Budapesti Corvinus Egyetem
Gazdálkodástudományi Doktori Iskola

Szelekciós torzítás és csökkentése az adósminősítési modelleknél

Ph.D. értekezés

Oravecz Beatrix

Budapest, 2008.

TARTALOMJEGYZÉK

Ábrák

Táblázatok

Bevezetés

Az utóbbi 15-20 évben forradalmi változás történt a pénzügyi szolgáltatások piacán. A bankok automatikus döntéshozói módszereket és döntéstámogatási modelleket kezdtek alkalmazni, hogy felgyorsíthassák a hitelengedélyezési döntéseket.
A hitelnyújtó és a hitelfelvevő között információs aszimmetria áll fenn. A bankok számára az egyik legnagyobb kockázat a *hitelezési kockázat*, amely annak a veszélyét fejezi ki, hogy a kölcsön adott tőkét és/vagy kamatait a hitelfelvevő részben vagy egészében nem fizeti vissza és emiatt a bankot veszteség éri. A bankoknak létérdeke, hogy minél több és jobb minőségű adatot szerezzenek az ügyfelekről, és ezekből különböző adatbányászati módszerek segítségével minél több információhoz jussanak az ügyfelek fizetési képességével és hajlandóságával kapcsolatban. Ezt a célt szolgálja az adósminősítéshez használt credit scoring is.
A credit scoringnak nagyon fontos szerepe volt a fogyasztói hitelek állományának robbanásszerű növekedésében. Egy pontos és automatizált kockázatelemző rendszer nélkül a bankok nem tudták volna ekkora ütemben növelni lakossági kihelyezéseiket.

A credit scoring módszerek széleskörű 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. 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 modell elveszíti prediktív erejét. Másrészt viszont, ha csak a befogadott ügyfelek adatait használják a modell frissítéséhez, akkor megkérdőjelezhető lesz az új 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.*

A dilemmára az elutasítottak jellemzőinek felhasználásával történő modellépítés (*reject inference*) jelenthet választ. Ez tulajdonképpen annak becslése, hogy hogyan viselkedett volna az elutasított kérelmező, ha megkapta volna a hitelt.
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.

A dolgozatban *a credit scoring modelleknél fellépő szelekciós torzítás csökkentésé*re használható módszerekkel foglalkozunk. A jelenség vizsgálata a magyarnyelvű szakirodalomból szinte teljesen hiányzik, csak említés szintjén találkozhattunk vele.
A témaválasztást elméleti érdekességén túl *gyakorlati jelentőssége* is indokolta. Hiszen ha csak egy kicsit is sikerül javítani a modellek teljesítményén, az óriási profitnövekedést, és/vagy kockázatcsökkenést eredményezhet *a bankok* számára, mivel nagy volumenű kihelyezésekről van szó. A kockázatok pontosabb értékelése ugyanakkor *az ügyfelek számára* is előnyös, mert a jó adósok számára a kockázati felár csökkentését teszi lehetővé, vagy megfelelő kockázati felárral olyanok is kaphatnak hitelt, akiket eddig elutasítottak.

A dolgozat felépítése:
Az adósminősítési modelleknél fellépő szelekciós torzítás adathiányból eredő probléma, hiszen a korábban elutasított banki ügyfelek esetén a hitelkockázatot (hitelvisszafizetést) leíró eredményváltozó értéke hiányzik (nem megfigyelhető), ezért az **I.** részben ***a hiányzó adatok*** típusait és kezelésük lehetséges módjait vesszük sorra.

A következő részben (**II.**) röviden áttekintjük a ***credit scoring*** feladatát és a leggyakrabban alkalmazott módszereket, valamint az ezek értékeléséhez használható mérőszámokat.

A **III.** részben ismertetjük a szakirodalomban fellelhető ***módszereket, amelyek a scoring modelleknél fellépő szelekciós torzítás csökkentését szolgálják***. Mindegyik módszer valamilyen módon felhasználja az elutasítottakról meglévő információkat.

Az elutasítottak tényleges visszafizetési adatát nem ismerjük, ezért – mivel a semmiből nem keletkezhet új információ-, ha fel akarjuk használni őket a modellépítéshez, akkor vagy *feltételezések*kel kell élnünk, vagy *pótlólagos*an *információ*t kell szerezni a visszafizetési viselkedésükről.

Ebben a részben bemutatjuk ezen (reject inference) technikák elméleti hátterét, kiemelve az alkalmazott feltételezéseket vagy a pótlólagos információ szerzésének és felhasználásának módját és összegezzük 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 mennyivel nagyobb. Az elutasítottak tényleges és imputált adatainak alkalmazásának haszna függ az elutasítási aránytól, a mintabeli és sokasági eloszlásoktól és az alkalmazott statisztikai feltételek teljesülésétől. Van néhány portfolió, ahol nagyon alacsony az elutasítottak aránya (ilyen például a jelzáloghitelek piaca). Ilyen esetekben felesleges lehet az elutasítottakkal foglalkozni, hiszen elhanyagolható az arányuk a populáción belül, így az általuk okozott torzítás sem igényel korrekciót. Másrészt a nagyobb kockázatú portfoliók esetén, 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 adatbázis-függő.

Néhány statisztikus szerint az elutasítottak hiányzó bedőlési adatainak megfelelő imputációjával megoldható a nem véletlen mintából való következtetés problémája (Joanes 1993/4, Donald 1995, Copas és Li 1997, Greene 1998). Külföldön a scorecard fejlesztők már alkalmaznak reject inference technikákat, amelyben statisztikai szoftver

csomagok (például SAS) is segítik őket. Ezek azonban sokszor fekete dobozként üzemelnek, mert a mögöttük lévő alapelvek és feltételezések nem világosak a felhasználók számára.

Ha elfogadhatónak tartunk bizonyos feltételezéseket és valamilyen imputációs eljárással felhasználjuk az elutasítottakat, akkor felmerül a kérdés, hogy hogyan validálhatjuk a modellünket és hogyan mérhetjük az általa elért javulást. Ebben a témában kevés releváns tanulmány született, mert a tesztelésre használt adatbázisok többsége nem teljes, vagy szimulált volt (Donald 1995, Feelders 1999, Manning et al. 1987).

Hand és Henley (1993/4) megmutatták, hogy az üzleti életben alkalmazott megoldások problematikusak, mert általában igen kétséges feltételezéseken alapulnak.

Az alkalmazott feltételek teljesülése azonban általánosságban nem tesztelhető, így - a szakirodalom áttanulmányozása után - arra a következtetésre jutottam, mint ők: *a torzítás csökkentésének egyetlen robusztus és megbízható módja, ha az elutasítottak egy részét ténylegesen meghitelezik és így figyelik meg viselkedésüket és 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 csökkentésének egy lehetséges módja a *résnyire nyitott kapu* alkalmazása egyfajta költségoptimális mintaelosztással.

Ez azt jelenti, hogy 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 kisebb. Így egy rétegzett mintát kapunk egyfajta költségoptimális mintaelosztással. Végül átsúlyozással kaphatunk egy a sokaságot valóban reprezentáló mintát anélkül, hogy vállalni kellett volna a mindenki beengedésével járó hatalmas költségeket.

Az utolsó (**IV.**) részben ***empirikus kutatás*** keretében egy valós banki adatbázison (lakossági hitelkártya adatokon) vizsgáljuk az ezzel a módszerrel elérhető javulást, annak költségeit és várható hasznait.

A vizsgált hipotézisek:

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

Az empirikus kutatás során azt tapasztaltuk, hogy *magas elutasítási arány (erőteljes és nem teljesen véletlenszerű szelekció) esetén gyengébb teljesítményű modellek építhetők,* mint kisebb arányú elutasítás esetén. Ennek egyik oka, hogy ekkor kevés rossz ügyfél kerül a portfolióba, 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 információ szerzés egyik módja, 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 terméken elérhető profit is nőtt.*

Azt tapasztaltuk, hogy ha a profitmaximalizálás a cél, akkor *jobb, ha az elméleti úton meghatározott cutoff 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, azoktól csak kicsit rosszabbnak tűnő ügyfelekből érdemes résnyire nyitott kapuval beengedni még ügyfeleket.*
Ez az elsőlépcsős nagymértékű modelljavulás és profitnövekedés valószínű 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 plusz minta költségei tervezhetőbbek, és kisebbek is, mintha egy távoli tartományból vennénk mintát.

A dolgozatban ismertetett technikák és elméleti -, gyakorlati megfontolások nem csak a banki adósminősítés területén hasznosak és alkalmazhatók, hanem sok más olyan adatbányászati probléma esetén is, amelyek hasonló mintaszelekciós mechanizmust tartalmaznak.

Köszönettel tartozom mindazoknak, akik segítettek abban, hogy elkészüljön ez a munka.
Köszönöm az értekezéstervezet bírálóinak, Herman Sándornak és Vita Lászlónak, hogy értékes észrevételeikkel és hasznos tanácsaikkal segítették a dolgozat jelen formájának kialakítását.
Kiemelten szeretnék köszönetet mondani témavezetőmnek, Hunyadi Lászlónak, hogy az elmúlt években tanácsaival segített, ösztönzött és bátorított, és idejét nem sajnálva mindig kész volt olvasni és megvitatni félkész gondolataimat.
Külön köszönöm családomnak a türelmet és a figyelmet, amellyel munkámat támogatták.

I. Hiányzó adatok és kezelésük a statisztikai elemzésekben

Egy általános adatmátrix sorai tartalmazzák a megfigyelési egységeket, vagy eseteket, az oszlopok pedig a változókat, amelyek értékét minden egység esetén ismerjük. Az adatmátrixban lévő adatok általában valós számok, amelyek vagy egy mennyiségi ismérv tényleges értékeit fejezik ki (például az életkor vagy a jövedelem), vagy egy minőségi ismérv kategóriáit reprezentálják (például az iskolai végzettség vagy a nem). A gyakorlatban azonban az a jellemző, hogy ez az adatmátrix nem teljes, bizonyos értékek hiányoznak.

