez használt credit scoring módszerek széles körű alkalmazásának ellenére, még mindig vannak a módszertannak olyan aspektusai, amelyek nem kaptak elegendő figyelmet sem a szakirodalomban, sem a gyakorlatban. A modellépítési minta reprezentativitásának kérdése például ilyen terület. Az adósminősítési modellek általában nem reprezentatív mintán épülnek, hiszen itt tipikusan csak azoknál az ügyfeleknél rendelkezünk teljes adatállománnyal, akik átestek egy hitelbírálati folyamaton, és elfogadták őket. Ha viszont a modellépítéshez csak a befogadott ügyfelek adatait használják, akkor megkérdőjelezhető lesz a modell érvényessége, hiszen a befogadottak és az elutasítottak eloszlása valószínűleg különbözik a szisztematikus elbírálási folyamat eredményeként, így a befogadottak nem reprezentálják a teljes sokaságot jelentő összes kérelmezőt. Ezt a jelenséget nevezzük *elutasítási torzításnak* (reject bias), vagy általánosabban *szelekciós torzításnak*. Ez a tanulmány *a credit scoring modelleknél fellépő szelekciós torzítás csökkentésére* használható módszereket tekinti át.

A credit scoring eljárás folyamán azt becsüljük, hogy egy adott hitelkérelmező milyen valószínűséggel lesz „rossz adós". A kérelmek elfogadására/elutasítására használt credit scoring modell idővel elveszti aktualitását, pontosságát, ezért újra kell építeni. Ha nem frissítik a modellt, akkor nem követi a populációban és a magyarázó változók hatásában bekövetkező változásokat, és az eredeti scorecard prediktív ereje csökken. A jó becsléshez olyan modellre van szükség, amely minden hitelkérelmező viselkedését reprezentálja. A scoringmodellek fejlesztésének egyik fő problémája éppen az, hogy csak azon ügyfelek teljesítéséről van múltbéli tapasztalati adatunk, akik korábban már kaptak hitelt a banknál. Azon kérelmezőknek, akiket elutasítottak, bizonyos tulajdonságaikat ismerjük, de nincs információnk arról, hogy jó vagy rossz adósok lettek volna. Ha ezeket az ügyfeleket nem veszik figyelembe az új modellépítés során, akkor a minta nem lesz reprezentatív, nem képviseli az „ajtón bejövő", valós sokaságot. Ez minden erre a mintára épülő klasszifikációs eljárás esetén torzítást okoz. Ezt a torzítást nevezik elutasítási torzításnak (reject bias), vagy általánosabban *szelekciós torzításnak*.

A credit scoring esetén fellépő szelekciós torzítás modellezhető egy kétlépcsős folyamatként, amint azt az alábbi ábra mutatja:

1. ábra

A credit scoring esetén fellépő szelekciós torzítás kétlépcsős folyamata



Az első lépcsőben a bank eldönti, hogy meghitelezi-e az ügyfelet, vagy sem. A második lépcsőben megfigyelhető, hogy az ügyfél a jó vagy rossz kockázati csoportba tartozik-e, de csak azoknál az ügyfeleknél, akiket meghiteleztek.[1]

A scoringmodellek építésénél fellépő szelekciós torzítás kiküszöbölésére számos módszert kipróbáltak és ajánlottak az utóbbi időkben, ezen technikák összefoglaló neve a *reject inference* (következtetés az elutasítottak felhasználásával). Ez tulajdonképpen azt jelenti: megpróbáljuk megbecsülni, hogyan viselkedtek volna az elutasítottak, ha meghiteleztük volna őket, és az elutasítottakat is felhasználjuk a modellépítéshez vagy az elfogadottakon épített modell kiigazításához.

Egy gyakran idézett példa a büntetett előélet. A büntetett előéletű kérelmezőket majdnem mindig elutasítják. Ha mindet elutasítanák, akkor reject inference nélkül a végső modellben nem jelenne meg ez az ismérv. Az a tény, hogy a többséget elutasítják, gyakran azt jelenti, hogy a kisebbség, akit elfogadnak, nagyon speciális tulajdonságokkal rendelkezik, és egyáltalán nem reprezentálja a büntetett előéletűeket általában. Így, ha csak az elfogadottak teljesítését modellezzük, akkor a végső modellünk túlzottan optimista lesz.

Hogyan lehet tehát felhasználni az elutasítottakról rendelkezésre álló, részleges információkat a scoringrendszer fejlesztéséhez? A kérdés megválaszolása előtt csoportosítsuk a lehetséges helyzeteket!

1 Itt természetesen application scorecard építésről van szó, azaz olyan ügyfelekről, akik most először jönnek a bankhoz, tehát nincs rájuk vonatkozó múltbéli visszafizetési adat.

1. A szelekciós torzítás mint hiányzóadat-probléma

A reject inference technikákat három csoportba oszthatjuk. Az *első csoportba* tartozik az az ideális helyzet, amikor a minta reprezentálja az egész populációt. A *második csoportban* – bár a minta csak az elfogadottakat tartalmazza – feltételezzük, hogy az elfogadottak eloszlásának jellemzői kiterjeszthetők az elutasítottakra is. Így az elutasítottakról lévő információk is beépíthetők a modellbe az elfogadottak információival készített, ismert függvények segítségével. A *harmadik csoportban* a minta az elfogadottakból származik, ami egy részsokasága a teljes populációnak, és feltételezzük, hogy eloszlása különbözik az elutasítottakétól. Ebben az esetben az elfogadottakból az elutasítottakra vonatkozó, direkt statisztikai következtetés megbízhatatlan.

A probléma felfogható a hiányzó adatok alapján készült statisztikai következtetés egy példájaként is. Ekkor a fenti csoportosítást megfeleltethetjük a *Little* és *Rubin*-féle adathiány-mechanizmusok három típusának[2]: 1. teljesen véletlen adathiány (Missing Completely at Random – MCAR), 2. véletlen adathiány (Missing at Random – MAR) és 3. nem véletlen adathiány (Not Missing at Random – NMAR).

Nézzük meg ezt egy kicsit közelebbről!

Először is különböztessük meg a *szelekciós mechanizmust*, amely meghatározza, hogy egy ügyfelet beenged vagy elutasít-e a hitelező, és az *eredménymechanizmust*, amely meghatározza, hogy az ügyfél jó vagy rossz-e. A szelekció tulajdonképpen maga az adathiány-mechanizmus, hiszen jelzi, hogy az ügyfélnél megfigyelhető-e az eredményváltozó (jó/rossz) értéke. A credit scoringban az elsődleges cél az eredménymechanizmus modellezése. A hitelező szeretne egy javított, frissített befogadási/elutasítási szabályt, amelyet az új ügyfeleknél alkalmazhat. A továbbiakban tegyük fel, hogy $\mathbf{x} = (x_1, \ldots, x_k)$ a magyarázó változók vektora, amelyik minden kérelmező esetén ismert. Ez tartalmazza azokat az információkat, amelyeket a hitelkérelmi nyomtatványban kitöltöttek, és esetleg egyebeket, amelyeket a bank még ismer a kérelmező hiteltörténetéből. Az y értéke csak az elfogadottakra ismert, az elutasítottakra hiányzik. Feltételezzük, hogy $y \in \{0,1\}$, ahol a 0 jelöli a jó hiteleket és az 1 a rosszakat. Továbbá definiáljunk egy a segédváltozót úgy, hogy $a = 1$ jelentse azt, ha befogadnak egy ügyfelet, és $a = 0$, ha elutasítják. Az y értéke tehát ismert, ha $a = 1$, és hiányzik, ha $a = 0$.

A továbbiakban a rövidség kedvéért néhol alkalmazzuk az alábbi jelöléseket:

- A: $a = 1$ (accept – elfogadás),
- R: $a = 0$ (reject – elutasítás),
- B: $y = 1$ (bad – rossz),
- G: $y = 0$ (good – jó).

Ez a betűs jelölés rövidebb, és talán könnyebben megjegyezhető.

Nézzük tehát az adathiány Little és Rubin-féle típusait!

2 A hiányzó adatok kezelésének legalkalmasabb módját akkor tudjuk megtalálni, ha ismerjük, hogy miként váltak hiányzóakká. Little és Rubin [1987] az adathiány három alapvető esetét különbözteti meg, attól függően, hogy milyen a kapcsolat a hiányzás és az adatbázisban lévő változók értékei között. Ezeket ők adathiány-mechanizmusnak nevezték el.

1.1. Teljesen véletlen adathiány (MCAR)

Az *y* értéke teljesen véletlenszerűen hiányzik (MCAR), ha annak a valószínűsége, hogy *y* megfigyelhető – azaz *A* (*a* = 1): a kérelmet elfogadták – nem függ sem az **x**, sem az *y* értékétől. Azaz: $P(A|\mathbf{x}, y) = P(A)$.[3]

Ez azt jelenti, hogy a kérelmek elfogadása és elutasítása véletlenszerű (például pénzfeldobással döntik el, ki kapjon hitelt). (Ez akkor történhet meg, ha így próbálnak meg információt vásárolni az egyébként elutasítandókról.) Akármi is az oka a teljesen véletlen adathiánynak, ebben az esetben semmi probléma nincs, mert az elfogadottakon épített modell megbízható és torzítatlan lesz az egész sokaságra nézve is.

1.2. Véletlen adathiány (MAR)

A visszafizetést leíró változó (*y*) értéke véletlenszerűen hiányzik (MAR), ha az elfogadás valószínűsége függ **x**-től, de feltéve, hogy **x**-et ismerjük, nem függ *y*-tól.

Azaz: $P(A|\mathbf{x}, y) = P(A|\mathbf{x})$.

Ez a helyzet már előfordulhat a gyakorlatban, hiszen egyre több helyen alkalmaznak formális szelekciós (credit scoring) modelleket. Ekkor *y* értékét csak akkor ismerhetjük, ha az **x** magyarázó változók valamilyen *s* függvénye egy küszöbérték alá süllyed, azaz $s(\mathbf{x}) \leq c$, ahol *c* a cut-off érték.[4]

Ekkor a fenti azonosságból következik, hogy

$$P(y = 1|\mathbf{x}, A) = P(y = 1|\mathbf{x}, R) = p(y = 1|\mathbf{x}),$$

azaz **x** minden rögzített értékére a megfigyelt és a hiányzó *y*-ok eloszlása megegyezik. Ez a MAR-feltételből következő, fontos tulajdonság, amelyet az erre az esetre alkalmazható modellek ki is használnak.

Bár az előbb azt mondtuk, hogy ez a feltétel teljesülhet a gyakorlatban, a valóságban azért inkább csak közelítőleg teljesül, mert a formális modelleket esetenként felülbírálhatják a befogadó/elutasító döntés meghozatalakor (override), azaz előfordul „kivételágon" való beengedés vagy ügyintézői elutasítás is.

3 Furcsának tűnhet, hogy feltételként szerepel a hitelkockázatot leíró változó (*y*), aminek értéke (jó vagy rossz) csak később derül ki, vagy az elutasítottaknál ki sem derül. Ez a hitelképesség vagy hitelkockázat azonban már meglévő tulajdonsága, jellemzője az ügyletnek, még ha nem is ismerjük az értékét. Igaz, hogy a hitelkockázat értékét az is befolyásolja, hogy a kérelmező megkapja-e a hitelt, de ettől a hatástól a dolgozatban eltekintünk.

4 Ez az *s* függvény a scoringmodell (bármilyen modell lehet, amit a bank alkalmaz: lineáris regresszió, logisztikus regresszió, klasszifikációs fa, neurális háló…).

1.3. Nem véletlen adathiány (NMAR)

A visszafizetést leíró változó (y) értéke nem véletlenszerűen hiányzik (NMAR), ha az elfogadás valószínűsége **x** mellett y-tól is függ.

Azaz: $P(A|\ \mathbf{x},y) \neq P(A|\ \mathbf{x})$.

Ez tipikusan akkor fordul elő, ha beengedés/elutasítás részben olyan jellemzőkön alapul, amelyeket nem rögzítettek az **x**-ben, mint például az ügyintéző általános benyomása a kérelmezőről. Ez a helyzet akkor is, ha a fent említett módon alapvetően a formális modell alapján döntenek, de előfordul, hogy felülbírálják a modell döntését (override) olyan jellemzők alapján, amelyek nem szerepelnek az **x**-ben. Ha ezek az $\mathbf{x}_{\text{látens}}$ jellemzők is pótlólagos hatással vannak az y-ra, akkor

$P(y = 1|\mathbf{x}, A) \neq P(y = 1|\mathbf{x}, R)$,

azaz minden rögzített **x** esetén az y eloszlása a befogadottak és az elutasítottak esetén eltérő. Ebben az esetben az adathiány-mechanizmust is be kell építeni a modellbe, hogy jó becsléseket kapjunk.

Ahogy elmozdulunk a MCAR-esettől a MAR-on keresztül a NMAR felé, az y megfigyelhető értékeivel rendelkező meghitelezettek csoportja egyre inkább szelektált és nem jellemző csoporttá válik a sokaságon belül, így a mintaszelekció problémája felerősödik.

2. Szelekciós torzítást csökkentő technikák

Nézzük tehát, hogyan lehet felhasználni az elutasítottakról rendelkezésre álló, részleges információkat a scoringrendszer fejlesztéséhez! A lehetséges módszereket a fentieknek megfelelő csoportosításban mutatjuk be. A módszerek besorolása nem teljesen egyértelmű, de mindig fel fogjuk tüntetni, hogy milyen feltételek mellett tartozik az adott módszer az egyik vagy másik csoportba.

2. 1. Módszerek MCAR esetén

Teljesen véletlenszerű adathiány esetén nincs szelekciós torzítás, így nincs szükség az elfogadottakon épített modell kiigazítására sem. A következőkben azt ismertetjük, hogyan érhető el ilyen szelekciós torzítást nem tartalmazó, véletlen (reprezentatív) minta. Ebbe a csoportba ideális, egyszerű, de igen drága megoldások tartoznak.

2.1.1 Nyitott kapu

A torzítás eltüntetésének a legegyszerűbb, leghatékonyabb módja egy olyan minta létrehozása, amely véletlenszerű kiválasztás eredménye (pénzfeldobással, kockadobással döntenek), vagy ha senkit nem utasítanak el. A csomagküldő cégek gyakran alkalmazzák ezt a

megoldást. Egy adott időszakban mindenkit kiszolgálnak, azzal a céllal, hogy az így nyert mintát majd használhassák a következő score card- építésnél.

A pénzügyi szervezetek számára azonban – az információszerzés túl magas költsége miatt – ez nem elfogadható megoldás. Hiszen egy vissza nem fizetett hitel összehasonlíthatatlanul nagyobb kárt okoz egy banknak, mint egy ki nem fizetett könyv vagy CD az internetes csomagküldő cégnek. Ráadásul további hátránya, hogy az esetleges szezonalitás miatt még mindig maradhat torzítás.

2.1.2. Résnyire nyitott kapu

A „nyitott kapu"-módszernek azonban vannak előnyei, és átalakítható úgy, hogy a pótlólagos információval elérhető, növekvő pontosság és haszon túlszárnyalja a költségeit.

Egy lehetséges átalakítás lehet, hogy véletlenszerű időszakokban a nyitott kaput használják, egyébként pedig a scoringfüggvény alapján döntenek.

Tovább finomítható a megoldás például úgy, ha minden – egyébként elutasítandó – ügyfélnek van esélye a mintába kerülésre, de nem egyforma valószínűséggel. Kis valószínűséggel kaphatnak hitelt azok, akiknél nagyobb a várható veszteség, és nagyobb valószínűséggel azok, akiknél ez a várható veszteség[5] kisebb. Így egy rétegzett mintát kapunk egyfajta költségoptimális mintaelosztással. Végül átsúlyozással nyerhetünk egy – a sokaságot valóban reprezentáló – mintát anélkül, hogy vállalnunk kellett volna a mindenki beengedésével járó, hatalmas költségeket.

A módszer hátránya, hogy csak jól összehangolt rendszerek és folyamatok esetén működhet jól, és ugyanaz a kérelem megismételt elbírálás esetén másképp viselkedhet.

(Ez a módszer a harmadik csoportba is tartozhat, ezért ott is felsoroljuk a *pótlólagos információk felhasználása* címszó alatt.)

Ha nem reprezentatív a minta, amelyre a scoringfüggvény épül (MAR és NMAR adathiány-mechanizmus), akkor az alább következő megoldásokat javasolja a szakirodalom.

2.2. Módszerek MAR esetén

Különböző megoldások alkalmazhatók, attól függően, hogy milyen a kapcsolat az elfogadó/elutasító döntéshez használt jellemzők ($\mathbf{x}_{\text{régi}}$) és az új scoringfüggvény építésénél elérhető jellemzők ($\mathbf{x}_{\text{új}}$) között.

Ha az $\mathbf{x}_{\text{régi}}$ részhalmaza az $\mathbf{x}_{\text{új}}$-nak, azaz minden jellemző most is elérhető, amelynek az alapján elfogadtak vagy elutasítottak egy kérelmet, akkor lesznek olyan csoportok, amelyeknél semmit nem tudunk a jó-rossz besorolásról (ott, ahol az $\mathbf{x}_{\text{régi}}$ alapján elutasították az ügyfelet). A jellemzők más értékkombinációi esetén viszont lesz információnk a jó/rossz arányról (az $\mathbf{x}_{\text{régi}}$ alapján az ilyen értékekkel rendelkezőket mind elfogadták).

Bonyolultabb a helyzet, ha az $\mathbf{x}_{\text{régi}}$ nem részhalmaza az $\mathbf{x}_{\text{új}}$-nak, azaz vannak olyan látens változók, amelyeket használtak a befogadó/elutasító döntés meghozatalánál, de nem rögzítették őket, így most nem elérhetők. (Ez már a NMAR-eset.)

5 A várható veszteség egyenesen arányos a bedőlés valószínűségével és a hitelösszeg nagyságával, és fordítottan arányos a fedezet nagyságával.

2.2.1. Augmentáció (vagy átsúlyozás)

Az augmentáció módszerét először *Hsia* [1978] vázolta fel. A módszer tulajdonképpen nem más, mint átsúlyozás; úgy, hogy a megfigyelt elemek reprezentálják a hozzájuk hasonló, nem megfigyelteket is. A hasonlóságot a hasonló score-ok jelentik.

Először építünk egy jó-rossz (good-bad) modellt a beengedett populáción, és becsüljük a $P(y = 1|\boldsymbol{x}, A)$ értékét, azaz annak a valószínűségét, hogy az ügylet rossz lesz, ha befogadták, és a jellemzőinek értéke $\boldsymbol{x}$. Ezután építünk egy beengedés-elutasítás (accept-reject) modellt, hasonló technikát alkalmazva, hogy megkapjuk $P(A|\boldsymbol{x}) = P(A|s(\boldsymbol{x}))= P(A|s)$-et, ahol s a befogadás-elutasítás score. Ha a múltban alkalmazott beengedés-elutasítás modell minden változója ismert, és mindenkit annak az alapján ítéltek meg, akkor a modell tökéletesen becsülhető, egyébként nem. A beengedés-elutasítás modellel becsült score-ok alapján kategóriákat alakítunk ki (osztályközös gyakorisági sort készítünk), ahol minden j katagóriában R_j elutasított és A_j beengedett ügyfél van. Az A_j beengedett ügyfélből G_j jó eset volt és B_j rossz (l. az *1. táblázatot*).

1. táblázat

Átsúlyozás

kategória (j)	jók száma	rosszak száma	beengedettek száma	elutasítottak száma	kategória súlya
1	G_1	B_1	$A_1 = G_1 + B_1$	R_1	$(R_1 + A_1) /A_1$
2	G_2	B_2	$A_2 = G_2 + B_2$	R_2	$(R_2 + A_2) /A_2$
.	.	.	.	.	.
.	.	.	.	.	.
k	G_k	B_k	$A_k = G_k + B_k$	R_k	$(R_k + A_k) /A_k$

Hsia ezután alkalmazza a következő feltételezést:

$P(B|s, R) = P(B|s, A)$,

azaz a csőd valószínűsége adott s befogadás-elutasítás score mellett a befogadott és az elutasított kérelmeknél megegyezik.[6]

Ez egyfajta átsúlyozása a mintabeli eloszlásnak úgy, hogy az s score-ral rendelkezők aránya $p(A, s)$ helyett $p(s)$ legyen.

Hiszen ha $p(G|s,R) = p(G|s,A)$, akkor $G_j/A_j = G_j^r/R_j$,

ahol G_j^r a jók imputált száma a j kategóriába eső elutasítottakra, és G_j^r/R_j pedig azon elutasítottak aránya a j kategóriában az elutasítottakon belül, akik jók lettek volna, ha befogadják őket.

6 Ez azt is jelenti, hogy $P(G| s, R) = P(G| s, A)$, mert $P(G) = 1–P(B)$

Így a j kategóriába eső A_j befogadott ügyfél akkora súlyt kap, hogy reprezentálja az A_j és R_j eseteket is, ez a súly $(R_j + A_j)/A_j$, ami a j kategóriában lévő elfogadási valószínűség reciproka.

Mivel a score-ok monoton kapcsolatban vannak az elfogadási valószínűséggel, akár helyettesíthetjük is a score-okat ezekkel a valószínűségekkel, és a kategóriák (osztályközök) helyett tekinthetünk egyedi értékeket, ahol *n* lehetséges érték van (mivel *n* esetünk van). Így minden sorhoz tartozik egy $P(A_i)$ elfogadási valószínűség (i = 1, 2, …n) és egy $1/P(A_i)$ súly.

Végül megépíthető az új jó-rossz scorecard a teljes mintán, amelyben már a korábban elutasítottak is szerepelnek oly módon, hogy az *s* score-ral rendelkező elutasítottak *P(G|s, A)* valószínűséggel lesznek jók . Ez tulajdonképpen azt jelenti, hogy az elfogadottakat $1/P(A_i)$-vel átsúlyozva építik a modellt.

A *fő probléma* ezzel a megoldással az, hogy azonosnak feltételezi a csőd valószínűségét az elfogadott és az elutasított kérelmek között (azonos elfogadási score mellett). Ez a feltétel viszont csak akkor teljesülhet, ha valóban MAR-mechanizmusról van szó, azaz a korábban a kiválasztáshoz alkalmazott látens (ma nem ismert) változók teljesen irrelevánsak voltak. Márpedig ez nem valószínű, valamilyen klasszifikáló erejük biztosan volt, ha már alkalmazták őket.

Ha nem véletlenszerű adathiánnyal találkozunk, a *p(G|s, R) = p(G|s, A)* feltétel nem teljesül. Ezt a problémát orvosolandó születtek más javaslatok a *p(G|s, R)* becslésére. Feltételezhetjük, hogy $p(G|s, R) \leq p(G|s, A)$, és ezt a kisebb valószínűséget szubjektíven választhatjuk. A csökkentés mértéke függhet bizonyos jellemzőktől (például attól, milyen típusú számláról volt szó, amikor nyitották).

Más megközelítésben lehet $p(G|s, R) = kp(G|s, A)$, ahol $k < 1$, és szintén a változók egy részének felhasználásával becsülhető. (Ezek a megoldások már a NMAR-esethez tartoznak).

Crook és *Banasik* [2002] tanulmányában azt találta, hogy az *augmentáció nem eredményezett jobb klasszifikációt*, mint a súlyozatlan, eredeti modell. Sőt, nagyobb elutasítási arány esetén (ami elvileg nagyobb szelekciós torzítást okoz) még rosszabb volt a teljesítménye. Ez azért lehet így, mert az átsúlyozás nem használja a sokasági jó-rossz arányról esetlegesen meglévő tudást.

A módszerekkel elérhető javulást azonban nehéz tesztelni, mert az augmentációnak minden formája szigorú feltételezésekre épül, amelyek a *p(G|s, R)* és *p(G|s, A)* eloszlására és a közöttük lévő kapcsolatra vonatkoznak. A gyakorlatban ezek a feltételek nem mindig teljesülnek, és nem is tesztelhetők.

2.2.2. Extrapoláció

Az *extrapolációnak* számos formája létezik, de alapvetően azt jelenti, hogy egy modellt a bedőlési valószínűségre illesztünk az olyan kombinációk esetére, amelyek mellett korábban befogadtak egy kérelmet (becslünk egy posterior valószínűséget), aztán ezt a modellt kiterjesztjük a korábban elutasítottakra is, majd egy cut-off érték felhasználásával klasszifikáljuk az elutasítottakat is a jó vagy a rossz csoportba. Végül egy új jó-rossz modellt építünk, most már a teljes (imputált) adatbázison (l. *Ash* és *Meester* [2002]).

Természetesen az elutasítási tartomány[7] nagysága meghatározza, hogy mennyire jó modellt tudunk illeszteni. A nagy elutasítási tartomány azt jelenti, hogy kevés információra támaszkodunk a modell építésekor. Előfordulhat például, hogy nem tudjuk pontosan specifikálni a függvényformát, és az elfogadottakon jobbnak tűnik egy lineáris függvény, holott a valóságos kapcsolat egy kvadratikus függvénnyel írható le. Ha ekkor extrapolálunk az elutasítási tartományra, nagyon pontatlan becsléseket kaphatunk, hiszen ott jóval nagyobb lehet a linearitástól való eltérés.

A felosztás (parcelling) is az extrapoláció egy formája. Feltételezi, hogy a jó/rossz odds arányosan változik az elfogadási tartomány mentén. A jó/rossz odds „parcellánkénti" változásának ütemét egy szakértői becslés adja meg. Ezek után a rosszakra is megbecsülhető a kimenet, és azokat is a mintához csatolva, felépíthető az új scoringmodell.

2.2.2.1 Az extrapoláció két lehetséges megközelítése

A célunk az eredménymechanizmus megismerése, azaz annak a modellezése, hogyan függ a rossz hitel valószínűsége az **x** jellemzőktől. Formálisan:

$$p(y|x) = f(\mathbf{x}).$$

Az $f(\mathbf{x})$ egy determinisztikus függvény, amely az **x** vektortér minden pontjára megadja

a rossz hitel valószínűségét. A klasszifikációs eljárás célja, hogy adjunk egy $\hat{f}(\mathbf{x})$ becslést az $f(\mathbf{x})$-re. Az ilyen becslés készítésének két alapvető megközelítési módja van: vagy közvetlenül a $p(y|x)$-re egy modell becslése – ez a *közvetlen becslés* (function estimation), vagy a $p(x, y)$ modellezése, majd a feltételes valószínűség definíciójának használata:

$$p(y|x) = \frac{p(x,y)}{\sum_y p(x,y)},$$

ez a *közvetett becslés* (density estimation). Vagy más elnevezéssel: az első megoldás a *diszkriminatív modellezés* (a modell különbséget tesz y=1 és y=0 között); a második pedig a *generatív modellezés*. Nézzük meg röviden mindkét megközelítést, mert reject inference esetén teljesen más következményeik lesznek (l. *Hand* és *Henley* [1993]).

- *Közvetlen becslés* (function estimation)
  Közvetlen becslés esetén csak az y adott **x** melletti feltételes eloszlására készítünk modelleket.
  Bináris klasszifikációs probléma esetén, általánosan:

  $$y \sim B(1, f(\mathbf{x}))$$

  azaz y Bernoulli-eloszlású véletlen változó, ahol a csőd (rossz kategória, y = 1) valószínűsége $f(\mathbf{x})$ és varianciája $\sigma_y^2(\mathbf{x}) = f(\mathbf{x})(1 - f(\mathbf{x}))$.

7 Elutasítási tartományon itt azt a score- (vagy becsült bedőlési valószínűség) tartományt értjük, amely esetén az ügyfeleket elutasítják (azaz nem hitelezik).

A legnépszerűbb technika, ami ezt a megközelítést alkalmazza, a logisztikus regreszszió, ahol

$$f(\mathbf{x}) = \Lambda(\mathbf{x}\alpha) = (1 + e^{-(\mathbf{x}\alpha)})^{-1},$$

ahol Λ(.) a logisztikus eloszlásfüggvény. A cél egy $\hat{f}(\mathbf{x}|T)$ elérése egy T minta felhasználásával. Fontos megjegyezni, hogy az **x** eloszlására vonatkozóan semmiféle feltételezéssel nem éltünk. A MAR feltétel esetén a megfigyelt *y* és a hiányzó *y* eloszlása megegyezik minden rögzített **x**-re. Ekkor a közvetlen becslés megközelítést alkalmazva, a csak az elfogadottakon épített modell is torzítatlan becslést ad $p(y = 1|\mathbf{x})$-re.
Az elutasítottak nem tartalmaznak semmilyen információt $p(y=1|\mathbf{x})$-re vonatkozóan, tehát semmi haszna nem lenne a modellbe foglalásuknak.
A közvetlen becslésen alapuló módszerek *előnye* az egyszerűség: egy standard statisztikai módszert (logisztikus regresszió) alkalmazhatunk, és pusztán az elfogadottakat kell felhasználnunk a modellépítéshez. *Hátrányuk* viszont, hogy nem használnak fel minden elérhető információt: az elutasítottakról meglévő információkat nem lehet beépíteni a modellbe.

- *Közvetett becslés* (density estimation)
Az $f(\mathbf{x})$ becslésének alternatív paradigmája közvetett becslésen alapul. Itt a Bayes-tételt:

$$f(x) = \frac{\pi_1 p_1(x)}{\pi_0 p_0(x) + \pi_1 p_1(x)}$$

alkalmazzák, ahol $p_i(x) = p(x|y = i)$ feltételes valószínűségi függvények, $\pi_i = p(y=i)$ pedig az osztályok feltétel nélküli (prior) valószínűsége. A mintát két részmintára osztjuk: $T = \{T_0, T_1\}$, ahol T_0 tartalmazza a jó hiteleket, T_1 pedig a rosszakat. Mindkét részmintán külön megbecsüljük a $\hat{p}_i(\mathbf{x}|T_i)$ feltételes eloszlást és a $\hat{\pi}_i$ prior valószínűséget. Aztán ezekből a Bayes-tétel alapján kaphatjuk meg az $\hat{f}(\mathbf{x}|T)$ becslést. Ezt a megközelítést alkalmazza a lineáris és kvadratikus diszkriminancia-analízis (*McLachlan* [1992]).

Legyen $T^A = \{T^A_0, T^A_1\}$ az elfogadottakat tartalmazó minta. Mivel a minta szétosztása függ **x**-től, $\hat{p}_i$ torzított lesz, és ha a rossz hitel valószínűsége függ **x**-től (amit nagyon remélünk), akkor a prior valószínűség becslése ($\hat{\pi}_i$) is torzított lesz.

A csak az elfogadottakat tartalmazó mintában a rosszak eloszlása közelebb van a jók eloszlásához, és a rosszak varianciája kisebb, mint a teljes sokaságban. A jók eloszlása nem nagyon változik, mivel elvileg csak kis arányban utasítják el őket. A rossz hitelek valószínűségét a sokaságban alulbecsüljük.