Például egy háztartási bevételeket és kiadásokat vizsgáló kutatás során a megkérdezettek megtagadhatják a jövedelemre vonatkozó kérdés megválaszolását, vagy egy fogyasztói preferenciákat vizsgáló kutatás során előfordulhat, hogy a válaszadó nem tud választani két termék közül, egyiket sem preferálja a másikkal szemben. Az első esetben a jövedelem értékét tekinthetjük hiányzónak, hiszen van mögötte egy tényleges érték, csak mi nem ismerjük. A második esetben azonban nem tekinthetjük a termékpreferenciát hiányzónak, mert nincs mögötte valós érték, a válaszadó nem megtagadta a választ, hanem nem tudott válaszolni. Ebben az esetben a „nincs preferencia" vagy „nem tudom" is egy válaszadói réteget jelöl. A legtöbb statisztikai szoftver tartalmaz egy vagy több speciális kódot az adathiány bevitelére. Egynél több kód lehetővé teszi a különböző jellegű adathiányok beazonosítását, mint „nem tudja", „válaszmegtagadás", „értelmetlen adat".

Felmerülhet a kérdés, miért kell egyáltalán a hiányzó adatokkal foglalkozni, ahelyett, hogy egyszerűen törölnénk őket a mintából. Válaszként álljon itt a következő (nem a banki hitelezés, hanem egészen más területről származó) példa:

1992. április 9-én a Konzervatív Párt megnyerte a brit választásokat, ami óriási bukást jelentett a közvéleménykutatási iparágnak. A választások napján a négy legnagyobb kövéleménykutató cég a Munkáspárt 0,9% pontos győzelmét várta. Ezzel szemben a Konzervatív Párt győzött 7,6% ponttal. Ez nagyon nagy 8,5% pontos hiba. Egy utólagos vizsgálat megállapította, hogy a hiba fő oka az volt, hogy a kutatás során nem foglalkoztak a válaszmegtagadásokkal és a „még nem tudom" típusú válaszokkal,

hanem egyszerűen törölték őket a mintából. Ez a gyakorlat végzetes volt az eredmények szempontjából, mert az utólagos kutatás megmutatta, hogy a konzervatív pártiak kevésbé tárták fel választási szándékukat. /Hasonló volt a helyzet a magyarországi 2002-es választások során is./

Az adathiány torzító hatása

Az adathiány mindennapos probléma. Az 1% vagy ez alatti adathiány ráta triviális, az 1-5% közötti kezelhető. Az 5-15% közötti adathiány kezelése már szofisztikált módszerek használatát igényli. 15% feletti adathiány pedig már súlyos interpretálási problémákat vet fel. (McDermit, 1999)

Ezt a problémát világítja meg az a szimuláció, amely a pártpreferenciákra való kérdésre adott válaszok terén a válaszolók és a nemválaszolók közötti szisztematikus eltérés hatását vizsgálja. (Rudas, 1998)

Az alábbi táblázat 1000 fős mintán mutatja meg, hogy a minta sikertelen lekérdezése esetén a meghiúsulás mértékének, illetve a válaszadók és a válaszmegtagadók véleményének eltérésének függvényében az eredeti megbízhatósági szint hogyan módosul:

Eltérés a válaszadók és a nemválaszolók tényleges értékei között	Válaszadási arány			
	100%	90%	80%	70%
0%	95%	95%	93%	91%
5%	95%	93%	88%	81%
10%	95%	90%	75%	54%
15%	95%	84%	54%	25%
20%	95%	76%	33%	7%

1. táblázat Nemvéletlen adathiány hatása a megbízhatóságra (Forrás: Rudas, 1998)

Ha például egy 70%-os válaszadási arányhoz, 20%-os eltérés figyelhető meg a válaszolók és a nem-válaszolók között, akkor a megfigyelt esetekből számított becslésünk megbízhatósága 95%-ról 7%-ra (!) esik.

A fentiekből is látható, hogy a hiányos adatbázisokból való következtetések torz képet adhatnak. Törekedni kell tehát az adathiány természetének megismerésére, majd ezen információk figyelembevételével a hiányzó adatok valamilyen kezelésére.

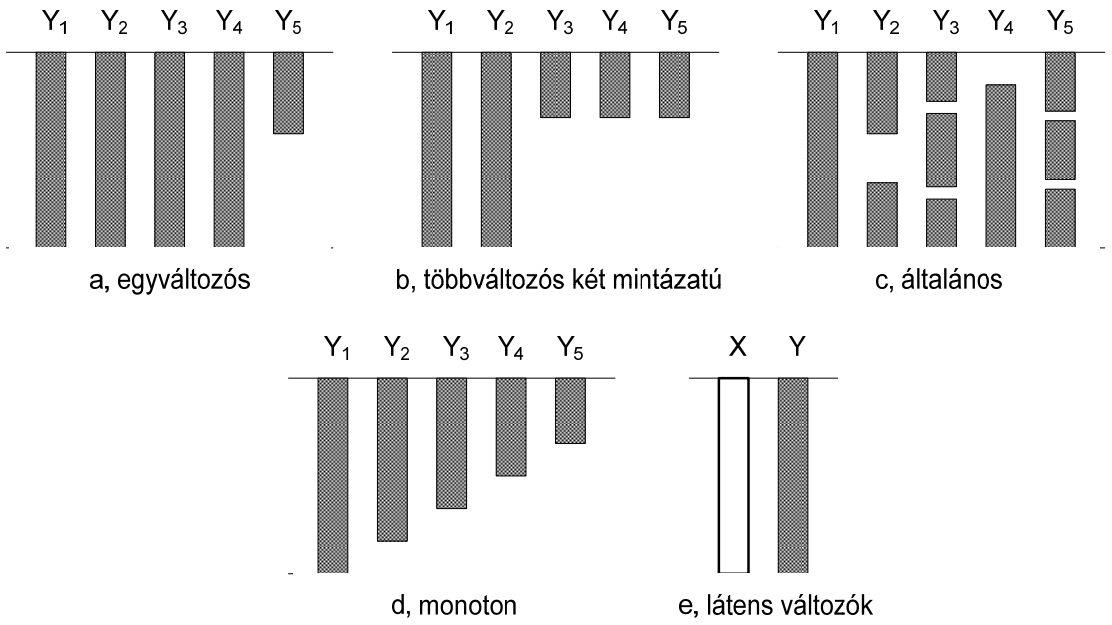
A hiányzó adatok sok kutatásnál jelentenek komoly problémát, mert a minta véletlenszerűségét rombolhatják le, pedig a legtöbb statisztikai módszer és következtetés alapja a véletlen minta. Ebben a fejezetben röviden áttekintjük a hiányzó adatok típusait és a kezelésükre használt legelterjedtebb módszereket, kiemelve fő előnyeiket és hátrányaikat.

1. Hiányzó adatok típusai

Az alábbiakban áttekintjük az adathiány típusait. A csoportosítás egyik szempontja az *adathiány mintázata*. A mintázat az írja le, hogy mely adatok a megfigyeltek és mely adatok hiányoznak az adatmátrixban. A másik csoportosítási szempont az *adathiány mechanizmus*, amely a hiányzás és az adatbázisban szereplő változók értékei közötti kapcsolatot veszi figyelembe.

1.1. Adathiány mintázat

Legyen $Y = (y_{ij})$ egy (n x K) általános adatmátrix, hiányzó adatok nélkül, amelynek i -dik sora $y_i = (y_{i1}, \ldots, y_{iK})$, ahol y_{ij} az Y_j változó értéke az i –dik egységnél. Hiányzó adatok esetén legyen $M = (m_{ij})$ az adathiány indikátor mátrix (Little és Rubin, 2002), ahol $m_{ij} = 1$, ha y_{ij} hiányzik és $m_{ij} = 0$, ha y_{ij} megfigyelt. Az M mátrix definiálja az adathiány mintázatot. Az 1. ábra mutat néhány példát az adathiány mintázatokra.(A megfigyelt y-ok sötéttel jelölve.)



1. ábra Adathiány mintázatok

Egyváltozós adathiány

Az 1a) ábra jelzi azt az esetet, amikor csak egyetlen változóban van adathiány, a többi változó teljesen megfigyelt. Ilyen mintázata lehet például a mezőgazdasági kontrollált kísérletek eredményének, ahol azt vizsgálhatják, hogy milyen a kapcsolat az Y_K eredményváltozó (terméshozam) és az $Y_1,\ldots,Y_{K-1}$ magyarázó változók (öntözővíz, hőmérséklet, műtrágya típusa, mennyisége) között. A magyarázó változók ekkor teljesen megfigyeltek, nincs hiányzó adat, a függő változóban viszont előfordulhat adathiány (például hibás vetőmag vagy rossz adatrögzítés miatt).

A dolgozatban is ezzel a típusú adathiánnyal fogunk foglalkozni, mert feltételezzük majd, hogy a hitelkérelmezők minden jellemzőjét ismerjük, amit a kérelemben megadtak, a hitelkockázatot leíró változót viszont csak azoknál az ügyfeleknél, akik korábban már kaptak hitelt, ezen eredményváltozó értéke tehát a korábban elutasított ügyfeleknél hiányzik.

Többváltozós kétmintázatú

Egy másik általános mintázat, amikor az előző példában szereplő egyetlen adathiányos változó (Y_K) helyett több adathiányos változónk van $(Y_{j+1},\ldots,Y_K)$, ahol mindegyik egyformán megfigyelt, vagy hiányzik ugyanazokra az esetekre. (Lásd az 1b) ábrát, ahol $K = 5$ és $J = 2$.)