Az ilyen megközelítést alkalmazó becslések esetén tehát torzított eredményeket kapunk, amennyiben csak az elfogadottakon építjük a modellt. Ezért itt valamilyen módon fel kell használni az elutasítottakat is a torzítás eltüntetéséhez. Ennek egyik lehetséges módja a *keverék eloszlások* alkalmazása, amit a következő pontban mutatunk be.

A közvetett becslésen alapuló eljárások *előnye*, hogy az elutasítottakban meglévő információt is tudják hasznosítani, *hátrányuk* viszont, hogy bonyolultabb számítási technikákat igényelnek, és jól kell specifikálni a komponenseloszlásokat.

- *Alkalmazások, eredmények*
  Crook és Banasik [2002] szerint *az extrapoláció nem javít a modelleken*.
  Meester [2000] két extrapolációs technikát vizsgált, és azt találta, hogy a módszerek sikere függ attól, hogy milyen termékről van szó.
  Hand és Henley [1993] rámutatott, hogy az extrapoláció jobban működik olyan módszerek esetén, amelyek direkt módon becslik $P(y|\mathbf{x})$-et (ilyen például a logisztikus regresszió), mint az olyan módszerek esetén, amelyek közvetve a $P(\mathbf{x}|y = 1)$ és a $P(\mathbf{x}|y = 0)$-n keresztül becsülnek (ilyen a diszkriminanciaanalízis). Például, ha egy normális eloszlású sokaságból csak az eloszlás egyik oldaláról veszünk mintát, akkor ez aszimmetrikus eloszláshoz vezet, így az olyan módszerek, amelyek feltételezik a normalitást (pl. a diszkriminanciaanalízis), torzítani fognak. A normalitás feltételezése egyébként sem tartható a credit scoring területén, hiszen nagyon sok diszkrét változót tartalmaznak a modellek.
  Feelders [1999] szimulációval hasonlította össze a közvetlen és a közvetett becslési modellek teljesítményét reject inference alkalmazás esetén, MAR-feltétel mellett. Kisebb mintáknál a közvetett becslési módszer relatív teljesítménye bizonyult jobbnak, különösen az elutasítási tartományban. A minta növekedésével ez a relatív előny eltűnt, mivel a minta növekedésével az előrejelzési hiba torzításkomponense nem csökken, a varianciakomponense viszont igen.
  Ha a befogadottakon épített jó-rossz modell regressziós koefficiensei alkalmazhatók az elutasítottakra is, akkor az eljárás valójában nem eredményez változásokat ezekben az együtthatókban, de a paraméterbecslések standard hibáinak alulbecsléséhez vezet, hiszen úgy tűnik, mintha nagyobb mintán épült volna a modell.

2.2.3. Keverék eloszlások

- *Elméleti háttér*
  A keverék eloszlások olyan eloszlások, amelyek kifejezhetők más eloszlások „súlyozott átlagaként” (McLachlan és *Basford* [1988]).
  Egy véges keverék általános felírása:

$$p(\mathbf{x}) = \Sigma\, \pi_i p_i(\mathbf{x},\theta_i) \qquad i\text{: } 1, \ldots, c,$$

  ahol c a komponensek számát, π_i a keverési súlyokat és θ_i a komponens paramétervektorokat jelöli.
  Itt feltételezzük, hogy a komponenseloszlások száma megegyezik a csoportok számával, és mindegyik egy, a csoportra való feltételes eloszlást jelöl.
  A credit scoring probléma esetén minden megfigyelésről azt feltételezzük, hogy egy kétkomponensű keverékből (jók és rosszak eloszlásának keverékéből) származik:

$$p(x) = \sum_y p(x,y) = p(y=0)\,p(x|y=0) + p(y=1)\,p(x|y=1)\ .$$

Ekkor, ha a *p(y=i)* keverési arányokat átnevezzük π_i-re, és a *p(x|y=i)* feltételes eloszlások jelölik a keverék komponenseit: $p(x|y=i) = p_i(x)$, akkor láthatjuk, hogy a fenti felírás valóban megfelel egy kétkomponensű keveréknek:

$$p(x) = \pi_0 p_0(x,\theta_0) + \pi_1 p_1(x,\theta_1)\ ,$$

ahol a komponenst az elfogadottaknál megfigyelhetjük, de az elutasítottaknál nem.

- *Alkalmazások, feltételek, eredmények*
  A keverék eloszlások megközelítése szerint tehát feltételezhetjük, hogy a sokaság két eloszlás keverékéből származik – a jók és a rosszak eloszlásából –, és ezen eloszlások típusa ismert. Ezt a megközelítést alkalmazta Feelders [1999]. Ha például az **x** jellemzőkkel rendelkezők aránya *p(***x***)*, akkor mondhatjuk, hogy

  $$p(\mathbf{x}) = p(\mathbf{x}|G)p_G + p(\mathbf{x}|B)p_B.$$

  Ekkor a bal oldal becsülhető a mintából. A p_G és p_B bizonyos feltételezett értékei, valamint a *p(***x***|G)* és *p(***x***|B)* paraméterei teljesen specifikálják a jobb oldalt. Ezek után olyan paramétereket kell választani, amelyek minimalizálják a két oldal közötti különbséget.

  A *p(***x***|G)* és a *p(***x***|B)* paramétereinek becsléséhez használhatók az elfogadottak és – az EM (expectation maximization) algoritmus segítségével – az elutasítottak is.

  Szokásos feltételezés, hogy *p(***x***|G)* és *p(***x***|B)* többváltozós normális eloszlásúak. Sajnos, ez a credit scoring területén nem túl realisztikus feltevés, hiszen a modellekben sok bináris vagy kategorikus változó is szerepel.

  Egy köztes megoldás ezen módszer és az augmentáció között, ha feltételezzük, hogy a jó-rossz score-ok és az elfogadás-elutasítás score-ok kétváltozós normális eloszlásúak mind a jók, mind a rosszak esetében. Ekkor először meg kell becsülni ezen eloszlások paramétereit az elfogadottakból, aztán e paraméterek felhasználásával becslést adni az elutasítottak bedőlési valószínűségére. Az elutasítottak becsült bedőlési valószínűségének felhasználásával újra kell becsülni a két eloszlás paramétereit. Ezt az iteratív eljárást addig kell folytatni, amíg a becsült paraméterértékek nem konvergálnak.

2.3. Módszerek NMAR esetén

Feltéve, hogy a credit scoring modellek jól specifikáltak és megfelelő klasszifikációs erővel bírnak, ráadásul alkalmaztak olyan látens (ma nem ismert) változókat, amelyek hatással vannak a nemfizetés valószínűségére, akkor NMAR típusú adathiánnyal van dolgunk. Ekkor az elfogadottak és az elutasítottak eloszlása tehát különböző. A *harmadik csoportba* sorolhatók azok a technikák, amelyek elfogadják és figyelembe veszik ezt a kiinduló pontot.

Látnunk kell, hogy általánosságban nem tudunk semmit a $p(y|x_0,A)$ és a $p(y|x_0,R)$ közötti kapcsolatról, de élhetünk bizonyos *feltételezésekkel*, amelyek – ha helyesek – csökkentik a modellünk torzítását.

2.3.1. Legyen rossz (önkényes besorolás)

Egy nagyon egyszerű megoldás, ha minden elutasítottat rossznak (csődösnek) definiálnak, azután az így „imputált" teljes adatbázison építik fel a klasszifikációs modellt. A megoldás elvi indoka, hogy biztosan volt valamilyen információ, amelynek az alapján korábban elutasították a kérelmezőt. Ez azonban nagyon durva kezelési mód, több hátránnyal. Probléma például, hogy megerősíti a múltbéli rossz előítéleteket. Ha a potenciális ügyfelek egy csoportját a múltban – tévesen – az „elutasítandó" kategóriába sorolták, akkor nincs lehetőségük onnan kikerülni. Ez a megoldás nemcsak statisztikai, hanem etikai szempontból is erőteljesen megkérdőjelezhető.

Kicsit finomítható a megoldás, ha csak a valóban nagyon rossznak tűnő, elutasított eseteket választjuk ki, és azokat elfogadottként kezeljük (de valójában nem hitelezzük meg őket!). Az így elfogadottként kezelt hitelekhez „rossz" besorolást rendelünk, és bevonjuk azokat a modellépítésbe. A legrosszabbnak tűnő eseteket kiválaszthatjuk az eddigi scoringfüggvény alapján (legmagasabb pontszámú egyedek), vagy egyéb negatív információ alapján (KHR-listás[8], vagy végrehajtás indult ellene). Ez utóbbi már pótlólagos külső információk felhasználását is jelenti. Még így is vannak hátrányai a módszernek. Azon túl, hogy a megoldás ad hoc jellegű, azt eredményezi, hogy a $P(y=1|\mathbf{x})=1$ vonatkozik a mintatér egy jelentős részére, amiről tudjuk, hogy nem igaz, és eltorzíthatja a modellt az elfogadott kérelmekre is.

2.3.2. Pótlólagos információk felhasználása

Meg lehet próbálni pótlólagos információt beszerezni az elutasítottak teljesítéséről. Ez történhet *külső* vagy *belső* forrásból. Külső forrás lehet a KHR-lista, a végrehajtási indítványok, hitelinformációs rendszerek, vagy ha például más hitelintézet nyújtott hitelt a kérelmezőnek, akkor tőlük is megpróbálhatjuk megszerezni a visszafizetésre vonatkozó adatokat. Ez ma Magyarországon az éles verseny és a banktitok megsértése miatt nem nagyon működhet, ráadásul nincs is olyan jó hitelinformációs rendszer, amely ezt lehetővé tenné.

Pótlólagos információt belső forrásból úgy kaphatunk, ha mintát veszünk az egyébként elutasítandókból, beengedjük őket, és megfigyeljük a viselkedésüket. Természetesen ennek nagy költsége van, amit figyelembe kell venni a módszer alkalmazásakor. A költségek csökkentésének egy lehetséges módját ismertettük a „résnyire nyitott kapu" címszó alatt.

Hand és Henley [1993] ezen pótlólagos információk használatát tartotta a legcélravezetőbbnek, ezt „kalibráló mintának" nevezte.

Ha tudjuk, hogy a kalibráló minta véletlen kiválasztás eredménye, akkor egyszerűen kombináljuk az elfogadottakkal, és egy olyan statisztikai technikát használunk, amelyik a rossz hitel posterior valószínűségén ($p(B|\mathbf{x})$) alapszik, mint a logisztikus regresszió. Ha nem vagyunk biztosak abban, hogy a kalibráló minta véletlen kiválasztás eredménye, akkor Hand és Henley [1993] három módszert javasolt a bennük lévő információk hasznosítására:

- A módszerhez több kalibráló mintára van szükség. (Mindegyik tartalmaz elfogadottakat és egyébként elutasítandókat is, és a bedőlés valószínűségének eloszlása is ismert mind az elfogadási, mind az elutasítási tartományban.) Ezek a minták származhatnak különböző időszakokból, különböző földrajzi helyekről, különböző hi-

8 Központi Hitelinformációs Rendszer

teltermékek, vagy akár egyetlen nagy minta felosztásából, de ekkor is fontos, hogy mind a minták száma, mind a mintákon belüli elemszám elegendő legyen megbízható modellek alkotásához.

- Így minden egyes mintában mindkét csoportra (elutasítottak-befogadottak) meghatározhatók az eloszlások bizonyos jellemző tulajdonságai (pl. rosszak aránya). Ha azután megvizsgáljuk ezeknek a jellemzőknek az összes mintán felvett értékeit, akkor megbecsülhetjük a két csoport értékei közötti kapcsolatot.
- Így az új mintában, ahol csak az elfogadottakat ismerjük, az elfogadottak jellemző értékének, valamint az elfogadottak és az elutasítottak értékei közötti kapcsolatnak az ismeretében megbecsülhetjük az elutasítottak jellemző értékét.
- A módszerhez csak egy kalibráló mintára van szükség. Első lépésként építünk egy scorecardot (1) az új mintán, ami csak az elfogadottakat tartalmazza. Aztán a kalibráló mintából csak az elfogadottakon építünk egy modellt (2), majd a teljes kalibráló mintán is építünk egy scorecardot (3). Így a két kalibráló modell (3-2) eltérése felhasználható az új scorecard (1) kiigazításához.
- Egy egyszerű példa: tegyük fel, hogy lineáris regresszióval készítünk scorecardot. Legyen az új elfogadottakra készített regresszió (1) paramétervektora: $\boldsymbol{\alpha}$: $[\alpha_1, \ldots, \alpha_n]$, a kalibráló elfogadottakon építetté (2) $\boldsymbol{\beta}$: $[\beta_1, \ldots, \beta_n]$, és a teljes kalibráló mintán építetté (3) $\boldsymbol{\gamma}$: $[\gamma_1, \ldots, \gamma_n]$. Ekkor a β-k és γ-k közötti kapcsolat leírható például egy **M** diagonális mátrixszal, aminek az i-edik diagonális eleme: γ_i/β_i. Ezt a mátrixot használva, az új teljes sokaságra vonatkozó regresszió együttható vektora $\mathbf{M}\boldsymbol{\alpha}$ lesz. Természetesen ez csak egy példa, és más kiigazítás is lehetséges.
- Szintén egy kalibráló mintára van szükség a keverékeloszlások-megközelítésű modell javításához. A kalibráló mintából ismerjük az ügyfelek egy részének valóságos jó-rossz osztályát (nem csak az elfogadottakét). Ezt az információt felhasználhatjuk, amikor a $p(\boldsymbol{x}|G)$ és $p(\boldsymbol{x}|B)$ eloszlások típusát kiválasztjuk.

A pótlólagos információk (credit bureau) beszerzésével imputált adatokon épített modell hatékonyságát vizsgálta Ash és Meester [2002]. A modell minden beengedési ráta esetén jobban becsülte a rosszak arányát, mint az a modell, amelyik csupán a befogadottakon épített.

Nem véletlen adathiány esetén az adathiány-mechanizmus nem mellőzhető. Ebben az esetben legalább egy nagyjából helytálló modellt kell specifikálni az adathiány modellezésére. Azok a reject inference modellek, amelyek nem írják le ezt a hiányzást, továbbra is torzítottak lehetnek.

2.3.3. Heckman kétlépcsős modellje

Heckman kétlépcsős kétváltozós probit modelljét (Heckman [1979]) is javasolták az elutasítottak modellbe építéséhez, mivel ez a modell nem feltételezi, hogy az elfogadási és az elutasítási tartományból származó minták eloszlása megegyezik. Technikailag a befogadási-elutasítási döntés (hiteldöntés) és a jó-rossz besorolás (csődmodell) leírható egy kétlépcsős modellel, részleges megfigyelhetőséggel.

A Heckman-modell alkalmazhatósága nagyon erősen támaszkodik a két egyenlet (hiteldöntés és csődmodell) teljes specifikálására.

Nézzük ezt a modellt!

A credit scoring esetén fellépő szelekciós torzítás modellezhető egy kétlépcsős folyamatként, amint azt már korábban láthattuk.

Az első lépcsőben a bank eldönti, hogy meghitelezi-e az ügyfelet, vagy sem. Egy szelekciós egyenlet specifikálásával írjuk le ezt a döntést. A második lépcsőben megfigyelhető, hogy az ügyfél a jó vagy rossz kockázati csoportba tartozik-e, de csak azoknál az ügyfeleknél, akiket meghiteleztek. Egy csődegyenlet specifikálásával megpróbáljuk leírni, hogyan hatnak a csőd valószínűségére az ügyfél bizonyos jellemzői. Ez az egyenlet, ha jól specifikálták, használható arra, hogy már a hitelezzük/ne hitelezzük döntés fázisában azonosítsa a várhatóan jó ügyfeleket.