Erre a mintázatra lehet példa a kérdőíves felméréseknél az egység szintű nemválaszolás. /Amennyiben az adathalmazból egy-egy elem teljesen hiányzik teljes (vagy egység szintű) nemválaszolásról (unit nonresponse) beszélünk./ Ez az egység szintű nemválaszolás előfordulhat azért, mert a kiküldött kérdőívet meg sem kapta a címzett, vagy megkapta, de megtagadta a válaszadást. Ekkor a kérdőívben szereplő változók lesznek az adathiányos változók. A teljes, adathiányt nem tartalmazó változók a minta tervezéséhez használt változók lesznek, amelyek mind a válaszadók, mind a nemválaszolók esetében előzetesen ismertek egy listáról (például név→nem, lakcím).

Általános mintázat

Ha csak bizonyos kérdésekre adott válaszok hiányoznak, akkor részleges (vagy tétel szintű) nemválaszolásról (item nonresponse) beszélünk. Ekkor az adathiány mintázata általában semmiféle specialitással nem rendelkezik. (Lásd 1c) ábra.)

Monoton adathiány

A longitudinális felmérések időről időre gyűjtenek be adatokat ugyanazon megfigyelési egységekről. Ezekben a felmérésekben gyakori jelenség a lemorzsolódás, ami azt jelenti, hogy a megfigyelési egység kiesik a mintából, még a kutatás befejezése előtt. Például háztartás panel esetén a család külföldre költözik, vagy klinikai kísérleteknél más gyógyszerek hatása, vagy egyéb betegség miatt a beteg nem tud tovább részt venni a kísérletekben. A lemorzsolódás egy példája a monoton mintázatú adathiányoknak. (Lásd 1d) ábra.) Ekkor a változókat lehet úgy sorba rendezni, hogy minden $Y_{j+1},\ldots,Y_K$ hiányzik, ha Y_j hiányzik. Vannak olyan módszerek, amelyek csak az ilyen mintázatú adathiányt tudják kezelni. Az ilyen mintázat a gyakorlatban ritkán fordul elő, közel monoton mintázat azonban már gyakrabban.

Látens változók

A nem megfigyelhető látens változókat is felfoghatjuk adathiány problémaként, csak ezeknél a látens változóknál speciálisan minden megfigyelési érték hiányzik. Az 1e) ábrán az X jelenti a látens változók csoportját, ahol minden érték hiányzik és Y pedig a teljesen megfigyelt változók csoportját. Ekkor természetesen bármiféle elemzéshez különböző feltételezésekkel kell élnünk. Látens változó lehet például a klinikai kísérleteknél a beteg gyógyulásba vetett hite, ha erre vonatkozóan nem szerepelnek adatok a mintában.

1.2. Adathiány mechanizmus

A hiányzó adatok kezelésének legalkalmasabb módját akkor tudjuk megtalálni, ha ismerjük, hogy miként lettek hiányzóak. 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 mechanizmus*nak nevezték el.

Intuitíve és formálisan is megadjuk az egyes csoportok definícióját. Legyen továbbra is az $Y = (y_{ij})$ a teljes adatmátrix és az $M = (m_{ij})$ az adathiány indikátor mátrix. Az adathiány mechanizmus jellemezhető az M adott Y melletti feltételes eloszlásával, az $f(M|Y,\theta)$-val, ahol θ ismeretlen paramétereket jelöl.

Teljesen véletlenszerű adathiány **(Missing Completely at Random (MCAR))**

A teljes adatállománnyal rendelkező egységek és a hiányzó adatokat tartalmazó egységek teljesen egyformák, ugyanazon eloszlásból származnak.

A hiányzás tehát nem függ az Y értékétől, sem a megfigyelt, sem a hiányzó adatokkal rendelkező változók értékétől, azaz:

$$f(M|Y, \theta) = f(M| \theta), \text{ minden } Y, \theta \text{ esetén.} \qquad (1)$$

Ez a mechanizmus például akkor fordulhat elő, ha minden válaszadó egy pénzérme feldobásával dönti el, hogy válaszol-e a kérdésre.

Véletlenszerű adathiány **(Missing at Random (MAR))**

A hiányzó adatokat tartalmazó egységek eltérnek a hiánytalan adatokkal bíró egységektől, de a hiány jellegzetességei nyomon követhetők, előre jelezhetők az adatbázis más változói segítségével. Az adathiány tehát más változókkal kapcsolatban van, de azzal a változóval, amelyikben a hiányzás felmerül nincs közvetlen kapcsolatban.

Legyen $Y_{megfigyelt}$ az a része Y-nak amelyben nincs adathiány és $Y_{hiányzó}$ az a rész, amelyben van adathiány. A véletlenszerű adathiány tehát az jelenti, hogy:

$$f(M|Y, \theta) = f(M|\ Y_{megfigyelt}, \theta), \text{ minden } Y_{hiányzó}, \theta \text{ esetén.} \qquad (2)$$

Ez a mechanizmus fordul elő például, ha a magasabb jövedelemmel rendelkezők nagyobb valószínűséggel tagadják meg a jövedelemre vonatkozó kérdések megválaszolását, de a jövedelemre következtetni tudunk a felmérés más változói (például: fogyasztási szokások, fogyasztás és megtakarítás egymáshoz való viszonya) alapján.

Nem véletlenszerű adathiány (Not Missing at Random[1] (NMAR))

Az adathiány nem véletlenszerű, és más változókkal sem becsülhető, mert közvetlenül az adathiányt tartalmazó változóval van kapcsolatban. Az *M* eloszlása tehát függ az *Y* hiányzó értékeitől (is). Ez az adathiány legveszélyesebb, legnehezebben kezelhető formája.

Ez a mechanizmus fordul elő például, ha a magasabb jövedelemmel rendelkezők nagyobb valószínűséggel tagadják meg a jövedelemre vonatkozó kérdések megválaszolását, és a jövedelemre nem tudunk következtetni a felmérés más változóiból.

A hiányzó adatok számos problémát okoznak. Ugyanazon az adatbázison különböző kutatók által végzett elemzések eredménye között inkonzisztenciát tapasztalhatunk, ha azok másképpen kezelték a hiányzó adatokat. A hiányzó adatok kezelésére pedig azért van szükség, mert a sokasági paraméterbecslések torzítottak lehetnek (mint ahogy az 1992-es brit választásoknál is történt), hacsak az adathiány nem teljesen véletlenszerű. A hiányzó adatok kezelésének célja éppen ennek a torzításnak az eltüntetése. Ezt a célt a különböző módszerek annak függvényében érik el, hogy mennyire helyesen sikerül azonosítani és modellezni az adathiány sajátosságait.

2. Hiányzó adatok kezelésére szolgáló módszerek

A hiányzó adatokkal való elemzés irodalma nem túl hosszú múltra tekint vissza. A szakirodalomban ajánlott és alkalmazott módszereket a következőképpen csoportosíthatjuk (Little és Rubin, 2002):

[1] Vagy másként: “nonignorable”.

1. Teljesen megfigyelt vagy elérhető egységek elemzésén alapuló eljárások
2. Átsúlyozás
3. Imputáció alapú eljárások
4. Modell alapú eljárások

A csoportok nem átfedésmentesek, de ebben a csoportosításban tekintjük át az alábbiakban a nemválaszolások kezelésének legelterjedtebb módszereit. A felsorolás nem tartalmaz minden alkalmazható módszert, csak a széles körben használt megközelítéseket.

2.1 Teljesen megfigyelt vagy elérhető egységek elemzésén alapuló eljárások

Adathiányt tartalmazó esetek törlése (Listwise vagy casewise adat törlés)

Ha egy megfigyelési egységnél akár csak egy változó tekintetében is hiányzik adat, az egész megfigyelést törlik az adatbázisból. Az eljárást számos statisztikai programcsomag tartalmazza alap megoldásként. A megoldás előnye az egyszerűsége, és hogy az egyváltozós statisztikák összehasonlíthatóak, mert mindegyik ugyanazon adatokon lett számítva. Hátránya viszont, hogy a nem teljes megfigyelésekben meglévő információt egyáltalán nem hasznosítja. Csak teljesen véletlenszerű eredetű adathiány esetén alkalmazható, azaz ha a hiányzó adatokat tartalmazó esetek az összes eseten belüli véletlenszerű almintának tekinthetők. Ha az adathiány nem MCAR, akkor a módszer torzítást okoz. Relatíve alacsony nemválaszolási arány mellett ésszerű lehet az alkalmazása, mert ekkor az egyszerűségből fakadó előnyök ellensúlyozhatják a néhány hiányzó adat által okozott információveszteséget és minimális torzítást.

Ha a banki hitelezésben a scoringfüggvény újjáépítését a meghitelezett ügyfelek adatain végzik és semmilyen módon nem veszik figyelembe az elutasítottakat, akkor tulajdonképpen ezt a módszert használják.

Elérhető adatok elemzése (available case analysis)

Az adatok törléséből származó információveszteség csökkenthető, ha minden változó elemzésekor az abban a változóban meglévő összes adatot használjuk. A módszer hátránya, hogy minden elemzés más adatbázison készül, így az eredmények összehasonlítása problémás lehet.

E módszer alkalmazásakor kétváltozós korreláció vagy kovariancia számításhoz mindig az adott két változó tekintetében elérhető adatpárokat használják (*Pairwise available case*). Számos statisztikai programcsomagban megtalálható. Előnye, hogy jobban kihasználja a meglévő adatokat, de az eredményeként létrejövő korrelációs mátrix nem feltétlen lesz pozitív definit. Márpedig a korrelációs mátrixon alapuló többváltozós regresszióhoz is szükséges a pozitív definit mátrix.