Alkalmazzuk a *kétváltozós probit modellt mintaszelekcióval.* A modell feltételezi, hogy létezik egy mögöttes kapcsolat (látens egyenlet):

$$y_i^* = \mathbf{x}_i \beta + v_i,$$

aminek mi csak a bináris kimenetét ismerjük (csődegyenlet), ahol
$y_i = 1$ csőd (rossz hitel) esetén ($y_i^* \geq 0$),
0 nem csőd (jó hitel) esetén ($y_i^* < 0$).

A szelekciós egyenlet:

$$a_i^* = \mathbf{z}_i \alpha + \varepsilon_i$$ [9],

aminek szintén csak a bináris kimenetét ismerjük:
$a_i = 1$ beengedett hitel esetén ($a_i^* \geq 0$),
0 elutasított hitel esetén ($a_i^* < 0$).

A csődegyenlet eredményváltozójának értéke (csőd vagy nem csőd) csak akkor megfigyelhető, ha $a_i = 1$.

Ahol feltesszük, hogy a hibatagok kétváltozós normális eloszlásúak:
$v \sim N(0,1)$,
$\varepsilon \sim N(0,1)$,
$corr(v,\varepsilon) = \rho$.

Az α együtthatók megmutatják, hogy a hitelelbírálók milyen mértékben támaszkodnak a befogadási döntés során az ügyfél megfigyelhető jellemzőire. A ρ korreláció pedig jelzi, hogy mennyire használnak általunk nem megfigyelhető, egyéb szempontokat.

A szelekciós egyenlet elvileg mindig becsülhető külön, hiszen az teljesen megfigyelt, de csak akkor lesz hatékony, ha $\rho = 0$ (*Meng* és *Schmidt* [1985]). Ha $\rho \neq 0$, akkor a standard probit- és logitmodellek direkt alkalmazása a csődegyenletben torzított paraméterbecslésekhez vezet. Meng és Schmidt [1985] megállapította, hogy a részleges megfigyelhetőség költsége a kétváltozós probit modellnél igen magas, ezért, ha lehetséges, érdemes pótlólagos információkat is beszerezni.

9 A szelekciós egyenletben azért jelöltük a magyarázó változókat **z**-vel, mert nem feltétlen egyeznek meg a csődegyenletben szereplő **x**-ekkel.

A credit scoring területén tehát erőteljesen kétséges a $\rho = 0$ feltételezés. Jobb, ha megpróbáljuk kideríteni, milyen döntési szabályokat alkalmaztak a korábbi modellépítés során, és megpróbáljuk megítélni, hogy mekkora hatása lehet a részleges megfigyelhetőségnek. Sajnos, a hatékonyságveszteséget nem lehet számszerűsíteni az adott adathalmazra vonatkozó referencia nélkül. Ezért ajánlott tehát, ha a paraméterek külön becslése helyett alkalmazzuk a kétváltozós probit modellt szelekcióval, hogy lássuk, szignifikáns-e a korreláció.

Ekkor – a modellnek megfelelően – háromféle megfigyelésünk van: elutasított hitelek, befogadott jó hitelek és befogadott rossz hitelek. Ezek valószínűsége:

$a = 0$: $P(a = 0) = 1 - \Phi(\mathbf{z}\alpha)$,
$a = 1, y = 0$: $P(a = 1, y = 0) = \Phi(\mathbf{z}\alpha) - \Phi_2(\mathbf{z}\alpha, \mathbf{x}\beta; \rho)$,
$a = 1, y = 1$: $P(a = 1, y = 1) = \Phi_2(\mathbf{z}\alpha, \mathbf{x}\beta; \rho)$,
ahol $\Phi(.)$ jelöli az egyváltozós standard normális eloszlásfüggvényt és $\Phi_2(., .; \rho)$ pedig a kétváltozós standard normális eloszlásfüggvényt ρ korrelációval.

Az ennek megfelelő loglikelihood függvény:
$\ln L(\alpha,\beta,\rho) = \Sigma(1 - a_i) \ln(1 - \Phi(\mathbf{z}_i\alpha)) +$
$\Sigma\, a_i\, (1 - y_i) \ln(\Phi(\mathbf{z}_i\alpha) - \Phi_2(\mathbf{z}_i\alpha, \mathbf{x}_i\beta; \rho)) +$
$\Sigma\, a_i\, y_i \ln(\Phi_2(\mathbf{z}_i\alpha, \mathbf{x}_i\beta; \rho))$.
Ezt maximálva kapjuk meg a modellek paramétereinek ML-becslését.

Ha az egyenleteket sikerült jól specifikálni (ez fontos feltétel, és nem biztos, hogy teljesül!), valamint a $\rho = 0$, akkor

$$P(y = 1 | \mathbf{x}, a = 1) = P(y = 1 | \mathbf{x}, a = 0),$$

azaz MAR-típusú adathiánnyal van dolgunk. Tehát nincs szelekciós torzítás a modellben a nem megfigyelhető változók miatt.

Másrészt, ha $\rho < 0$, akkor

$$P(y = 1 | \mathbf{x}, a = 1) < P(y = 1 | \mathbf{x}, a = 0),$$

azaz minden rögzített **x** esetén a rossz hitel valószínűsége az elfogadottak esetén kisebb, mint az elutasítottak között. Ezt várjuk, ha a hitelügyintézők a döntési szabályok felülbírálása során tendenciózusan jó irányba döntenek, bár a döntés okát nem ismerjük, mert nincs rögzítve **x**-ben.

Végül, ha $\rho > 0$, akkor

$$P(y = 1 | \mathbf{x}, a = 1) > P(y = 1 | \mathbf{x}, a = 0),$$

vagyis az a furcsa helyzet áll elő, hogy minden rögzített **x** esetén a rossz hitel valószínűsége az elfogadottak esetén nagyobb, mint az elutasítottak között, ami azt jelentheti, hogy a hitelügyintézők a döntési szabályok felülbírálása során általában rossz irányba döntenek.

- *Alkalmazások, eredmények*

  Meglepő módon *Jacobson* és *Roszbach* [1998], *Boyes* et al. [1989] és *Greene* [1992, 1998] szignifikáns pozitív korrelációt talált a két hibatag között. (A talált ρ értékek rendre: +0,9234; +0,353 és +0,1178.[10]) Jacobson és Roszbach [1998] arra a következtetésre jutott, hogy a vizsgálatba bevont bankok nem akarták minimalizálni a bedőlési kockázatot. Ezt nem csak a pozitív korreláció alapján gondolták; véleményüket alátámasztotta az a tény, hogy a kiválasztáshoz használt változók között voltak olyanok is, amelyek nem csökkentették a bedőlési valószínűséget (nem voltak szignifikánsak a csődegyenletben, vagy éppen ellentétes hatást kifejező előjellel szerepeltek).

  Boyes et al. [1989] is hasonló eredményeket kapott. Ő azonban azzal a hipotézissel magyarázta az eredményeket, hogy a bankok kiválogatnak nagyobb kockázatú hiteleket is, mert ezek nagyobb mérete miatt nagyobb megtérülésre számíthatnak. Ha már magyarázni akarjuk ezt az eredményt, nem gondolom, hogy a nagyobb méretre kell gondolnunk; sokkal inkább arról lehet szó, hogy a nagyobb kockázatú hiteleket magasabb kockázati felárral kompenzálták, így valóban jövedelmezőbbek lehetnek. Jacobson és Roszbach eredményei is ellentmondanak Boyes hipotézisének, mert ők azt találták, hogy a hitel mérete nincs hatással a hitel kockázatára.

  Chen és *Astebro* [2001] nem tudta elvetni a $\rho = 0$ hipotézist, ami azt jelezte, hogy csak gyenge szelekciós torzítás lehetett a mintában a nem megfigyelhető változók miatt. Ez szintén egy adatbázis-specifikus eredmény, hiszen ők a kis kezdővállalkozások hitelezésénél fellépő torzítást vizsgálták. Ezeknél a vállalkozásoknál a bankok elsősorban a tulajdonos hitelképessége alapján döntenek a hitelezésről. *Caouette* et al. [1998] szerint a személyi és a vállalati hitelképesség eltérő, és a kettő közötti korreláció igen alacsony, tehát amikor a bankok a tulajdonos hitelképessége alapján döntenek a vállalkozás hitelképességéről, akkor közel járhatnak a véletlen kiválasztáshoz.

  Ezzel a véleménnyel nem értek egyet. Az általam látott adósminősítési modellekben és a vizsgált adatbázisokban szignifikáns kapcsolatot találtam a tulajdonos hitelképessége és a vállalkozás hitelvisszafizetési képessége között.

  Gyakorlati szempontból a leglényegesebb kérdés, hogy a szelekciós mechanizmus modellezése jobb (nagyobb besorolási pontosságú) csődegyenletet eredményez-e. Sajnos, ezt a kérdést valós hiteladatokon nehéz megválaszolni, mivel az elutasítottak tényleges teljesítménye ismeretlen.

  Banasik et al. [2001] úgy találta, hogy a kétváltozós probit módszer csak minimális javulást jelent a csak az elfogadottakon épített modellhez képest. Ash és Meester [2002] is hasonló következtetésekre jutott.

  Tehát a Heckman-eljárás – bár elméletileg jól hangzó technika – nem képes megfelelően kontrolálni a szelekciós torzítást, megbízhatatlan és nagyon érzékeny, ráadásul támaszkodik a normalitásra, ami gyakran nem teljesül.

10 Az alkalmazandó legjobb reject inference módszer esetenként más-más lehet. A különböző vizsgálatokban nagyon eltérő eredmények adódtak, ez is azt jelzi, hogy az adatbázisok erősen eltérő jellegzetességekkel bírnak. Ugyanakkor az instabil eredmények a módszer kritikájaként is felfoghatók, főleg, hogy túl sokszor találkozunk ezzel a + előjellel, ami ellentétes az előzetes várakozásainkkal.

2.3.4. Három csoport

A mintát három csoportra oszthatjuk: jók, rosszak és elutasítottak. A probléma viszont az, hogy a jövőben mi csak két csoportba szeretnénk osztani a kérelmezőket: a jók (akiket beengedünk) és a rosszak (akiket elutasítunk) csoportjába.

Nem világos, hogy mit tehetünk azokkal, akiket elutasítottként klasszifikáltunk. Ha elutasítjuk őket, akkor az eljárás ekvivalens a „minden elutasított legyen rossz" megoldással. *Thomas*, *Edelman* és Crook [2002] szerint az eljárás egyetlen előnye, hogy klasszikus lineáris diszkriminanciaanalízis esetén, ha három csoportba klasszifikálunk, akkor feltételezzük, hogy mindhárom csoportnak közös kovarianciamátrixa van. Így ez egy módja lehet annak, hogy felhasználjuk az elutasítottakat is a kovarianciamátrix becslésének javítására. (Valójában ez az előny is kérdéses credit scoring esetén, mert nem valószínű, hogy tartható ez a közös kovarianciafeltétel, hiszen az elutasítási döntés, amely a csoportokat képzi, korrelál az ügyfelek megfelelő jellemzőivel.)

2.3.5. Bayesi határ és összecsukás (Bound and Collapse)

Legyen továbbra is $y = j$ a hitelkockázat kimenetele (j=0: jó; j=1: rossz), $s=i$ ($i=1, \ldots, r$) pedig a credit score. (Vannak olyan scoringalkalmazások, ahol a credit score folytonos változó egy alsó és egy felső határ között. Ekkor egyszerűen osztályközökre bontjuk az eloszlást és ezeket az osztályközöket jelöljük i-vel $[i=1, \ldots, r]$.)

NMAR esetén a $P(y,s)$ és $P(y)$ valószínűségek becslései és posterior varianciájuk kiszámítható (*Sebastiani* és *Ramoni* [2000]). A $P(y,s)$ együttes valószínűség- és a $P(y)$ peremvalószínűség posterior eloszlását azonban általában igen bonyolult kifejezések adják meg. Az egyik leggyakrabban alkalmazott módszer, a Gibbs-mintavétel az MCMC (Markov–Chain–Monte-Carlo) módszereket hívja segítségül, és a hiányzó értékeket ismeretlenként kezeli, amelyekből empirikus becslések és megbízhatósági intervallumok számíthatók.

Sebastiani és Ramoni [2000] viszont egy másik módszertani keretet javasolt, amely *„határ és összecsukás" (Bound and Collapse)* néven ismert. A módszer lényege, hogy megállapíthatjuk a hiányzó adatok lehetséges becsléseinek *határait* néhány extrém eloszlás által definiált intervallumon belül, függetlenül az adathiány-mechanizmustól. Az adathalmaz hiányzó értékektől mentes része szolgáltatja az intervallum határait.[11] Ha az adathiány-mechanizmusról van elérhető információ (vagy feltételezés), akkor az beépíthető egy nemválaszolási valószínűségi modellbe, és használható arra, hogy egyetlenegy becslést kiválasszunk. A BC-módszer második lépésként tehát *összecsukja* az intervallumot egyetlen értékké. A módszer tehát egy véletlenszerűen imputált adatot tesz a hiányzó adat helyére.

Ezt a bayesi alapú[12] eljárást javasolta Chen és Astebro [2003] a reject inference problémához. Ez a technika egyrészt beépíti az adatforrás hatását azáltal, hogy a függő változó hiányzó értékeit a becsült hiányzási valószínűségen alapulva imputálja, másrészt lehetővé teszi az elutasítási tartományról elérhető, pótlólagos külső információk felhasználását is a modell kiigazításához.

11 Például ha van 20 elutasított és 100 meghitelezett ügyfél, a meghitelezetteken belül 10 rossz és 90 jó, akkor a rosszak arányára előzetesen felállítható határok: 10/120 és 30/120. (Az extrém eloszlások: az elutasítottakon belül mindenki jó vagy mindenki rossz.)

12 A bayesi gondolat: az adathiány-mechanizmusról meglévő információk vagy feltételezések beépítése a modellbe.

A módszer egy egyszerű alkalmazásánál a hitelek beengedése csak az eredeti scoring modellen alapul (*s* függvény *h* cut-off értékkel), nincs más beengedési szabály.

- *A modell alkalmazásához kapcsolódó kérdések*
  A módszer alkalmazásához látnunk kell, hogyan lehet becsülni az adathiány-mechanizmust.

  Chen és Astebro [2003] az adathiány-mechanizmus becslésére[13] a bedőlés valószínűségét (annak a valószínűségét, hogy az adott eset rossz, csődös lesz) javasolja. Annak a valószínűsége, hogy az adott ügylet rossz lesz, egyenlő annak a valószínűségével, hogy az adott ügyletet (kérelmet) elutasítják. Így a score felfogható a hiányzás valószínűségének mértékeként.

  Ebben az egyszerű alkalmazásban az eredeti credit score tartalmaz minden „külső" információt az adathiány-mechanizmus becsléséhez. (Ennél összetettebb elbírálási folyamat esetén azonban pótlólagos információk is szükségesek a mechanizmus leírásához.) Másrészt „belső" információként az elfogadott hiteleken belüli rosszak aránya szintén felhasználható az adathiány-mechanizmus becsléséhez. A becsléshez használható például lineáris vagy exponenciális extrapoláció.

  Chen és Astebro [2003] a külső és belső információk súlyozott átlagát használta az adathiány-mechanizmus leírásához.

  A módszer előnye a kidolgozott elméleti háttér, a relatíve egyszerű alkalmazás, és hogy könnyen kiterjeszthető többszörös imputáció alkalmazására is.

  A szerzőpár eredményei szerint ez a módszer – nem véletlen adathiány okozta szelekciós torzítás esetén – javítja a modell klasszifikációs erejét. A módszer igényli az adathiány-mechanizmus becslését. A credit scoring modell klasszifikációs ereje növelhető azáltal, hogy a hiányzó értékek imputációjához felhasználják a rosszak arányáról elérhető információkat (*belső információ* az elfogadottakból, *külső információ* a régi scoring építésénél felhasznált teljes adatokból, megfelelően súlyozva). Azt találták, hogy NMAR esetén (a tréningadatokon) ez a bayesi módszer jobb, mint a Heckman-féle kétváltozós kétlépcsős modell. Ha viszont az adathiány inkább MAR-jellegű (a tesztadatokon), akkor *a modell gyengébb teljesítményű, mint a pusztán az elfogadottakon épített modell.*

  Az eredményeik alapján – bár nagyobb elutasítási arány esetén nagyobb szükség van a szelekciós torzítás csökkentésére –, a kipróbált módszerek hatékonysága és prediktív ereje csökkent erősebb szelekció esetén.

2.3.6. Maximum likelihood alapú módszer

A Chen és Astebro [2006] által javasolt modell a hagyományos maximum likelihood megközelítésen alapul. Bár ők logitmodellnél használták, a módszer alkalmazható minden maximum likelihood alapú eljárás esetén. Ez a reject inference technika beépíti az adathiány-mechanizmus okozta bizonytalanságot a modellépítésbe.

A logitmodellek feltételezik, hogy létezik egy y^* mögöttes eredményváltozó, amit egy regressziós kapcsolat határoz meg: $y^* = \beta' x_i + u_i$, ahol x_i a magyarázó változók egy vektora,

13 Azaz a hiányzás valószínűségének leírására.

β a paraméterek vektora, u_i a hibatényező és y^* nem megfigyelhető. Csak az y dummyváltozót figyelhetjük meg, ami $y = 1$, ha $y^*>0$ és $y = 0$ egyébként (*Maddala* [1983]). Ekkor $P(y = 1) = P(u_i > -\beta' x_i) = 1 - F(-\beta' x_i)$, ahol F az u_i eloszlásfüggvénye. A megfelelő likelihood függvény pedig:

$$L(\beta) = \prod_{y_i=0} F(-\beta x_i) \prod_{y_i=1} (1 - F(-\beta x_i)) \ , \tag{1}$$

ennek a loglikelihoodja:

$$\log L(\beta) = \sum_{y_i=0} \ln(F(-\beta' x_i)) \sum_{y_i=1} \ln(1 - F(-\beta' x_i)) \ . \tag{2}$$

A logitmodell feltételezi, hogy az u_i eloszlásfüggvénye logisztikus. Ekkor:

$$p_i(y = 1) = 1 - F(-\beta x_i) = 1 - \frac{\exp(-\beta x_i)}{1 + \exp(-\beta x_i)} = \frac{1}{1 + \exp(-\beta x_i)}$$

és

$$p_i(y = 0) = F(-\beta' x_i) = \frac{\exp(-\beta' x_i)}{1 + \exp(-\beta' x_i)}, \tag{3}$$

ahol $p_i(y=j)$ az $y=j$ (j= 1 v. 0) becsült valószínűsége az i megfigyelés esetén.

Ebben a modellben az y eredményváltozó értékét minden esetben ismerjük, nincs hiányzó adat. A credit scoring területén azonban az elutasítottak esetében nem tudjuk megfigyelni a hitelkockázatot leíró eredményváltozó értékét.

Legyen λ_i a hiányzás valószínűsége az i esetre, ahol a hitelkockázat (eredményváltozó) nem megfigyelhető. Jelölje továbbra is $y = 1$ a rossz hiteleket, $y = 0$ pedig a jókat. Ekkor háromféle mintaelemünk lesz: a jók, a rosszak és az elutasítottak, akiknél nem ismerjük az y-t. Ekkor, ha az előző megoldáshoz (a BC-modellhez) hasonlóan feltételezzük, hogy a hiányzás valószínűsége megegyezik a bedőlés valószínűségével, a loglikelihood várható értéke a következő lesz:

$$\log L(\beta) = \sum_{y_i=0} \ln(F(-\beta' x_i)) \sum_{y_i=1} \ln(1 - F(-\beta' x_i)) + \tag{4}$$

$$\sum_{y_i=hiányzó} ((1 - \lambda_i) \ln(F(-\beta x_i)) + \lambda_i \ln(1 - F(-\beta x_i)))$$

Ezek után már csak az a kérdés, hogyan modellezzük az adathiány-mechanizmust.

Ha az adathiány-mechanizmus MAR lenne, akkor a modell paramétereit megkaphatnánk az EM-algoritmussal (*Dempster*, *Laird*, Rubin [1977]). Mivel az EM-algoritmus feltételezi a véletlen adathiányt, csak a hiányzás valószínűsége becsülhető az elfogadottakból.

Ha azonban NMAR-adathiánnyal van dolgunk, akkor a csak az elfogadottakat tartalmazó mintából nem becsülhető a hiányzás valószínűsége, mert ez a minta nem reprezentálja az

egész sokaságot. NMAR esetén tehát a (4) egyenlet adja a loglikelihood függvény korrekt formáját.

A módszer sikeres alkalmazásának kulcsa egyedül a λ hiányzási valószínűség megfelelő becslése.

Chen és Astebro [2006] most is (akárcsak a bayesi BC-modellnél) a külső és belső információk súlyozott átlagát használta az adathiány-mechanizmus leírásához. Most is az eredeti credit score-t tekintették „külső" információforrásnak az adathiány-mechanizmus becsléséhez. Másrészt „belső" információként az elfogadott hiteleken belüli rosszak arányát használták az adathiány-mechanizmus becsléséhez. A becsléshez használható például lineáris vagy exponenciális extrapoláció.

Eredményeik szerint a javasolt maximum likelihood módszer a többi technikához képest jobban teljesített a modellépítési mintán, de a külön tesztelésre szánt mintán már *nem*. Ennek véleményük szerint az lehetett az oka, hogy NMAR-adathiány helyett inkább MAR-adathiánnyal volt dolguk, ekkor pedig elégséges csak a megfigyelteken modellt építeni, hiszen az is torzítatlan és hatásos lesz.

3. A szakirodalomból levonható következtetések

Az előzőekben bemutattuk azokat a szakirodalomban fellelhető módszereket, amelyek a scoringmodelleknél fellépő szelekciós torzítás csökkentését szolgálják. Mindegyik módszer valamilyen módon felhasználja az elutasítottakról meglévő információkat.

Az elutasítottak tényleges visszafizetési adatait nem ismerjük, ezért – mivel a semmiből nem keletkezhet új információ –, ha fel akarjuk használni azokat a modellépítéshez, akkor vagy *feltételezésekkel* kell élnünk, vagy *pótlólagos információ*t kell szereznünk a visszafizetési viselkedésükről.

Megmutattuk ezen (reject inference) technikák elméleti hátterét, kiemelve az alkalmazott feltételezéseket vagy a pótlólagos információ szerzésének és felhasználásának módját, és összefoglaltuk az eddigi gyakorlati tapasztalatokat.

Összegezve elmondható, hogy az elutasítottak alkalmazása a modellépítés során csak akkor lehet értelmes és hasznos megoldás, ha *bizonyos feltételek teljesülnek* az elfogadott és az elutasított sokaságra. A gyakorlatban működhetnek ezek a megoldások, mert a feltételezések sokszor indokoltak, vagy legalábbis jó irányba mutatnak. Például ésszerű feltételezés, hogy a rosszak aránya nagyobb az elutasítottakon belül, mint az elfogadottakon belül (azonos score mellett is), még akkor is, ha nem tudjuk korrekten számszerűsíteni, hogy menynyivel nagyobb. Az elutasítottak tényleges és imputált adatai alkalmazásának haszna függ az elutasítási aránytól, a mintabeli és sokasági eloszlásoktól és az alkalmazott statisztikai feltételek teljesülésétől. Van néhány portfólió, ahol nagyon alacsony az elutasítottak aránya (ilyen például a jelzáloghitelek piaca). Ekkor felesleges az elutasítottakkal foglalkozni, mert elhanyagolható az arányuk a populáción belül, így az általuk okozott torzítás sem igényel korrekciót. A nagyobb kockázatú portfóliók esetén viszont – például a kis- és kezdő vállalkozások hitelezésénél – az elutasítási arány igen nagy lehet, így a szelekciós torzítást már nem lehet figyelmen kívül hagyni.

Az alkalmazandó legjobb megoldás esetenként (ügyfélcsoportonként, termékenként) más-más lehet. Nincs kidolgozott elméleti háttér arra vonatkozóan, hogy milyen feltételek esetén okoz az elutasítottak kimaradása a modellből jelentős torzítást a paraméterbecslésekben. Nehéz is lenne ilyen általános alapelveket lefektetni, mert a torzítás erősen adatbázisfüggő.

Külföldön a scorecard-fejlesztők már alkalmaznak reject inference technikákat, amelyben statisztikai szoftvercsomagok (például SAS) is segítik őket. Ezek azonban sokszor fekete dobozként üzemelnek, mert a mögöttük lévő alapelvek és feltételezések nem világosak a felhasználók számára.

Az üzleti életben alkalmazott megoldások sokszor kétséges feltételezéseken alapulnak, amelyek teljesülése általánosságban nem tesztelhető, így – a szakirodalom áttanulmányozása után – arra a következtetésre jutottunk, hogy *a torzítás csökkentésének egyetlen robusztus és megbízható módja, ha az elutasítottak egy részét ténylegesen meghitelezik, és így figyelik meg viselkedésüket, valamint esetleges bedőlésüket.*

Gyakorlati szempontból jó lenne elkerülni a nem véletlen szelekciós mechanizmust. A hitelezők általában tudják, hogy milyen szabályokat alkalmaztak a múltban a befogadási döntések meghozatalakor, ezeket rögzíteni kell a későbbi elemzések érdekében. A scoringfüggvény felülbírálása (override) esetén – mind a kivételágon való beengedés, mind az ügyintézői elutasítás esetén – megérné a fáradságot a scoringfüggvény felülbírálásánál alkalmazott indokok, okok, jellemzők összegyűjtése és az adatbázisban való rögzítése. Ekkor persze további problémákat jelenthet a szubjektivitás és az adatok minőségének kérdése.

Ezeknek a statisztikai, ökonometriai modelleknek olyan pénzügyi szolgáltatásoknál van létjogosultsága, ahol *tömegszerű* kiszolgálás történik, azaz főleg a lakossági és kisvállalkozási (relatíve kis összegű és nagy számosságú) hitelek esetében. Ezeknél a hiteleknél viszont meglehetősen ritka a modellek felülbírálata. Tehát a gyakorlatban ebben a szegmensben leginkább véletlenszerű adathiánnyal (MAR) találkozunk.

Pótlólagos információkra azonban még akkor is szükségünk lehet, ha tökéletesen le tudjuk írni a szelekciós mechanizmust a meglévő változóinkkal, azaz véletlenszerű adathiányunk van (MAR). A kérelmek elfogadására/elutasítására használt credit scoring modell ugyanis idővel elveszti aktualitását, pontosságát, ezért újra kell építeni. Ha az eredeti modellünk a kérelmezők valamelyik (egy ismérv alapján képzett) csoportját mindig elutasította (például a büntetett előéletűeket), akkor reject inference nélkül a végső scorecardban nem jelenne meg ez az ismérv. Mi azonban tudjuk, hogy ez az ismérv is fontos volt a múltban (mivel rögzítünk minden változót, amit a múltban használtunk), és beépítenénk a modellbe. De ha nem vagyunk biztosak abban, hogy továbbra is minden büntetett előéletűt el kell utasítani (hiszen időközben megváltozhatott a magyarázó változók hatása), akkor vagy feltételezésekkel élünk, vagy szükségünk van ebből a csoportból is megfigyelésekre, azaz pótlólagos információkat kell használnunk.

A következő részben egy valós banki adatbázison megvizsgáljuk az általunk legjobbnak tartott módszer, a pótlólagos információszerzés által elérhető modelljavulást, annak költségeit és várható hasznait. A pótlólagos információkat a *résnyire nyitott kapu* módszerével, költségoptimális mintaelosztással szerezzük.

4. A szelekciós torzítás csökkentésének empirikus vizsgálata

A szelekciós torzítás csökkentésének egyetlen robusztus és megbízható módja az, ha az elutasítottak egy részét ténylegesen meghitelezik, és így figyelik meg viselkedésüket, valamint esetleges bedőlésüket.

Kétségtelen, hogy *pótlólagos információk felhasználásával* minden szempontból javítani tudunk a modellen, hiszen ekkor valóban több információra támaszkodunk a modellépítés során. Ezt az utat azonban nem mindig lehet megvalósítani a megoldás pénz- és időigényes volta miatt. Az eljárás költségei csökkenthetők a *résnyire nyitott kapu* alkalmazásával.

A szelekciós torzítás csökkentésére szolgáló módszerek vizsgálatához szükség van egy olyan adatbázisra, amelyben senkit nem utasítanak el. Az ezen a teljes mintán épített scoringmodell az *etalonmodell*. Ez az elméletileg létező legjobb modell, amit a valóságban (ha vannak elutasítottak) nem ismerünk. Majd az elutasítást szimulálva, létrehozunk egy csak az elfogadottakat tartalmazó mintát. Az ezen a mintán épített scoringmodell lesz a *kiinduló modell*. Az etalonmodell jobb lesz, többek között azért, mert nem tartalmaz szelekciós torzítást. Ezek után alkalmazhatjuk a *résnyire nyitott kapu* módszert, s az így létrejött, új adatbázist megfelelően súlyozva építhetjük a javított *nyitott kapu modelleket*, hogy csökkentsük a szelekció által okozott torzítást a kiinduló modellben. Ezután tesztelhetjük a javulás mértékét, vagyis azt, hogy mennyire sikerült közelíteni a kiinduló modellt az etalonhoz. Végül megvizsgáljuk a módszer költségeit és várható hasznát. A következő hipotéziseket fogjuk vizsgálni.

4.1. Hipotézisek

1. Erősebb szelekció (magasabb elutasítási arány) esetén gyengébb teljesítményű modellek építhetők.
2. A résnyire nyitott kapu módszerrel javítani lehet a modelleket.
3. A modelljavulás által elérhető többlethaszon egy bizonyos üzemméret (portfólióvolumen) fölött meghaladja az információszerzés költségeit.

4.2. Adatbázis

A kutatáshoz egy magyarországi bank bocsátotta rendelkezésünkre a fenti elvárásoknak eleget tevő adatbázisát. Az adatbázis egy olyan lakossági hiteltermék (hitelkártya) fogyasztóiról tartalmaz adatokat, amelynél egy adott időszakban majdnem mindenkit beengedtek (éppen scoringépítési céllal). A nagyon kis arányú elutasítás miatt teljesnek tekinthetjük az adatbázist. Ez a *teljes minta* 2279 ügyfél adatait tartalmazza, akik közül 381 volt rossz (nem fizető), a többi 1898 pedig jó ügyfél. Az adatbázisban csak kategóriás változók szerepelnek. (A kategóriás változók a modellekben dummyváltozókkal szerepelnek a kategóriák felsorolásának sorrendjében, mindig az utolsó kategória a referenciacsoport.)