Nézzük a következő példát (Little és Rubin, 2002), ami három változóra vonatkozóan 12 megfigyelést tartalmaz, a „?" hiányzó adatot jelent:

Y_1	1	2	3	4	1	2	3	4	?	?	?	?
Y_2	1	2	3	4	?	?	?	?	1	2	3	4
Y_3	?	?	?	?	1	2	3	4	4	3	2	1

Ebben a mintában az elérhető adatpárokat használva a mintából számított korrelációs együtthatók $r_{12} = 1$, $r_{13} = 1$, $r_{23} = -1$. Ezek a becslések nem jók, mert a sokasági korrelációs együtthatóknál $\rho_{12} = \rho_{13} = 1$ –ből az következik, hogy $\rho_{23} = 1$, nem -1.

Mivel az elérhető adatokat használó módszerek több információra támaszkodnak, azt várnánk, hogy hatékonyabbak, mint a csak a teljes adatokat használók. Kim és Curry (1977) is ezt találták MCAR és gyenge korreláció esetén. Erősebb korreláció esetén viszont a teljes adatokat használó módszerek bizonyultak jobbnak (Azen és Van Guilder, 1981).

2.2. Átsúlyozás

Az átsúlyozásos módszerek abból indulnak ki, hogy válaszmegtagadás esetén a válaszmegtagadó elemhez hasonló nem adathiányos esetek (vele azonos kategóriában vagy rétegben szereplő elemek) arányosan több sokasági elemet képviselnek, azaz

nagyobb súlyt kell kapjanak. Általában, ha a *j*-dik alcsoportban (kategóriában) a válaszadók aránya p_j, akkor az itt szereplő elemek $\frac{1}{p_j}$ súlyt kapnak, azaz itt mindegyik elem ennyiszer több sokasági elemet képvisel.

Véletlen mintákból való következtetésnél, amikor az elemek kiválasztása nem azonos valószínűséggel történik, gyakran súlyozzák a megfigyelési elemeket a tartalmazási valószínűségük[2] inverzével (Hunyadi, 2001). Legyen például y_i az *Y* változó értéke az *i*-dik megfigyelési egységre. Ekkor, ha nincs hiányzó adat a sokasági átlag Horvitz – Thompson becslőfüggvénye:

$$\hat{\bar{Y}}_{HT} = \frac{\sum_{i=1}^{n} \frac{y_i}{\pi_i}}{\sum_{i=1}^{n} \frac{1}{\pi_i}} \qquad (3)$$

ahol π_i az *i*-dik egység ismert tartalmazási valószínűsége, a szumma pedig a megkérdezettekre vonatkozik.

Hiányzó adatok esetén az átsúlyozás úgy módosítja a súlyokat, mintha a nemválaszolás is a mintavételi terv része lett volna, ekkor a fenti becslőfüggvény a következőképpen módosul:

$$\hat{\bar{Y}}_{HTm} = \frac{\sum_{i=1}^{n} \frac{y_i}{\pi_i \hat{p}_i}}{\sum_{i=1}^{n} \frac{1}{\pi_i \hat{p}_i}} \qquad (4)$$

Itt a szumma nem a megkérdezettekre, hanem a ténylegesen válaszolókra vonatkozik, a $\hat{p}_i$ pedig az *i*-dik egység becsült válaszadási valószínűsége (általában a válaszadási arány a minta egy alcsoportjában).

A módszer alapelve tehát egyszerű, de többdimenziós feladatoknál már bonyolultabb lehet a kivitelezése. Ráadásul a túlságosan szóródó súlyok nagy korrekciót jelentenek, ami megnöveli a feltételezések szerepét a becslésekben (Hunyadi, 2001).

Az átsúlyozás mögött az a feltételezés húzódik meg, hogy az adott rétegen belül a válaszadók a megkérdezettek véletlen almintájának tekinthetők, azaz a rétegen belül az adathiány MCAR jellegű.

Az átsúlyozott mintából sokszor relatíve egyszerű a sokasági paraméterek pontbecsléseit elkészíteni. Az intervallumbecslésekhez szükséges standard hibák

[2] probability of inclusion, azaz a minták hány százaléka tartalmazza az adott elemet.

számítása már korántsem ilyen egyszerű. A statisztikai programcsomagok lehetővé teszik aszimptotikus standard hibák számítását összetettebb mintavételi tervek esetén, beleértve az átsúlyozást, rétegzést is. Ezek a programok azonban tipikusan fixnek, ismertnek tartják a súlyokat, pedig adathiány esetén a válaszadási aránnyal arányos súlyok maguk is mintavételi ingadozásnak vannak kitéve.
Egyszerű véletlen mintára vannak képletek a hibaszámításhoz, komplexebb esetekhez azonban a minta mesterséges újrahasznosításán alapuló számítógép-intenzív módszerek (jackknife, bootstrap, kiegyensúlyozott ismétlések) alkalmazására van szükség.

2.3. Imputáció alapú eljárások

Az imputáció azt jelenti, hogy a hiányzó adatot utólag mesterségesen pótolják egy ahhoz vélhetően hasonló értékkel. Ezután az így létrejött „teljes" adatbázison elvégezhetők a standard statisztikai elemzések. A helyes következtetéshez azonban módosítani kell a standard elemzéseket, valahogyan meg kell különböztetni a valódi és az imputált értékeket, hiszen ez utóbbiak újabb bizonytalansági faktort jelentenek. Ezt a bizonytalansági tényezőt építi be a modellbe például a többszörös imputáció (multiple imputation).

***Logikai imputáció* (data editing)**

Ha a hiányzó értékek más adatokból, vagy korábbi felvételekből logikailag következnek, akkor azokkal pótolják őket. A nem például nem változik, és a hiányzó életkorra is következtethetünk, ha egy korábbi felmérésnél megadták. A módszer előnye, hogy nem csökkenti az adatokban lévő tényleges változékonyságot.

Átlaggal való pótlás

Az adott változóban meglévő adatok átlagával[3] helyettesítik a hiányzó értékeket. Az átlaggal való imputálás előnye az egyszerűsége, és könnyű alkalmazhatósága. Hátránya viszont, hogy bár teljesen véletlenszerű adathiány esetén várható érték

[3] Átlag helyett más középérték is használható (módusz, medián).

szempontjából nem torzít, az elemek változékonyságát alulbecsli. Ez javítható, ha a megfigyeléseket homogénebb csoportokra bontjuk és csoportokon belüli részátlagokkal imputálunk, de a standard hibákat és a becslések konfidencia intervallumát még így is alulbecsüljük. Sőt a szabadságfoknál is figyelembe kell venni, hogy az imputált átlagok nem függetlenek a többi megfigyelés értékétől.
Ez a módszer tulajdonképpen az átsúlyozással azonos eredményt ad.

Regressziós módszerek

A teljes megfigyeléseken építenek egy regressziót a hiányzó értéket tartalmazó változót eredményváltozóként, a többit magyarázóváltozóként kezelve. Aztán azokra az esetekre, ahol az eredményváltozó értéke hiányzik, a regresszió segítségével becslést készítenek. A módszer továbbfejlesztéseként a *sztochasztikus regressziós imputálások* esetén egy véletlen változót is adnak a becslésekhez, mert e nélkül a változók közötti kapcsolat a későbbi elemzésekben szorosabbnak mutatkozna, mint amilyen valójában lehet.

Hot deck imputáció

A hiányzó adatot tartalmazó megfigyeléshez leginkább hasonló hiánymentes esetet megkeresik és ennek Y értékével pótolják a hiányos eset hiányzó Y értékét. A hasonlóság mértékének megítélésére különböző módszerek használhatók. A hot deck módszer előnye a fogalmi egyszerűsége mellett, hogy megőrzi a változók eredeti mérési szintjét (a kategóriás kimenetelű változók kategóriások maradnak, a folytonosak pedig folytonosak). A módszer hátránya, hogy nehéz az esetek hasonlóságát definiálni és az elemzőnek esetleg saját programot kell készítenie a donor egységek kiválasztásához. Ezenkívül a standard hibák számítása is nehézségekbe ütközhet (Roth, Switzer, 1995). A nehézségek ellenére a hot deck imputáció igen népszerű technika, számos hivatalos statisztikai felmérésben is ezt a módszert alkalmazták[4]. Vannak modellek, amelyek több hasonló esetet keresnek és azokból véletlenszerűen választják ki a donor megfigyelést, vagy, ha az megfelelő, az átlagukat számítják az imputációhoz.

[4] Például: British Census (Baker, Harris, O'Brien, 1989), Statistics Canada (Rubin, 1987)

Hot deck (belső) módszereken sokszor tágabb értelemben az olyan adatpótlást értik, amely csak az adott mintát használja az imputációhoz, cold-deck (külső) módszerek esetén pedig más, külső forrásokat[5] is felhasználnak.

Közelítő Bayes-i bootstrap (Approximate Bayesian Bootstrap (ABB))

A módszer logisztikus regressziót alkalmaz, hogy az Y függő változóban a válaszolás / nemválaszolás valószínűségét becsülje az X_i változók segítségével.[6] A megfigyelési egységek az így kapott hiányzás hajlamossági score-ok alapján képzett kvantilisekbe csoportosíthatók. A csoportokon belül a nem hiányos esetekből visszatevéses mintavétel segítségével lehet imputálni a hiányzó értékeket. Az eljárás minden hiányzó adatot tartalmazó változóra megismétlődik. A módszer a hot deck imputáció egy formája, ahol a hasonlóságot a hiányzás hajlamossági score-ok határozzák meg.

Léteznek úgynevezett *kompozit módszerek* (composite methods) is, amelyek különböző módszerek alapelemeit ötvözik. Például a hot deck és a regressziós imputáció keveréke, amely először regresszióval számítja a becsült átlagokat, majd ezekhez hozzáadja egy véletlenszerűen kiválasztott empirikus reziduum értékét.

A nemválaszolás miatti bizonytalanság pótlólagos varianciaforrást jelent, amit valahogyan be kell építeni a becslésekbe. Ez megoldható például a minták másodlagos hasznosításán alapuló számítógép intenzív módszerek alkalmazásával, amelyekkel bonyolult mintavételi terv és imputációs technika esetén is becsülhető a becslőfüggvények varianciája.
Több imputált adatbázis létrehozásával és azok eredményeinek összesítésével szintén beépíthető az adathiány okozta pótlólagos bizonytalanság a rendszerbe.
A következőkben ezt a többszörös imputációt tekintjük át.

[5] Az adott mintához képest külső, például múltbeli hasonló felmérések adatai.

[6] Ilyen logisztikus regressziós módszert alkalmaz György Erika (2004) a munkaerő felvételben szereplő nemválaszolás kezelésére.

Többszörös imputáció **(Multiple Imputation (MI))**

Többszörös imputáció esetén minden hiányzó elem helyére több lehetséges értéket imputálnak, ezáltal több (általában 3-10) „teljes" adatbázist készítenek az eredeti hiányos adatbázisból. Az elemző mindegyik adatbázison elvégzi a megfelelő statisztikai módszerekkel a kívánt elemzéseket, a kapott eredményeket összegyűjti és összekombinálja egyetlen elemzésbe. Ez utóbbi sokszor nem egyszerű feladat. A többszörös imputáció egy lépéssel tovább megy, azzal, hogy bevezeti a statisztikai bizonytalanságot a modellbe, azért, hogy egy teljes adatbázisban meglévő változékonyságot közelítse az imputációval teljessé tett adatbázis is.

A többszörös imputációt először Rubin (1978) javasolta a hiányzó adatok kezelésére.

Furcsának tűnhet, hogy viszonylag kevés (3-10) imputációval is érzékeltetni lehet a pótlások bizonytalanságát. Rubin (1987) megmutatta, hogy *m* imputáción alapuló becslés relatív hatékonysága végtelen számú imputáció hatékonyságához képest nagyjából

$\left(1+\frac{\gamma}{m}\right)^{-1}$, ahol γ a hiányzó információk aránya (számítását lásd később). Az *m* és γ különböző értékei mellett elérhető hatékonyságokat mutatja az alábbi táblázat:

	γ				
m	0.1	0.3	0.5	0.7	0.9
3	97	91	86	81	77
5	98	94	91	88	85
10	99	97	95	93	92
20	100	99	98	97	96

2. táblázat Többszörös imputációval elérhető relatív hatékonyság (%)

Ha a hiányzó információk aránya nem túl magas, akkor igen kevés javulást jelent néhánynál több imputált adatbázis készítése és elemzése.

Az *m* db imputált adatbázison elvégzett elemzések eredményeinek összegzésére Rubin (1987) a következő módszert ajánlotta:

Minden elemzésből mentsük el a becsült paraméterek és standard hibák értékét. Legyen $\hat{\theta}_j$ a becsülni kívánt paraméter értéke (például egy regressziós együttható) a j-edik adathalmazból (j=1,2,...,m). U_j pedig legyen a $\hat{\theta}_j$ varianciája. Az összesítés utáni becslés az egyedi becslések átlaga lesz:

$$\bar{\theta} = \frac{1}{m}\sum_{j=1}^{m}\hat{\theta}_j \tag{5}$$

Ezen becslés standard hibájához először az átlagos imputáción belüli varianciát:

$$\bar{U} = \frac{1}{m}\sum_{j=1}^{m}U_j . \tag{6}$$

és az imputációk közötti varianciát kell kiszámolni:

$$B = \frac{1}{m-1}\sum_{j=1}^{m}(\hat{\theta}_j - \bar{\theta})^2 , \tag{7}$$

A teljes variancia:

$$T = \bar{U} + \left(1 + \frac{1}{m}\right)B. \tag{8}$$

ahol az $(1+\frac{1}{m})$ a véges m miatti korrekciós tényező.

Az együttes standard hiba pedig $\sqrt{T}$ lesz.

A $\gamma = \frac{(1+m^{-1})B}{T}$ a nemválaszolás miatt a θ–ról hiányzó információk becsült aránya.

Nagy minták esetén a θ–ra vonatkozó szignifikancia tesztelése a $t = \frac{\theta - \bar{\theta}}{\sqrt{T}}$

próbafüggvénnyel történhet, ami a nullhipotézis alatt Student-féle t-eloszlást követ az alábbi szabadságfokkal:

$$\nu = (m-1)\left(1 + \frac{1}{m+1}\frac{\bar{U}}{B}\right)^2 \tag{9}$$

ami a Satterthwaite közelítésen alapul (Rubin 1987).

A szabadságfok javított értéke kis mintákra:

$$\nu' = (\nu^{-1} + \hat{\nu}_{megfigy}^{-1})^{-1}, \tag{10}$$

ahol

$$\hat{\nu}_{megfigy} = (1-\gamma)\left(\frac{\nu_{telj}+1}{\nu_{telj}+3}\right)\nu_{telj}, \tag{11}$$

és ν_{telj} az adathiányt nem tartalmazó adatbázis esetén alkalmazandó szabadságfok. (Barnard és Rubin, 1999).[7]

Intervallumbecslés a paraméterre szintén ezek felhasználásával készülhet. További módszereket ismertet az eredmények összesítésére többszörös imputáció esetén Schafer (1997, 4. fejezet).

A többszörös imputáció különböző módszerekkel történhet attól függően, hogy milyen jellegzetességekkel bír az adathiány.

A többszörös imputációt besorolhatnánk a modell alapú eljárások közé is, mert legtöbbször Bayes-i eljáráson alapul: szükség van egy parametrikus modellre a teljes adatokra vonatkozóan és prior eloszlásra az ismeretlen modell paraméterekre (ez esetlegesen lehet nem-informatív), aztán a hiányzó adatokra készít m független szimulációt a hiányzó adatok feltételes eloszlását használva (Bayes-tétel). Bonyolultabb parametrikus modellek esetén speciális számítási technikákra is szükség lehet, ezek közül leggyakrabban a Markov lánc Monte Carlo[8] (Markov chain Monte Carlo (MCMC)) szimulációt használják.

Rubin (2003) egy olyan MCMC szimulációt és beágyazott többszörös imputációt alkalmazó modellt ír le, amelyet három változó esetén a következő módon lehet illusztrálni: Legyen a három változónk *X*, *Y* és *Z*. Kezdjük azzal, hogy valahogyan kitöltjük az *Y* és *Z* hiányzó értékeit, majd a megfigyelt *X*-kel építünk egy *X|Y,Z* modellt és e modell segítségével imputáljuk a hiányzó *X*-eket. Ezek után dobjuk ki az imputált *Y* értékeket és illesszünk egy *Y|X,Z* modellt a megfigyelt *Y*-okra , majd ezzel a modellel imputáljuk a hiányzó *Y*-okat. Aztán dobjuk ki az imputált *Z* értékeket és illesszünk egy *Z|X,Y* modellt a megfigyelt *Z*-kre , majd ezzel a modellel imputáljuk a hiányzó *Z*-ket. Az iteratív eljárás mindaddig folytatja a fenti lépések ismétlését, míg a kapott paraméterek nem konvergálnak.

A többszörös imputáció előnyei között megemlítendő, hogy könnyen érthető és elég robusztus a változók normalitási feltételének sérülése esetén is. Még például bináris,

[7] A fenti képletek pontos elméleti háttere megtalálható a hivatkozott művekben.

[8] A Markov lánc véletlen változók sorozata, amelyben minden egyes elem eloszlása az előző értékétől függ. A módszert eredetileg a fizikában használták egymással kölcsönhatásba lépő molekulák egyensúlyi eloszlásának feltárására. A statisztikai alkalmazások során többdimenziós, más módszerekkel megfoghatatlan eloszlások generálására használják.

vagy ordinális skálán mérő kategóriás változók esetén is gyakran elfogadható a normalitási feltétel melletti imputáció, majd a kapott folytonos imputált érték kerekíthető a legközelebbi kategóriára. Az erőteljes aszimmetriával rendelkező eloszlások közel normálissá transzformálhatók (például logaritmizálással), majd imputáció után visszatranszformálhatók az eredeti skálára. Hátránya viszont, hogy időigényes a három-tíz adatbázis imputálása, majd külön-külön az elemzések elvégzése, végül ezek összegzése. Ráadásul az összegzés módszertana még nincs minden statisztikai modellre kidolgozva. A többszörös imputációt több statisztikai szoftverbe is beépítették (például: a SAS enterprise Miner-hez írt IMISS –Intelligent Multiple Imputation Software System) ezek használatával az eljárás időigénye csökkent és sok kutató számára vonzó megoldássá vált.[9]

2.4. Modell alapú eljárások

A modell alapú eljárások egy modellt definiálnak a megfigyelt adatokra és a becsléseket a modell melletti posterior valószínűségekre, vagy likelihoodra alapozzák. A megközelítés előnye a rugalmasság, a modellnél alkalmazott feltételezések explicit volta és az adathiányt is beépítő varianciabecslések elérhetősége.

Ilyen modell alapú becslés a Maximum Likelihood becslés, ami nagyon jó nagymintás tulajdonságokkal rendelkezik (konzisztens, aszimptotikusan hatásos, határeloszlása normális) (Hunyadi és Vita, 2002). A hiányzó adatok mintázata azonban nem mindig teszi lehetővé az ML becslések explicit számítását.