2. táblázat

Az adatbázisban szereplő változók

APPLICATION ID	Azonosító
CSALÁDI ÁLLAPOT	Családi állapot (egyedülálló / élettársi kapcs. / elvált/ házas / özvegy)
FOGLALKOZÁS	Foglalkozás (alkalmazott vezető / fizikai alkalmazott/ közalkalmazott, köztisztviselő / vállalkozás tulajdonosa / nyugdíjas / szellemi alkalmazott)
FSZLAVEZETŐ BANK	Számlavezető bank (0: ez a bank / 1: másik bank)
ISKOLAI VÉGZETTSÉG	Iskolai végzettség (8 általános vagy kevesebb / érettségi / felsőfokú/ szakképesítés)
LAKÁS JOGCÍM	Lakásjogcím (bérlő / családtag / egyéb / tulajdonos)
NEM	Nem (0: nő / 1: férfi)
ÜGYFÉLTÍPUS	Kártyatípus (0: A / 1: B)
BUDGET JÖVEDELEM	Jövedelem kategorizálva (kvintilisek)
ÉLETKOR	Életkor kategorizálva (kvintilisek)
DEFAULT	Visszafizetés (1: rossz adós / 0: jó adós)

4.3. A modellezés folyamata

Az adatbázist szétválasztjuk modellépítésre (*tréning*) és ellenőrzésre (*teszt*) használt részre (2/3–1/3 arányban véletlen kiválasztással), így elkerülhető, hogy a modellek jóságát vagy a javulás mértékét a ténylegesnél többre értékeljük. Minden modellt a tréningadatbázison (vagy annak egy részén) építünk, de a modellek teljesítményét a tesztadatokon mérjük.

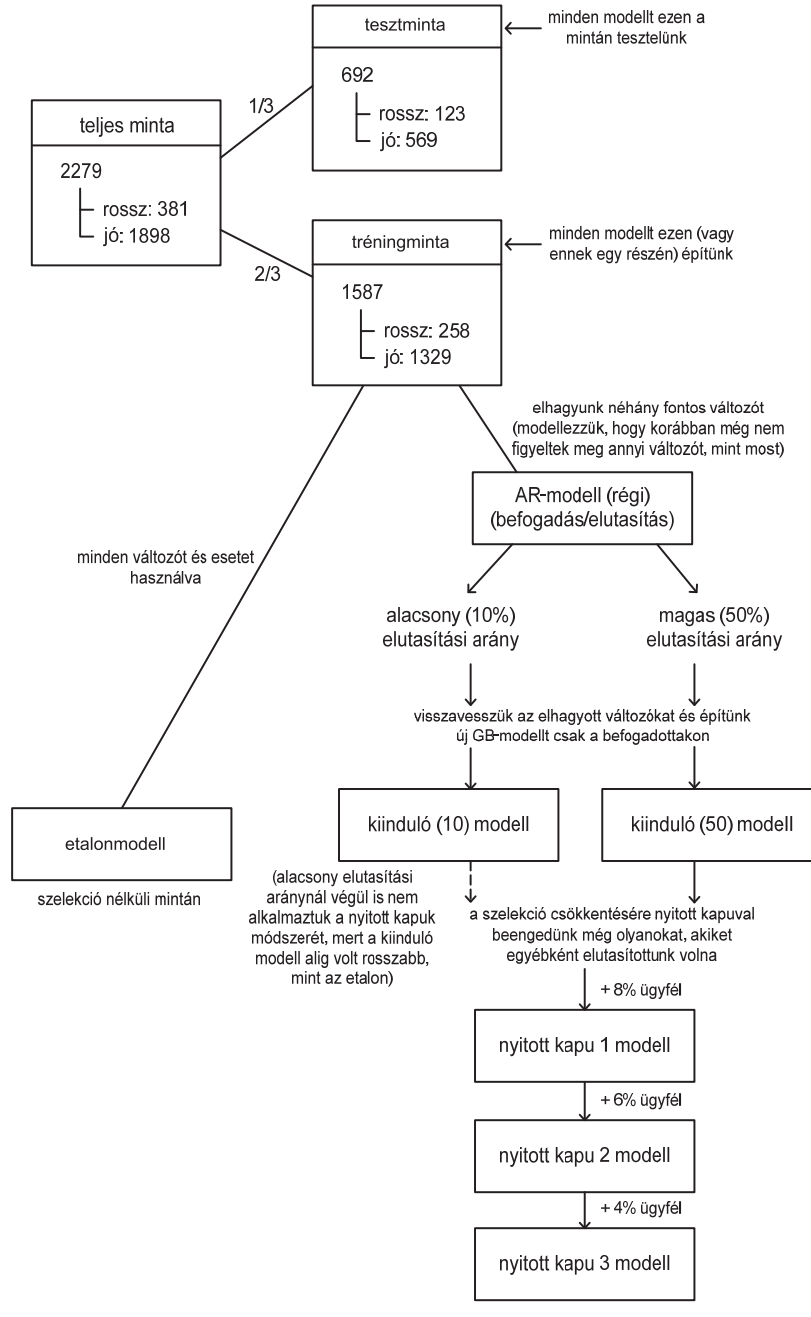
A modelleket SPSS-programcsomag segítségével, logisztikus regresszióval[14] építjük, mert ez a módszer alkalmas a kategóriás változók kezelésére, ráadásul napjainkban ez a leggyakrabban használt klasszifikációs eljárás a credit scoring területén. Minden modellt ugyanazzal az algoritmussal építünk (backward stepwise likelihood ratio 5%-os beléptetési, 10%-os kiléptetési szignifikanciaszint beállításával), így a modellek közötti különbségek csak a minta különbözőségének tudhatók be.

A *2. ábra* mutatja a modellezés folyamatát:

14 A logisztikus regresszióról l. például Hajdu [2003].

2. ábra

A modellezés folyamata



4.4. Eredmények

A tréningmintán megépítettük az *etalonmodellt*; ez most számunkra a létező legjobb modell, mert egy teljesen véletlen adatbázison épült.

Az etalonmodell paramétereit és illeszkedési mutatóit láthatjuk az alábbi táblázatokban.[15]

3. táblázat

Az etalonmodell paraméterei

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 5(a)	NEM(1)	,801	,159	25,366	1	,000	2,228
	Életkor			61,212	4	,000	
	Életkor (1)	1,803	,298	36,563	1	,000	6,071
	Életkor (2)	1,378	,305	20,359	1	,000	3,967
	Életkor (3)	,605	,328	3,395	1	,065	1,831
	Életkor (4)	,569	,338	2,835	1	,092	1,766
	isk. végzettség			50,598	3	,000	
	isk. végzettség (1)	–,990	1,048	,893	1	,345	,372
	isk. végzettség (2)	–,436	,165	6,956	1	,008	,647
	isk. végzettség (3)	–2,399	,340	49,771	1	,000	,091
	BUDGET			16,132	4	,003	
	BUDGET (1)	–,300	,275	1,190	1	,275	,741
	BUDGET (2)	–,541	,252	4,615	1	,032	,582
	BUDGET (3)	–,933	,261	12,735	1	,000	,393
	BUDGET (4)	–,664	,259	6,573	1	,010	,515
	kártyatípus (1)	,580	,181	10,286	1	,001	1,786
	Constant	–2,354	,375	39,332	1	,000	,095

Tehát szignifikáns magyarázó változók lettek: a nem, életkor, iskolai végzettség, jövedelem (BUDGET) és a kártyatípus változók. Például az életkor (1) dummyváltozó B=1,803-as paramétere így értelmezhető: Exp(B) = 6,071, ami azt jelenti, hogy a legfiatalabbak esetén a p/(1–p) odds értéke várhatóan 6,071-szeresére nő a legöregebbekéhez képest minden más magyarázó változó változatlansága esetén.[16]

15 Csak erre az egy modellre vesszük részletesen végig az outputokat, a többi modellnél csupán egy összefoglaló táblázatot közlünk.

16 A legfiatalabbak az életkor szerinti első kvintilisbe esők, a legöregebbek az ötödik kvintilisbe esők, a *p* a nemfizetés becsült valószínűsége (PREPD).

4. táblázat

Az etalonmodell illeszkedési mutatói a tréningadatbázison

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1132,707 [a]	,158	,269
2	1132,812 [a]	,158	,269
3	1136,158 [a]	,156	,266
4	1141,521 [b]	,153	,261
5	1148,883 [b]	,149	,254

a. Estimation terminated at iteration number 7 because parameter estimates changed by less than ,001.

b. Estimation terminated at iteration number 6 because parameter estimates changed by less than ,001.

A logitmodell paramétereinek becslése maximum likelihood módszerrel történik. A likelihood maximalizálása ekvivalens a -2 Log likelihood minimalizálásával, tehát minél kisebb a -2 Log likelihood, annál jobb a modell. A likelihood értékét az üres modelléhez kell hasonlítani. Ezt teszi a Cox–Snell R^2, amelynél a nagyobb értékek jelzik a jobb modellt. Ez a mutató hasonló a sokváltozós lineáris regressziónál alkalmazott R^2 mutatóhoz, de a maximuma nem 1. A Nagelkerke R^2 már 0–1 közötti értékeket vehet fel, és hasonlóan értelmezhető, mint a többszörös determinációs együttható.[17] Az etalonmodellünk magyarázó ereje tehát 25,4%-os.[18]

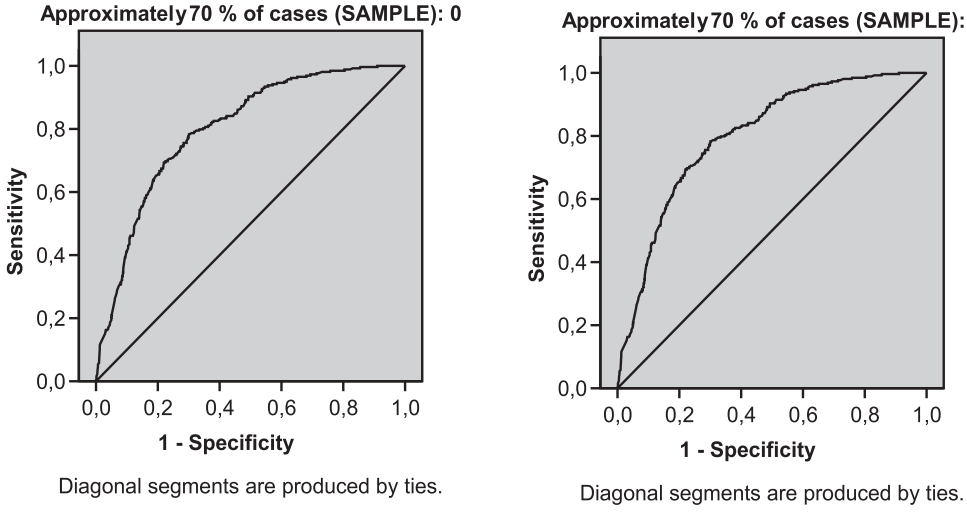
A scoringmodellek értékelésére a gyakorlatban leginkább a ROC-görbét, illetve a görbe alatti területet (AUROC) alkalmazzák.

17 $R^2_{Cox-Snell} = 1 - \left(\frac{L_{null}}{L_{aktuális}} \right)^{2/n}$, $R^2_{Nagelkerke} = \frac{R^2_{Cox-Snell}}{\max R^2_{Cox-Snell}}$

18 Mindig az utolsó lépésbeli mutatókat kell néznünk. Mivel backward eljárással építettük fel a modellt, az első lépésben a legnagyobb az R^2, mert ott van a legtöbb magyarázó változó; ha elhagyunk magyarázó változókat, csökken (nem nő) az R^2.

3. ábra

ROC-görbék az etalonmodellre a tréning- és a tesztadatokon



5. táblázat

Az AUROC értéke az etalonmodellnél a tréning- és a tesztadatokon[19]

Test Result Variable(s): Predicted probability

Approximately 70 % of cases (SAMPLE)	Area	Std. Error [a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
				Lower Bound	Upper Bound
0	,800	,014	,000	,773	,828
1	,782	,022	,000	,738	,826

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

c. For split file Approximately 70 % of cases (SAMPLE) = 0, the test result variable(s): Predicted probability has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

d. For split file Approximately 70 % of cases (SAMPLE) = 1, the test result variable(s): Predicted probability has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

Az AUROC értékének minimuma 0,5, maximuma 1; minél nagyobb, annál jobb a modell. Az etalonmodell esetében a modellépítésre szánt mintán a mutató értéke 0,8, a tesztelésre szánt mintán kicsit kisebb, 0,782. Azt a nullhipotézist, amely szerint az aktuális mo-

19 0: tréning, 1: teszt.

dellünk nem különbözik szignifikánsan a véletlenszerű besorolást jelentő (üres) modelltől, minden szokásos szignifikanciaszinten elvethetjük (p = 0,000).

A modellek teljesítményét a tesztadatbázison fogjuk összehasonlítani, ezért a leggyakrabban alkalmazott AUROC-on kívül kiszámítottuk a szakirodalom által leginkább ajánlott Brier-score és logaritmikus score értékét is a tesztadatokon.[20]

A Brier-score (etalon) = 0,122, a logaritmikus score (etalon) = 0,415. Mindkét mutatónál a kisebb értékek jelentik a jobb modellt.

A további modellek teljesítményét ehhez az etalonmodellhez fogjuk hasonlítani.

Ezek után *szimuláljuk a szelekciót*, azaz úgy dolgozunk, mintha a bank alkalmazott volna valamilyen szűrőt (beengedés/elutasítás vagy AR-modellt) az ügyfelek beengedésénél. Ezt a szűrőt úgy készítjük el, hogy a (tréning) adatbázison építünk egy logitmodellt, de úgy, hogy ne szerepeljenek benne a *jövedelem* és a *kártyatípus* változók. Ezen változók elhagyásával azt szeretnénk szimulálni, hogy a múltban még nem figyeltek meg annyi változót, mint most.[21] Az így létrejött AR-modell outputjait már nem mutatjuk be külön részletesen, hanem a *6. táblázatban* összefoglaljuk az összes modell jellemzőit:

6. táblázat

A modellek és jellemzőik összefoglaló táblázata

Modellek jellemzőik	Etalon	AR	K10	K50	NYK1	NYK2	NYK3
Nagelkerke R2	0,254	0,234	0,247	0,135	0,151	0,17	0,225
AUROC (trénig)	0,8	0,785	0,79	0,742	0,773	0,751	0,776
AUROC (teszt)	0,782	0,769	0,786	0,694	0,782	0, 791	0,786
Brier score (teszt)	0,122	0,127	0,123	0,141	0,131	0,122	0,123
Logaritmikus score (teszt)	0,415	0,403	0,417	0,483	0,411	0,413	0,421
optimális cutoff (tréningen)				0,1	0,08	0,15	0,14
profit (teszten)				3,3	13,7	15,1	15,2
profit (teszten)a 0,1-es cutoff mellett				3,3	14,9	15,9	15,9
a kapu nyitás költsége a tréningen					7,3	6,4	8,1

20 A modellek teljesítményének értékeléséhez használható mutatók megtalálhatók például Oravecz [2007] cikkében.

21 A valóságban nemcsak a változók, hanem az esetek is mások voltak a régi modell építésénél, és az eltelt időszak alatt a kapcsolat jellege is változhatott, ennek hatására a valóságban nagyobb lehet a különbség a régen épített AR-modell és a most építhető legjobb etalonmodell között, de ennek vizsgálata nem célja a kutatásnak, és nem is tudnánk beépíteni a modellezésbe, mert csak egy időszakból vannak adataink.

7. táblázat

A modellek magyarázó változói

Modellek jellemzőik	Etalon	AR	K10	K50	NYK1	NYK2	NYK3
nem	√	√	√		√	√	√
életkor	√	√	√	√	√	√	√
foglalkozás		√		√			
iskolai végzettség	√	√	√		√	√	√
jövedelem	√		√		√	√	√
kártyatípus	√		√	√	√	√	√
családi állapot							√
számlavezető bank							
lakásjogcím						√	

Szimulációnk szerint tehát az AR-modellt használta a bank az ügyfelek beengedésére/elutasítására. Ezek után kétféle beengedést modellezünk, egy *alacsony* (nagyjából 10%-os) és egy *magas* (nagyjából 50%-os)[22] *elutasítási arány* melletti beengedést.