Tegyük fel, hogy van egy modellünk az Y-ra, melynek eloszlását az $f(Y|\theta)$ sűrűségfüggvénnyel írhatjuk le, ahol θ ismeretlen paraméter. Legyen $Y = (Y_{megfigyelt}, Y_{hiányzó})$, ekkor $f(Y|\theta) = f(Y_{megfigyelt}, Y_{hiányzó} |\theta)$ az $Y_{megfigyelt}$,és az $Y_{hiányzó}$ együttes eloszlását leíró sűrűségfüggvény, az $Y_{megfigyelt}$ peremeloszlása pedig :

$$f(Y_{megfigyelt}|\theta) = \int f(Y_{megfigyelt}, Y_{hiányzó} |\theta)\, dY_{hiányzó}$$

Ekkor MAR adathiány esetén a likelihood:

[9] Forgarty (2005) ABB eljáráson alapuló többszörös imputációt alkalmazott a scoring modelleknél elutasítottak pótlására.

$$L(\theta \mid Y_{megfigyelt}) = \int f(Y_{megfigyelt},\ Y_{hiányzó} \mid \theta)\ dY_{hiányzó}$$

Ekkor a ML becslés a következő egyenlet megoldásával kapható:

$$D_{\ell}(\theta \mid Y_{megfigyelt}) = \frac{\partial \ln L(\theta \mid Y_{megfigyelt})}{\partial \theta} = 0$$

Ha nincs a fenti egyenletnek zárt alakú megoldása, akkor iteratív módszerek alkalmazására van szükség.

Ilyen iteratív módszer a Newton-Raphson algoritmus is.

Legyen $\theta^{(0)}$ a θ kiinduló becslése (például a teljesen megfigyelt egységeken alapuló becslés), és $\theta^{(t)}$ a t-edik iteráció becslése θ-ra. A Newton-Raphson algoritmust a következő egyenlet írja le:

$\theta^{(t+1)} = \theta^{(t)} + I^{1}(\theta^{(t)} \mid Y_{megfigyelt})\, D_{\ell}(\theta^{(t)} \mid Y_{megfigyelt})$,

ahol $I(\theta \mid Y_{megfigyelt})$ a megfigyelt információ:

$$I(\theta \mid Y_{megfigyelt}) = \frac{\partial^2 \ln L(\theta \mid Y_{megfigyelt})}{\partial\theta\partial\theta}$$

és $D_{\ell}(\theta^{(t)} \mid Y_{megfigyelt})$ pedig a loglikelihood függvény θ-szerinti első deriváltja.

Ha a loglikelihood függvény konkáv és egymóduszú, akkor a $\theta^{(t)}$ iterációk sorozata konvergál a θ ML becsléséhez.(Little és Rubin, 2002.)

Egy alternatív módszer a hiányzó adatokkal való becslések készítéséhez a várakozás maximalizáció (EM - expectation maximization), ami nem igényli a második deriváltak számítását, így nincs szükség olyan komplex programozási megoldásra, mint a Newton-Raphson algoritmust alkalmazó módszerek esetén.

A következőkben ezt a módszert mutatjuk be, mert a gyakorlatban nagyon elterjedt az alkalmazása.

Várakozás maximalizáció (Expectation Maximization (EM))

A várakozás maximalizáció egy általános módszer maximum likelihood becslésre MAR típusú adathiány esetén. A módszer egy iteratív eljárás, amely két lépésből áll. Először, a várakozási lépésben (E) kiszámítják a teljes adatokat tartalmazó állományra a loglikelihood várható értékét. Aztán a maximalizáló lépésben (M) a kapott várható értékeket behelyettesítik a hiányzó értékek helyére és maximalizálják a likelihood függvényt, mintha nem lett volna hiányzó adat, így új paraméterbecsléseket kapnak.

Ez az iteratív eljárás mindaddig folytatja a fenti két lépés ismétlését, míg a kapott paraméterek nem konvergálnak. Konvergenciáról akkor beszélhetünk, ha a paraméterbecslések változása lépésről lépésre egyre kisebb lesz mígnem teljesen elhanyagolhatóvá válik. A konvergenciához annál több iteráció szükséges, minél több a hiányzó adat.

Nézzük meg egy egyszerű példán, hogyan működik az EM módszer. A becslés elvégzéséhez igazából nincs szükség az EM algoritmusra, csak a szemléltetés kedvéért választottuk.

Tegyük fel, hogy négyszer egymás után feldobunk egy pénzérmét, aminek az eredménye: (Fej,Fej,Írás,?), ahol a ? azt jelenti, hogy a negyedik dobás eredményét valamilyen oknál fogva nem ismerjük. Legyen a becsülni kívánt sokasági paraméter a Fej dobás valószínűsége, π. A teljes Y adatállományt felbontjuk megfigyelt és hiányzó részre: $Y = (Y_{megfigyelt}, Y_{hiányzó})$, a megfigyelt adatok valószínűségét a következő módon kapjuk:

$P(Y_{megfigyelt}|\pi) = \sum_{Yhiányzó} P(Y|\pi) = P((F,F,Í,Í)|\pi) + P((F,F,Í,F)|\pi) = \pi^2(1-\pi)^2 + \pi^3(1-\pi) = \pi^2(1-\pi)$

A megfigyelt adatok valószínűsége tehát ugyanaz, mintha a negyedik dobást egyáltalán nem vennénk figyelembe. Ekkor a π maximum likelihood becslése:

$L(\pi|Y_{megfigyelt}) = P(Y_{megfigyelt}|\pi) = \pi^2(1-\pi)$

$$D_L(\pi|Y_{megfigyelt}) = \frac{\partial L(\pi|Y_{megfigyelt})}{\partial \pi} = 2\pi(1-\pi) - \pi^2 = 2\pi - 3\pi^2 = 0 \rightarrow \hat{\pi}_{ML} = \frac{2}{3}$$

Az illusztráció kedvéért nézzük, hogyan kaptuk volna meg ezt az eredményt az EM módszer segítségével! Az E várakozási lépésben felírjuk a teljes adatok loglikelihoodjának várható értékét a jelenlegi $\pi^{(t)}$ becslés mellett.

$$Q(\pi|\pi^{(t)}) = \pi^{(t)}(3\ln\pi + \ln(1-\pi)) + (1-\pi^{(t)})(2\ln\pi + 2\ln(1-\pi))$$

Az M maximalizálási lépésben keressük Q maximumát π szerint, hogy megkapjuk $\pi^{(t+1)}$ –et. Ebben az egyszerű esetben zárt formát kapunk az iterációra: $\pi^{(t+1)} = 0{,}5+0{,}25\pi^{(t)}$

Ha a kiinduló becslésünk mondjuk $\pi^{(0)} = 0{,}25$, akkor az iterációk sorozata: 0,2500; 0,5625; 0,6406; 0,6602; 0,6650; 0,6663; …, ami konvergál a 2/3-hoz.

Az EM megközelítés előnye, hogy jól ismert statisztikai tulajdonságai vannak és általában jobban működik, mint az egyszerűbb listwise és pairwise adattörlések, az átlaggal való helyettesítés, vagy a regressziós imputálás (Little, 1979, Donner és Rosner, 1982, Lee és Chiu, 1990). Monte Carlo szimulációk is hasonló eredményeket mutattak (Malhotra, 1987, Graham és Donaldson, 1993). Ugyanakkor ez az előny sokszor igen kicsi lehet (Donner és Rosner, 1982). A módszer hátránya viszonylagos bonyolultsága, ami miatt inkább csak statisztikusok számára vonzó megoldás.

A legfontosabb gyengéje a módszernek, hogy a becsült adathoz nem ad egy bizonytalansági komponenst. A gyakorlatban ez azt jelenti, hogy míg a paraméterbecslések torzítatlanok lesznek, addig a standard hibák és a kapcsolódó tesztek nem megbízhatóak. Ez a hiányosság arra késztette a statisztikusokat, hogy újabb likelihood alapú módszereket fejlesszenek ki. Ilyenek a teljes információs maximum likelihood módszer vagy a fent már tárgyalt többszörös imputáció alkalmazása.

A *teljes információs maximum likelihood* (Full Information Maximum Likelihood (FIML) vagy Raw Maximum Likelihood) minden elérhető adatot használ, hogy maximum likelihood alapú becsléseket készítsen. A módszert részletesen ismerteti például Wothke (1998).

A módszer előnye az EM módszerrel szemben, hogy lehetővé teszi a standard hibák és teszt statisztikák közvetlen számítását. Hátránya, hogy feltételezi a változók együttes többváltozós normális eloszlását és nem készít új adatmátrixot az imputált értékekkel.

A maximum likelihood módszer MAR típusú adathiányt feltételez, de a listwise és pairwise törlésekhez képest még nem véletlenszerű adathiány esetében is jobb eredményeket ad (Wothke, 1998).

Modellek nem véletlenszerű adathiány kezelésére

A fentiekben ismertetett eljárások alkalmazásának szükséges feltétele a véletlenszerű adathiány (MAR). Vannak azonban olyan körülmények, amelyek esetén ez a feltételezés nem tartható, mert az adathiány kapcsolatban van a hiányt tartalmazó változóval.

Ekkor az adathiány jellegét figyelembe vevő modellek alkalmazására van szükség.

A NMAR adathiánnyal foglalkozó kutatások fundamentálisan eltérő megközelítésük alapján két csoportra bonthatók: *szelekciós modellek* és *mintázat-keverék (pattern-mixture) modellek*. Ezek a modellek az együttes valószínűséget eltérő módon bontják fel. A *szelekciós modellek* a $P(y_{hiányzó}, y_{megfigyelt}) = P(y_{hiányzó}| y_{megfigyelt}) P(y_{megfigyelt})$ felbontást használják. A szelekciós modellek feltételezik, hogy az adathiányt tartalmazó változó akkor és csak akkor figyelhető meg, ha egy másik változó (ami nem megfigyelhető) átlép egy küszöbértéket. Ilyen modellt alkalmazott Heckman (1976). Ezt a Heckman-féle kétlépcsős modellt a reject inference módszerek között (III. rész) részletesen ismertetjük. A szelekciós modellek esetén a likelihood szokatlan eloszlású lehet, mert a paraméterek becsléséhez sokszor kevés információ áll rendelkezésre (Schafer és Graham, 2002).

A megoldás alternatívájaként alkalmazhatók a *mintázat-keverék* modellek, amelyek a $P(y_{hiányzó}, y_{megfigyelt}) = P(y_{megfigyelt}| y_{hiányzó}) P(y_{hiányzó})$ felbontást alkalmazzák.

A mintázat-keverék (pattern-mixture) modellekkel foglalkozó tanulmányok: Hedeker és Gibbons (1997), Little és Schenker (1994), Little (1993), és Glynn, Laird és Rubin (1986). Ezek a modellek kategorizálják a hiányzó értékek különböző mintázatait egy magyarázó változóba és ezt a magyarázó változót beépítik az adott statisztikai modellbe. Ezek után meghatározható, hogy az adathiány jellegzetességének van-e prediktív ereje akár önállóan (közvetlen hatás), akár más változókkal együttesen (interakciós hatás). A módszer előnye, hogy nem feltételezi a véletlenszerű adathiányt és részben használhatók hozzá statisztikai szoftverek, például a SAS MIXED proc. (pl. Hedeker és Gibbons (1997)), hátránya viszont, hogy az elemzőnek magának kell bizonyos lépéseket leprogramozni. Ha a megfigyelések számához képest sok változó esetén van relatíve sokféle eredetű adathiány, akkor a módszer elegendő adat hiányában nem működik.

3. Összegzés

A hiányzó adatok kezelésére nem létezik tehát egyetemesen legjobb megoldás. Pontosabban a legjobb gyógymód itt is a megelőzés.[10] Ez sajnos nem mindig lehetséges, így ha már van adathiány, és az nem teljesen véletlenszerű, akkor valamilyen módon kezelni kell.

[10] „*The only real cure for missing data is to not have any*." (Anderson, Basilevsky, Hum, 1983)

Összességében elmondható, hogy az általánosan használt egyszerű adathiány kezelési eljárásoknál (listwise és pairwise törlés, átlag imputálás) a hot deck, a maximum likelihood alapú és a többszörös imputációs eljárások a legtöbb esetben jobban teljesítenek. Mivel egyre szélesebb körben elérhető és könnyen használható szoftverek is tartalmazzák ezeket az eljárásokat, így az elméleti szerepükön túl az alkalmazásuk is egyre gyakoribb. Ezen módszerek mindegyike feltételezi a véletlenszerű adathiányt, vannak azonban újabb statisztikai modellek a nem véletlenszerű adathiány kezelésére is. Ezekhez is használhatók (részben) az ismert statisztikai programcsomagok.

Az eljárások közötti választásban fontos szerepe van annak, hogy a cél *paraméterbecslések és teszt statisztikák készítése*, vagy *konkrét megfigyelések hiányzó adatának becslése*.

Az első esetben az adatbázis felhasználója kezeli a hiányzó adatokat és választhatja a saját elemzéséhez leginkább megfelelő metódust. A második esetben, ha például statisztikai hivatalok, kormányzati szervek nyilvánosságnak szánt adatbázisairól van szó, vagy olyan vállalati adatbázisokról, amelyeket sokféle belső kutatáshoz használnak, akkor olyan megoldást kell választani, ami nem igényel túl komplex bánásmódot a végső elemzések elvégzésekor. Ekkor például nem nagyon alkalmazható a többszörös imputáció.

Fontos, hogy a választott imputációs eljárás kompatibilis legyen az imputált adatbázison később elvégzendő elemzésekkel. Az imputációs modellel szembeni elvárás, hogy megőrizze a későbbi vizsgálat tárgyát képező változók közötti kapcsolatokat. Ha például az Y változót egy olyan modellel imputálták, amelyik csak az X_1 változót tartalmazta, majd imputáció után a kutató egy lineáris regressziós modellt illeszt Y-ra X_1 és X_2 változók felhasználásával, akkor az X_2 együtthatója torzított lesz 0 felé, a helytelen imputáció következtében. Hasonló okokból panel felvételeknél például a keresztmetszeti kapcsolatokon kívül az adott változó korábbi hullámbeli tényleges vagy imputált értékét is figyelembe kell venni.

Az imputált adatbázisokhoz mellékelni kell az imputáló által alkalmazott modellt, mert így az elemző láthatja, hogy milyen változókat vontak be a modellbe és mely változók közötti kapcsolatokat tekintettek implicite 0-nak.

Sokan az imputációt egyfajta *statisztikai alkímiának* tartják, amelyben a semmiből valahogyan új információ keletkezik. Ez a felvetés helytálló lehet az olyan imputációs eljárásokkal kapcsolatban, amelyek az imputált értékeket ugyanúgy kezelik, mint a ténylegesen megfigyelteket. Ha viszont pontosan közlik az alkalmazott módszert és a

hiányzó adatok bizonytalansága is megjelenik, akkor a hiányzó adatok megfelelő kezelésével eltüntethető, vagy legalábbis csökkenthető a nem teljesen véletlen adathiányból eredő torzítás.

Ebben a fejezetben tehát áttekintettük a hiányzó adatok típusait és a kezelésükre leggyakrabban alkalmazott módszereket.

A hitelkérelmek elbírálása után az elutasított kérelmezők esetében a legtöbb változót ismerjük, de a hitelkockázatot leíró visszafizetési adatok hiányoznak. Az adósminősítéshez használt modellek esetében ez éppen az eredményváltozó.

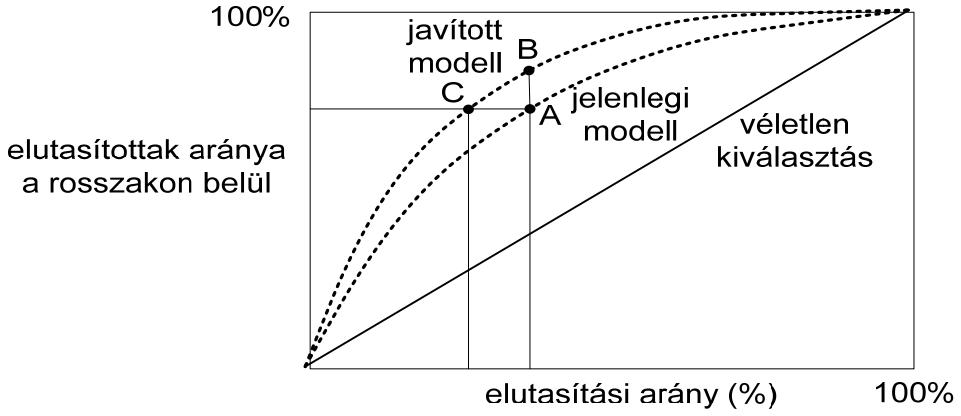
A szelekció okozta torzítással és csökkentésének lehetséges módszereivel foglalkozunk majd a III. fejezetben.

Mielőtt rátérnénk erre a problémára, tekintsük át röviden a credit scoring feladatát és alkalmazott módszereit!

II. Credit scoring

Az utóbbi 15-20 évben forradalmi változás történt a pénzügyi szolgáltatások piacán. A bankok automatikus döntéshozói módszereket és döntéstámogatási modelleket kezdtek alkalmazni, hogy felgyorsíthassák a hitelengedélyezési döntéseket.

A credit scoringnak nagyon fontos szerepe volt a fogyasztói hitelek állományának robbanásszerű növekedésében. Egy pontos és automatizált kockázatelemző rendszer nélkül a bankok nem tudták volna ekkora ütemben növelni lakossági kihelyezéseiket. A credit scoring azért nagyon vonzó kutatási téma, mert ha csak egy picit is sikerül javítani a modellek teljesítményén, az óriási profitnövekedést vagy kockázatcsökkenést eredményezhet a bank számára, hiszen nagy volumenű kihelyezésekről van szó (lásd 2. ábra). A kockázatok szofisztikáltabb értékelése, pontosabb megítélése az ügyfelek számára is előnyös, mert a jó adósok számára a kockázati felár csökkentését teszi lehetővé.



2. ábra A scoring modell javítása

Az ábrán az A pont jelöli a jelenlegi hitelezési gyakorlat melletti elutasítási arányt és azt, hogy a rosszak mekkora arányát sikerül kiszűrni. Ha egy picit sikerül javítani a modellen (felső szaggatott vonal), akkor a jelenlegi elutasítási arány mellett növelhető a rosszak elutasítási aránya (B pont) (ugyanakkora hitelezési volumen mellett csökkenthető a rosszak hitelezéséből adódó veszteség). Vagy ugyanannyi rossz hitel mellett növelhető a hitelezési volumen (csökkenthető az elutasítási arány) (C pont).