Úgy gondoljuk, hogy ez a modell időközben elavult, ma már több változót is ismerünk, ezért frissíteni akarjuk ezt a régi (AR-) modellt, és egy új (GB-) modellt készítenénk. Ha tehát elutasították volna a kérelmezők egy részét (10 vagy 50%-át), akkor az adatbázisunkból hiányozna az esetek 10 vagy 50%-ában a visszafizetést leíró eredményváltozó értéke, és ez az adathiány jelen esetben nem teljesen véletlenszerű (nem MCAR), hanem a szelekciós modell használata miatt MAR-jellegű.[23] Ezen a mintán tehát építünk egy új modellt, amelyhez már minden elérhető magyarázó változót felhasználunk. (Pontosabban két modellt építünk, mert egy alacsony és egy magas elutasítási arányú szcenáriót is megvizsgálunk.) Ez a (két) új modell (kiinduló [GB-] modell) azonban szelektált mintán épül, és a szelekció nem teljesen véletlenszerű.

Az előző összefoglaló táblázat tartalmazza a kiinduló modellek (K10 és K50) jellemzőit is.

Az első hipotézisünk az volt, hogy erősebb szelekció (magasabb elutasítási arány) esetén gyengébb teljesítményű modellek építhetők. A hipotézis helyességét az *eredményeink is alátámasztják*.

Alacsony elutasítási arány (10%) esetén a kiinduló modell (K10) teljesítménymutatóinak az értéke a tesztadatbázison hasonló az etalonmodell értékeihez, tehát a modell teljesítmé-

22 Nem tudjuk pontosan 10 és 50%-ra beállítani az elutasítási arányt, mert kategóriás változóink vannak, és sok az egyforma eset.

23 A formális szelekciós modell felülbírálata (override) manapság lakossági ügyfelek esetén nem olyan nagy volumenű, ezért ennek a modellezésétől eltekintünk, így a nem véletlen adathiány (NMAR) modellezésétől is.

nyén nincs mit javítani. (Az AUROC értéke [0,786] még jobb is, mint az etalonmodell esetén [0,782], de a különbség nem szignifikáns.)

Magas elutasítási arány (50%) esetén már rosszabb a kiinduló modellünk (K50) teljesítménye, mint az etaloné és a K10 modellé. Az AUROC, a Brier-score és a logaritmikus score szerint is gyengébb a teljesítmény.

Nézzük, mi lehet ennek az oka! A múltban fontos volt a nem, életkor, foglalkozás és iskolai végzettség (ezek az AR-modell szignifikáns változói). Tudjuk, hogy most is fontosak, és ezekhez adódnak az újonnan megfigyelt változók is (jövedelem, kártyatípus), vagyis az etalonmodell szignifikáns változói.

A K50 modell akkor lenne jó, ha ugyanazokat a változókat tartalmazná, mint az etalon. Azok közül viszont csak az életkor és a kártyatípus vált szignifikánssá. Ennek lehet pusztán az az oka, hogy feleakkora a minta, és egyszerűen a kisebb elemszám miatt nem tűnnek szignifikánsnak a paraméterek. Az is lehet az ok, hogy az AR-modell erősen szelektált, és például az életkor szerinti kockázatosabb csoportból (a fiatalokból) alig engedett be valakit, és az így szelektált mintán már nem szignifikáns a változó. Továbbá lehet azért is, mert az erős szelekció miatt kevés a rossz eset, így nem építhető jó modell.

A K50 modell tehát gyengébb, mint az etalon és a K10, ezért érdemes lehet megpróbálkozni a javításával.[24]

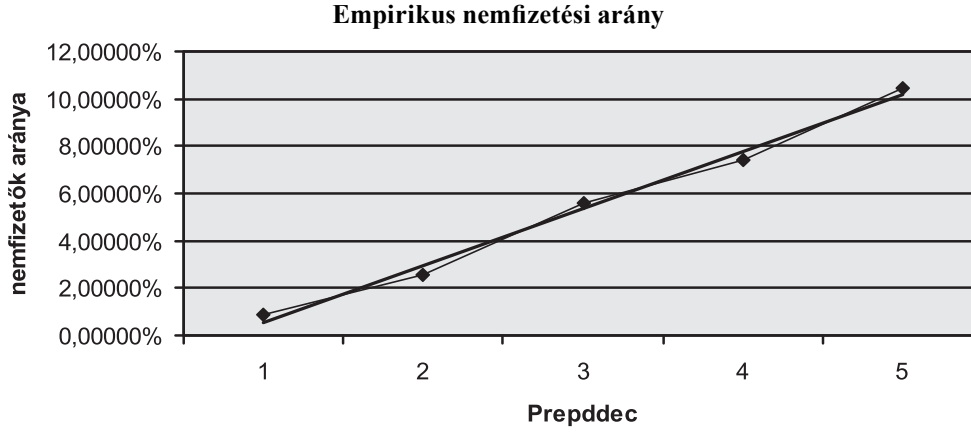
Azt láttuk, hogy a modellek javításának egyetlen megbízható és robusztus módja a pótlólagos információk felhasználása. Erre a *résnyire nyitott kapu* módszert fogjuk használni, költségoptimális mintaelosztással. Azaz minden, egyébként elutasított ügyfélnek van esélye a mintába kerülésre, de nem egyforma valószínűséggel. Nagyobb valószínűséggel kapnak hitelt azok, akiknél a várható veszteség kisebb, és kisebb valószínűséggel azok, akiknél ez a várható veszteség nagyobb. A várható veszteség függ a nemfizetés valószínűségétől, a hitelösszeg és a fedezet nagyságától. A vizsgált hitelkártya a termék jellegéből adódóan fedezetlen hitel, és a kártya hitelkerete azonos minden ügyfél esetében, tehát a várható veszteség ebben a kutatásban csak a nemfizetési valószínűségtől függ.[25]

Ha a régi AR-modell által prediktált nemfizetési valószínűség (Prepd) szerint (növekvő) sorba rendezzük az ügyleteket, és megnézzük, hogy az egyes decilisekben mennyi volt a tényleges nemfizetési arány, akkor az első 5 decilisre van megfigyelésünk, hiszen 50% volt a beengedési arány.

24 A valóságban az etalonmodellt soha nem ismerjük. Most azért készítettük el, hogy lássuk, egyáltalán mekkora az elérhető maximális javulás, és ennek tükrében értékeljük majd a javított modelljeinket.

25 Az adatbázis nem tartalmazza, hogy a hitelkeretből mekkora részt használtak fel, ezért feltételezzük, hogy mindig a teljes hitelkeretet kihasználják.

4. ábra

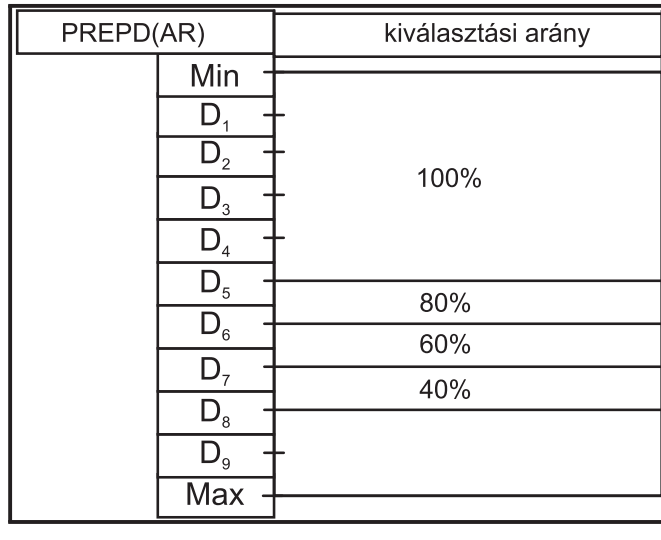


Azt látjuk, hogy az empirikus nemfizetési arány közelítőleg lineárisan nő a megfigyelhető decilisekben, és feltételezzük, hogy ez a tendencia az elutasítási tartományban is folytatódik (lineáris extrapoláció).[26]

Nyissuk ki a kaput résnyire, és az egyébként elutasítandók egy részét is hitelezzük meg úgy, hogy a mintába kerülés valószínűsége csökkenjen, ha a nemfizetés valószínűsége nő.

5. ábra

A mintába kerülés valószínűsége a prediktált bedőlési valószínűség függvényében



26 A valóságban csak feltételezhetjük ezt a tendenciát, most viszont, mivel az egyébként elutasítandókról is van adatunk, ellenőrizhetjük, hogy helytálló-e a feltételezés. (A lineáris tendencia folytatódott.) Valójában itt tehát mégis alkalmazunk egy feltételezést, amelynek a helyessége csak a beengedés után tesztelhető.

Úgy választottuk meg a mintába kerülés valószínűségét, hogy az lineárisan csökkenjen az elutasítási tartomány mentén. Ez csak egy lehetséges elosztás. Ha a bank erre a célra kihelyezhető tőkéje kisebb, akkor ennél kisebb kiválasztási arányokat kell beállítanunk; ha nagyobb, akkor lehet nagyobbakat.

A résnyire nyitott kapun való beengedést három fokozatban hajtjuk végre.

Az első esetben csak +8% ügyfelet engedünk be, az eddig beengedettekhez közelálló esetek 80%-át (az ötödik és hatodik decilis közötti tartományból véletlen kiválasztással, 80%-os kiválasztási aránnyal). Az így beengedett +134 ügyfélből 111 lett jó és 23 rossz. Ez lesz a NYK1-es minta. Ezt a mintát hozzárakjuk a kiinduló modellünk adatbázisához, és az így létrejött mintán megépítjük a NYK1-es modellt.

A második fokozatban az előzőhöz képest beengedünk még 6%-nyi ügyfelet (a hatodik és hetedik decilis közötti tartományból véletlen kiválasztással, 60%-os kiválasztási aránnyal). Az így beengedett ügyfelek közül 80 volt jó és 18 rossz. Ezekkel az esetekkel bővítjük az előző adatbázist, és megépítjük a NYK2-es modellt.

A harmadik fokozatban az előzőhöz képest beengedünk még 4%-nyi ügyfelet (a hetedik és nyolcadik decilis közötti tartományból véletlen kiválasztással, 40%-os kiválasztási aránnyal). Ezekkel az esetekkel (39 jó és 15 rossz) bővítjük az előző adatbázist, és megépítjük a NYK3-as modellt.

A legrosszabbnak tűnő esetekből (a prediktált bedőlési valószínűség szerinti felső 20%-ból) nem engedünk be eseteket.

Az NY modellek építése előtt *az adatokat át kell súlyoznunk*, mert tudjuk, hogy a nyitott kapuval beengedett esetek arányosan több ügyfelet képviselnek, ezért első körben minden megfigyelést átsúlyozunk a bekerülési valószínűség reciprokával. Ekkor viszont a kapott súlyok összege nagyobb lesz, mint a tényleges esetszám, ezért minden súlyszámot beszorzunk a tényleges esetszám és a kapott súlyok összegének hányadosával, így kapjuk az alkalmazandó végleges súlyokat.

A fenti súlyokkal épített nyitott kapu modellek (NYK1, NYK2, NYK3) jellemzőit is tartalmazza a korábbi összefoglaló táblázat.

A vizsgálandó második hipotézisünk az volt, hogy a résnyire nyitott kapu módszerrel javítani lehet a modelleket.

Pusztán elméleti alapon is azt várjuk, hogy nagyobb mintán jobb modellt lehet építeni. Itt azonban már a kiinduló modell elemszáma is elég nagy, így nem biztos, hogy pusztán az elemszám növelésével sokat lehet javítani a modellen. Ráadásul a modellek jóságát nem a modellépítési adatbázison, hanem egy attól eltérő tesztadatbázison vizsgáljuk. Tehát egyáltalán nem biztos, hogy javulni fog a modellünk. Láthattuk például, hogy a 10%-os elutasításnál a kiinduló modell (K10) nem rosszabb, mint az etalonmodell, pedig a mintanagyságban 10%-os eltérés volt.

Most a kiinduló (K50) és az első résnyire nyitott kapu (NYK1) modell mintanagysága között csak 8%-os eltérés van. A modell teljesítménye viszont sokat javult. Az AUROC értéke 0,694-ről 0,782-re nőtt, ami igen jelentős javulásnak tűnik, bár a különbség 5%-os szignifikanciaszinten nem szignifikáns (összeérnek a konfidenciaintervallumok), de 10%-on már igen.

A Brier-score értéke 0,141-ről 0,131-re csökkent, a logaritmikus score pedig 0,483-ról 0,411-re, ami szintén javulást jelent.

Azt látjuk tehát, hogy a kapu kinyitásával javult a modellünk.

Ugyanakkor az is látható, hogy a modellünk teljesítménye nagyon közel került az etalonmodell teljesítményéhez, tehát ezek után már hiába nyitjuk a kaput, sokat nem fog javulni a modellünk. Sőt, ez az eredmény azt is jelzi, hogy ha az első hipotézis vizsgálatakor az alacsony és magas elutasítási arány hatásának vizsgálatához nem a 10 és 50%-ot, hanem mondjuk a 10 és 30 vagy a 10 és 40%-ot választottuk volna, akkor nem lett volna nagy különbség a modellek teljesítménye között. *Tehát csak a valóban magas (50% feletti) elutasítási arány esetén kell számolnunk a modellek teljesítményének romlásával.*

A második nyitott kapu modellhez (NYK2) nagyobbra nyitjuk a kaput, és további 6%-nyi (az előzőeknél kicsit rosszabbnak tűnő) ügyfelet engedünk, be. Az AUROC és a Brier-score szerint kicsit javult, a logaritmikus score szerint nem változott (picit romlott) a modell. A különbség az NYK1-hez képest nem szignifikáns.

Az NYK3 modellhez további 4%-nyi kicsit rosszabb ügyfelet engedtünk be. Itt már mindhárom mutató romlást mutat[27], de a különbség igen kicsi, nem szignifikáns.

Az NYK1-hez képest tehát nem jelentett javulást az NYK2 és NYK3 modell építése, de *a kiinduló modellhez képest mindhárom nyitott kapu modellnek jobb a teljesítménye.*

A statisztikus vagy modellező tehát örülhet, mert a nyitott kapu módszerrel sikerült javítani a modellek teljesítményét. De mit szólnak mindehhez a bank tulajdonosai, jelent-e számukra hasznot a modellek javulása?

A modellek javításához ugyanis többletinformációra volt szükségünk az egyébként elutasítandók visszafizetési viselkedéséről. Ez pedig plusz költséget jelentett, mert sok rossz ügyfelet is meghiteleztünk. Megéri-e ezt a többletköltséget felvállalni a jövőbeni többletprofit reményében?

Erre a kérdésre vonatkozik a *harmadik hipotézisünk:*

A modelljavulás által elérhető többlethaszon egy bizonyos üzemméret (portfóliómennyiség) fölött meghaladja az információszerzés költségeit.

Az információszerzés költségeinek kiszámításához meg kell néznünk, hogy a nyitott kapuval milyen ügyfélből mennyit engedtünk be. A plusz beengedett ügyfelek:

8. táblázat

Nyitott kapu módszerrel beengedett ügyfelek

	jó	**rossz**
NYK1	111	23
NYK2	80	18
NYK3	39	15

A költségek számszerűsítéséhez tudnunk kell, hogy mekkora a bank haszna a jó hiteleken, és mekkora a vesztesége a rossz hiteleken, azaz szükségünk van egy haszon- (vagy költség-) mátrixra. Feltételeztük, hogy a jó hiteleken a bank haszna 10%, a rossz hiteleken a vesztesége 80%, azaz a haszonmátrix az alábbi:

27 A romlás oka lehet, hogy az itt beengedett ügyfelek a modellépítés során nagy súlyt kaptak (a kis kiválasztási arány miatt), ezért egy-egy, a sokasági tendenciától eltérő ügyfélnek nagy lehet a modellre gyakorolt hatása.

9. táblázat

A bank haszonmátrixa

haszon		**valóságos kategória**	
		jó (G)	rossz (B)
a modell által besorolt kategória	jó (elfogadás) (A)	0,1*x*	–0,8*x*
	rossz (elutasítás) (R)	0	0

Mivel az *x* hitelösszeget most minden ügyfél esetén egyformának feltételezzük, tekinthetjük egységnyinek. Így a plusz ügyfelek által okozott veszteség (a többletinformáció költsége):

NYK1: 7,3 (egység), NYK2: 6,4 egység, NYK3: 8,1 egység.

Ahhoz, hogy megnézzük, mekkora haszonnövekményre számíthatunk a nyitott kapu alkalmazásának köszönhetően, meg kell határoznunk minden modell esetében a cut-off értékét, azaz, hogy milyen becsült nemfizetési valószínűség alatt engedjük be az ügyfeleket.

Mivel a K50 modellt javítottuk a nyitott kapu segítségével, ezért a K50 és az NYK modellekre kell meghatároznunk a profitmaximalizáló cut-off értékét.[28]

A cut-off érték kiválasztása több módon is lehetséges.

A *gyakorlatban* általában a cut-off értékek lehetséges tartományán megvizsgálják a modellépítési mintán a különböző cut-off értékekhez tartozó profit- (vagy hozam-) értékeket, és azt a cut-off értéket választják, amely mellett a mintán maximális a profit.

Elméletileg viszont akkor érdemes befogadni egy kérelmet, ha annak várható haszna pozitív (nagyobb, mint az elutasítás várható haszna), azaz a fenti haszonmátrix és *p* bedőlési valószínűség esetén, ha $(1–p)0,\ 1x + p(–0,8)x > 0$. Jelen esetben a $p < 0,0909$ bedőlési valószínűségű hiteleket érdemes beengedni.

Mindkét megoldás mellett megvizsgáltam az elérhető profitot. A *gyakorlati* módszer esetén készítettem egy Excel-fájlt, amely tetszőleges haszonmátrix esetén kiszámítja a 0–100%-ig[29] terjedő cut-off értékek mellett elérhető profitot. A fenti haszonmátrix mellett a vizsgálandó modelleknél kiválasztottam a profitmaximalizáló cut-off értéket a tréning adatbázison. (Az optimális cut-off értékek megtalálhatók a korábbi összefoglaló táblázatban és az alábbi táblázatban is.) Ha több maximumhelye volt a profitgörbének, akkor (a nagyobb piaci részesedés miatt) a nagyobbat választottam.

Az elméleti megoldás szerint a cut-off 0,0909, ami azt jelenti, hogy a 9%-os bedőlési valószínűségű ügyfeleket még be kell fogadni, a 10%-osakat el kell utasítani. Mivel 1%-os lépésközű profitszámítást készítettem, így itt az elméleti cut-off 10% (0,1).

Az így meghatározott optimális cut-off értékek mellett kiszámítottam a tesztadatbázison elérhető profitot:

28 Most elkülönülten csak az ezen az egy terméken elérhető profitot akarjuk maximalizálni.
29 1%-os lépésközzel.

10. táblázat

Költség és haszon

	K50	NYK1	NYK2	NYK3
optimális cutoff (tréningen)	0,1	0,08	0,15	0,14
profit (teszten)	3,3	13,7	15,1	15,2
profit (teszten)a 0,1-es cutoff mellett	3,3	14,9	15,9	15,9
a kapu nyitás költsége a tréningen		7,3	6,4	8,1

Azt látjuk, hogy *a kiinduló modellhez (K50) képest a nyitott kapuval óriási profitnövekedést értünk el (3,3-ról 13,7-re vagy 3,3-ról 14,9-re)*, majd a további kapunyitással tovább nőtt a profit, de már nem ilyen mértékben. Az eredményekből látható, hogy *az elméleti* 10%-os *cut-off alkalmazásával* minden esetben[30] *nagyobb profitot lehetett elérni*, *mint a tréningen empirikusan meghatározottal*. A gyakorlatban (és az oktatásban is) elterjedt megoldással szemben tehát könnyebb és érdemesebb is ezt használni.

De térjünk vissza erre a szinte hihetetlen profitnövekedésre! Ennek oka a mintában keresendő, amin a modellek épültek, illetve amilyen a valóság (teszt). A minták elemszáma és nemfizetés szerinti megoszlása az alábbi:

11. táblázat

Elemszám és nemfizetés szerinti megoszlás

	jó	rossz	összes
K50	766	37	803
NYK1	877	60	937
NYK2	957	78	1035
NYK3	966	93	1089
teszt	569	123	692

Láthatjuk, hogy az AR-modell 50%-os elutasítás mellett a rossz ügyfelek legnagyobb részét kiszelektálta, ezért a kiinduló (K50) modell adatbázisában csak 37 rossz ügyfél szerepel. Ilyen kevés rossz adóssal pedig nem lehet jó modellt építeni. A modell javításához tehát rossz ügyfelek adataira van szükség, és a nyitott kapuval sikerült is szert tenni ilyen rossz ügyfelekre.

Olcsóbb megoldás lenne persze, ha az ilyen ügyfelek jellemzőinek megismerését nem nekünk kellene finanszíroznunk, hanem a költséget megosztva, más bankoktól vagy hitelin-

30 Kivéve a K50 modellnél, mert itt a kétféle cut-off egyezik.

formációs rendszerekből megvásárolhatnánk. Amíg ez a módszer nem járható, addig marad a saját költségen való adatgyűjtés.

A 803 elemű mintához a +134 ügyfél beengedése a NYK1 modell építéséhez 7,3 egységbe került, de ez már egy egészen pici jövőbeli portfólión is megtérül, hiszen a 692 elemű tesztadatbázison a modelljavulás következtében 3,3-ról 14,9-re nőtt a profit.[31]

A NYK2 modellhez +98 ügyfelet engedtünk be, ami 6,4 egységnyi költséget jelentett, és a modelljavulás hatására további 1 egységgel nőtt a profit a 692 elemű tesztadatbázison. Hogy a költségek megtérüljenek, egy 6,4-szer ekkora jövőbeli portfólióra van szükség. Az 1035 fős modellezési mintához képest tehát 4,3-szor akkora jövőbeli várható portfólió mellett már megtérülnek a költségek.

Az NYK3-mal már nem sikerült javítani a modellt és növelni a profitot, tehát a plusz költség semmilyen volumen mellett nem térül meg.

5. Összefoglalás

Ha az adósminősítési modellek építéséhez csak a meghitelezett ügyfelek adatait használjuk – ami egy szelektált, nem reprezentatív mintát jelent –, akkor a modellünk túlzottan optimista lesz.

A dilemmára az elutasítottak jellemzőinek felhasználásával történő modellépítés (reject inference) jelenthet választ.

Az elutasítottak tényleges visszafizetési adatait nem ismerjük, ezért – mivel a semmiből nem keletkezhet új információ –, ha fel akarjuk használni őket a modellépítéshez, akkor vagy *feltételezésekkel* kell élnünk, vagy *pótlólagos információ*t kell szereznünk a visszafizetési viselkedésükről.

Az elutasítottak alkalmazása a modellépítés során csak akkor lehet értelmes és hasznos megoldás, ha *bizonyos feltételek teljesülnek* az elfogadott és az elutasított sokaságra. A gyakorlatban működhetnek ezek a megoldások, mert a feltételezések sokszor indokoltak, vagy legalábbis jó irányba mutatnak.

Az üzleti életben alkalmazott megoldások azonban sokszor kétséges feltételezéseken alapulnak, amelyeknek a teljesülése általánosságban nem tesztelhető, így *a torzítás csökkentésének egyetlen robusztus és megbízható módja, ha az elutasítottak egy részét ténylegesen meghitelezik, és így figyelik meg viselkedésüket, valamint esetleges bedőlésüket.*

Pótlólagos információk felhasználásával minden szempontból javítani tudunk a modellen, hiszen ekkor valóban több információra támaszkodunk a modellépítés során. A pótlólagos információ megszerzése azonban pénz- és időigényes megoldás. Az eljárás költségei csökkenthetők a *résnyire nyitott kapu* alkalmazásával, egyfajta költségoptimális mintaelosztással.

Ebben a tanulmányban egy valós banki adatbázison (lakossági hitelkártyaadatokon) vizsgáltuk az ezzel a módszerrel elérhető javulást, annak költségeit és várható hasznát.

Az empirikus kutatás során azt tapasztaltuk, hogy *magas elutasítási arány (erőteljes és nem teljesen véletlenszerű szelekció) mellett gyengébb teljesítményű modellek építhetők,*

31 Valójában ez egy jövőbeli profit, tehát diszkontálnunk kellene, mert a költségeket viszont most kell vállalnunk.

mint kisebb arányú elutasítás esetén. Ennek egyik oka, hogy ekkor kevés rossz ügyfél kerül a portfólióba, ami megnehezíti a modellek számára a rosszak karakterisztikáinak megismerését. Másik oka, hogy a szelekció hatására egyébként szignifikáns magyarázó változók bizonyos értékei nem kerülnek a mintába, aminek következtében a magyarázó változó már nem lesz szignifikáns.

Ilyen esetekben segíthet a pótlólagos információszerzés egyik módja: az, ha belső forrásból, a résnyire nyitott kapu alkalmazásával nyerünk új megfigyeléseket. Azt láttuk, hogy a *nyitott kapu módszerrel javult a modellek teljesítménye, és ennek következtében a termé-ken elérhető profit is nőtt.*

Azt tapasztaltuk, hogy ha a profitmaximalizálás a cél, akkor *jobb, ha az elméleti úton meghatározott cut-off értéket használjuk*, szemben a gyakorlatban elterjedt empirikus meghatározási móddal.

Eredményeink szerint a modelljavulás és a profitnövekedés mértéke az első lépcsőben volt a legnagyobb. Tehát *leginkább az egyébként befogadandókhoz közel álló, azoknál csak kicsit rosszabbnak tűnő ügyfelek közül érdemes résnyire nyitott kapuval beengedni még továbbiakat.*

Ez az első lépcsős, nagymértékű modelljavulás és profitnövekedés valószínűleg csak az adatbázis sajátossága, de egyéb, általános érvényű megfontolások is ezt a stratégiát sugallják. Az elfogadási tartományhoz közelre még jobbak a becsléseink. Ide még valószínűleg jól tudjuk becsülni a rosszak arányát, ezáltal a többletminta költségei tervezhetőbbek és kisebbek is, mintha egy távoli tartományból vennénk mintát.

Végezetül elmondhatjuk, hogy a tanulmányban ismertetett technikák és elméleti-gyakorlati megfontolások nemcsak a credit scoring területén hasznosak és alkalmazhatók, hanem sok más olyan adatbányászati probléma esetén is, amelyek hasonló mintaszelekciós mechanizmust tartalmaznak.

Irodalomjegyzék

Ash, D.–Meester, S. [2002]: Best Practices in Reject Interference, Presentation at Credit Risk Modeling and Decision Conference, Wharton Financial Institutions Center, Philadelphia, 2002. május

Banasik, J. B.–Crook, J. N.–Thomas, L. C. [2003]: Sample Selection Bias in Credit Scoring Models, *Journal of the Operational Research Society*, 54. szám, 822–832. o.

Boyes, W. J.–Hoffman, D. L.–Low, S. A. [1989]: An Econometric Analysis of the Bank Credit Scoring Problem, *Journal of Econometrics*, 40. évf., 3–14. o.

Caouette, J. B.–Altman, E. I.–Narayanan, P. [1998]: Managing Credit Risk: The Next Great Financial Challenge, John Wiley & Sons, New York

Chen, G.–Astebro, T. [2001]: The Economic Value of Reject Inference in Credit Scoring, presented at the Conference of Credit Scoring and Credit Control, Credit Research Centre, University of Edinburgh (EK), 2001. szeptember

Chen, G.–Astebro, T. [2003]: Bound and Collapse Bayesian Reject Inference When Data are Missing not at Random, in Astebro, T., Beling, P., Hand, D., Oliver, B. és Thomas, L. B. (eds.): Mathematical Approaches to Credit Risk Management, Conference Proceedings, Banff International Research Station for Mathematical Innovation and Discovery, 2003. október 11–16.

Chen, G.–Astebro, T. [2006]: A Maximum Likelihood Approach for Reject Inference in Credit Scoring, 2006. november 25., kézirat (available at SSRN: http://ssrn.com/abstract=872541

Crook, J.–Banasik, J. [2002]: Does Reject Inference Really Improve the Performance of Application Scoring Models? Credit Research Centre, Working Paper 02/3, University of Edinburg (EK)
Feelders, A. J. [1999]: Credit Scoring and Reject Inference with Mixture Models, *International Journal of Intelligent System in Accounting, Finance and Management*, 8. évf. 271–279.o.
Greene, W. H. [1992]: A Statistical Model for Credit Scoring, Working Paper EC–92–29, Leonard N. Stern School of Business, New York
Greene, W. H. [1998]: Sample Selection in Credit-scoring Models, *Japan and the World Economy*, 10. évf. 299–316. o.
Hajdu Ottó [2003]: Többváltozós statisztikai számítások, KSH, Budapest
Hand, D. J.–Henley, W. E. [1993/4]: Can Reject Inference Ever Work?, *IMA Journal of Mathematics Applied in Business & Industry*, 5. évf. 4. sz., 45–55. o.
Heckman, J. J. [1979]: Sample Selection Bias as a Specification Error, *Econometrica*, 47. évf. 153–161.o.
Hsia, D. C. [1978]: Credit Scoring and the Equal Credit Opportunity Act, *The Hastings Law Journal*, 30. évf. november, 371–448. o.
Jacobson, T.–Roszbach, K. [1998]: Bank Lending Policy, Credit Scoring and Value at Risk, SSE/EFI Working Paper Series in Economics and Finance 260, Stockholm School of Economics
Little, R. J. A.–Rubin, D. B. [1987]: Statistical Analysis with Missing Data, John Wiley & Sons, New York
Maddala, G. S. [1983]: Limited Dependent and Qualitative Variables in Econometrics, Cambridge University Press, Cambridge (EK)
McLachlan, G. J.–Basford, K. E. [1988]: Mixture Models, Inference and Applications to Clustering, Marker Dekker, New York
McLachlan, G. J [1992]: Discriminant Analysis and Statistical Pattern Recognition, Wiley & Sons, New York
McLeod, R.W. et al. [1993]: Predicting Credit Risk: A Neural Network Approach, *Journal of Retail Banking,* 15. évf. 3.szám, 37–40. o.
Meng, C. L.–Schmidt, P. [1985]: On the Cost of Partial Observation in the Bivariate Probit Model, *International Economic Review*, 26. évf. 1. szám, 1985. február, 71–85. o.,
Oravecz Beatrix [2007]: Credit scoring modellek és teljesítményük mérése, *Hitelintézeti Szemle*, 6. évf. 6. sz., 607–627. o.
Sebastiani, P.–Ramoni, M. [2000]: Bayesian Inference with Missing Data Using Bound and Collapse, *Journal of Computational and Graphical Statistics*, 9. évf. 4. szám, 779–800.o.
Thomas, L. C.–Edelman, D. B.–Crook, J. N. [2002]: Credit Scoring and Its Applications, Society for Industrial and Applied Mathematics, Philadelphia