1. Mi a credit scoring?

A credit scoring magyarul pontozásos hitel (vagy hitelkérelem) minősítést jelent, olyan döntési modellek és mögöttes módszereik együttese, amelyek segítik a hitelezőt a hitelnyújtásban. A credit scoring technikák megbecsülik egy adott ügyfél hitelezésének kockázatát. Mondhatnánk úgy is, hogy a credit scoring felméri egy adott ügyfél hitelképességét. A hitelképesség azonban nem túl szerencsés kifejezés, mert az nem egy tulajdonsága az egyénnek, mint a magasság, életkor, vagy akár a jövedelem. Egyes bankok hitelképesnek tarthatják a potenciális ügyfelet, mások nem, attól függően, hogy milyen a kockázatvállalási hajlandóságuk, a hitelezési politikájuk[11] vagy a meglévő portfoliójuk.

Minden technika alkalmazása egy nagy adatbázisra épül, amely tartalmazza a korábbi ügyfelek adatait (olyan jellemzőket, amelyeket a kérelemben rögzítettek) és a hiteltörténetüket (problémamentes, vagy akadtak késedelmek, nemfizetések). A módszerek (különböző technikákkal) megpróbálják feltárni a kapcsolatot az ügyfél jellemzői és fizetési hajlandósága (képessége) között. Egyes módszerek egy scorecard-ot (pontozásos kártyát) eredményeznek. Minden tulajdonság kap valamilyen pontszámot (score) és a pontszámok összege alapján eldönthető, hogy az ügyfélnél nagy-e a nemfizetés kockázata. Más technikák nem eredményeznek ilyen scorecard-ot, hanem direkt megmutatják a nemfizetés valószínűségét. Mégis ezen módszerek és modellek összefoglaló neve a credit scoring lett.

Hitelkérelem elbírálási rendszerről először Dunham (1938) tett említést. Ő a fontos ismérveket szakértői alapon választotta ki, ehhez nem használt statisztikai módszereket. Durand 1941-ben kezdte el vizsgálni, hogy a szakértői (tapasztalati) alapon fontosnak tartott jellemzők megfoghatók –e statisztikailag is. Vizsgálataihoz diszkriminancia analízist használt. Javaslatokat fogalmazott meg a hitelkockázat elemzésére is, így az ő munkája tekinthető a mai credit scoring előfutárának.

A credit scoring napjainkra a mindennapi banki hitelezési gyakorlat egyik legfontosabb eszközévé vált. Használják a személyi -, ingatlan -, vállalati hitelek, hitelkártyák és lakossági lízing ügyletek elbírálásánál is, valamint minden olyan

[11] Az *óvatos (konzervatív)* hitelezési politika célja, hogy lehetőleg minden hitelt visszafizessenek. Ezt támogathatja például, ha a kockázatkezelőket a portfolió minősége alapján premizálják.
Az *agresszívebb (liberális)* hitelezési politikára nagyobb kockázatvállalás jellemző, amelynek célja lehet a piaci részesedés növelése. Ezt támogathatja, ha a kockázatkezelők premizálásában szerepet kap a portfolió nagysága is.

pénzügyi szolgáltatásnál, ahol *tömegszerű* kiszolgálás történik. A hangsúly itt a tömegszerűségen van. Ezeknek az adósminősítő rendszereknek a lakossági - és kisvállalkozási - (relatíve kisösszegű és nagyszámosságú) hitelek esetében van nagyobb jelentősége.

A credit scoring módszerek széles skálája alakult ki napjainkig. Ebben többek között szerepet játszott, hogy egyrészt újabb és újabb banki termékek jelentek meg a piacon, másrészt az eredményesebb piaci szereplés érdekében az ügyfeleket is szegmentálták a bankok. Ezek az egymástól eltérő termékcsoportok és ügyfélszegmensek más-más sajátosságokkal rendelkeznek, így más vizsgálati szempontokat is igényeltek. Például az új kérelmezőknél alkalmazható (application) scorecard nem tartalmaz olyan változókat, amelyeket egy régebbi ügyfél újabb igénylésénél már ismerünk (például az eddigi hitelek rendben megtérültek-e, mennyire használta ki az eddigi hitelkeretet). A már meglévő ügyfelekre vonatkozó viselkedési (behavioral) scoring ezeket a visszafizetési és használati szokásokat jellemző adatokat is felhasználja.

2. A credit scoringban alkalmazott módszerek

A credit scoring modellek feladata, hogy támogassák a hitelezési döntést. A modelltől elvárjuk, hogy megmondja milyen változókat kell figyelembe venni a döntés során (szignifikáns magyarázó változók) és ezek felhasználásával megadja a döntési módszert is.

A hitelkérelmeket a modellek segítségével szeretnénk két vagy több kockázati csoportba besorolni.

Az erre a célra leggyakrabban alkalmazott módszerek egy lehetséges csoportosítása (Kiss F., 2003):

A) Hagyományos módszerek:

- lineáris valószínűségi modell,
- probit és logit modellek,
- diszkriminancia analízis,
- klasszifikációs fák (rekurzív felosztási algoritmusok),
- lineáris programozás
- k-legközelebbi szomszéd

B) Mesterséges intelligencia módszerek:

neurális hálózatok,

szakértői rendszerek,

genetikus algoritmusok.

Az egyszerűség kedvéért mi két osztályba soroljuk az ügyfeleket. Az egyikbe tartoznak azok, akik nagy valószínűséggel nem fogják visszafizetni a hitelt (a „rosszak"), a másikba azok, akiknél ez a valószínűség alacsony (a „jók"). Feltételezzük, hogy minden ügyfél besorolható a két csoport valamelyikébe, de csak az egyikbe, és hogy ezek a csoportok fixek.

Bármilyen modellt építünk, először el kell döntenünk, hogyan definiáljuk a jó és a rossz hiteleket. Tarthatjuk rossz hitelnek például a több mint 6 hónapja lejárt követeléseket, vagy a legalább három törlesztőrészlettel elmaradt ügyfeleket, de alkalmazhatunk komplexebb definiciókat is , ha az adatbázisunk lehetővé teszi.
Ha kész az adatbázis a megfelelően definiált függő változóval, akkor elkezdhetjük a modellépítést.
Tekintsük át a credit scoringban alkalmazott módszereket! (A fenti felsorolásban szereplő legközelebbi szomszéd - és genetikus algoritmus módszerek alkalmazása a gyakorlatban nem elterjedt, ezért ezek ismertetésétől itt eltekintünk.)

2.1 Lineáris valószínűségi modell

Ez egy lineáris regressziós modell, ahol az eredményváltozó értéke 1, ha az ügylet rossz hitelnek bizonyult, és 0, egyébként. A modell:

$$y = \boldsymbol{\beta}'\mathbf{x} + \varepsilon$$

ahol : y a fizetési problémák meglétét leíró eredményváltozó,

$\mathbf{x}$ a magyarázó változók vektora,

$\boldsymbol{\beta}$ a regressziós paraméterek vektora,

ε pedig a véletlen változó

Ekkor az $\hat{y} = \hat{\beta}'\mathbf{x}$ becslés felfogható az adott $\mathbf{x}$ jellemzőkkel bíró kérelmek esetén becsült nemfizetési valószínűségként ($\hat{p}(y=1)$). A hitelezési döntés meghozatalakor az így kapott becsült nemfizetési valószínűséget kell összehasonlítani egy

meghatározott küszöbértékkel (vágási ponthatár – cut-off -value). A cutoff értéknél kisebb becsült nemfizetési valószínűségű kérelmeket el lehet fogadni, a nagyobbakat pedig el kell utasítani.

A modell gyakorlati megvalósítása során több probléma is felmerül. Például a maradéktag gyakran nem normális eloszlású és varianciája nem állandó, tehát a modell heteroszkedasztikus. (Lásd például Chhikara, 1989.) Ennek következményeként egyrészt a paraméterek hagyományos legkisebb négyzetek (OLS) becslése nem lesz BLUE (legjobb lineáris torzítatlan becslés), és nem lesz hatásos (bár torzítatlan és konzisztens marad), másrészt a regressziós együtthatók becsült varianciái és kovarianciái torzítottak és inkonzisztensek, így a szokásos tesztek (*t*- és *F*-próbák) nem érvényesek.

A legfőbb problémát viszont az jelenti, hogy a becsült bedőlési valószínűség értéke kívül eshet az értelmes [0;1] intervallumon, ezért ilyen lineáris regressziós modelleket a gyakorlatban ritkán alkalmaznak.

2. 2 Logit és probit modellek

A lineáris regressziós modell fent említett hiányossága arra késztette a kutatókat, hogy jobb megoldásokat keressenek. A becsült bedőlési valószínűség [0;1] intervallumba való kerülését egy alkalmas transzformációval lehet biztosítani. Az eloszlásfüggvények alkalmazása jó megoldásnak tűnt, mivel ezek monoton transzformációk és értékkészletük a [0;1] intervallum.

Választható például a standard normális eloszlás a becsült bedőlési valószínűség leírására:

$$\hat{p} = \Phi(\beta' x) = \int_{-\infty}^{\beta' x} \varphi(z)dz$$

ahol $\Phi(.)$ a standard normális eloszlás eloszlásfüggvénye, $\varphi(.)$ pedig a sűrűségfüggvénye.

Ez adja a *probit modellt*.

Ha a logisztikus eloszlásfüggvényt választjuk a bedőlési valószínűség leírására, akkor *logit modellt* kapunk. Ekkor:

$$\hat{p} = F(\beta' x) = \frac{1}{1 + e^{-\beta' x}}$$

A normális eloszlásfüggvénnyel szemben a logisztikus eloszlásfüggvénynek van zárt alakja.

Logisztikus regresszió esetén a becsült bedőlési valószínűség oddsának[12] természetes alapú logaritmusa írható le a magyarázó változók lineáris függvényeként:

$$\text{Ln (odds)} = \ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k = \beta' x$$

A regressziós paraméterek értelmezését az e^{β_j} (odds ratio) faktor szolgáltatja, amely az x_j magyarázó változó egységnyi abszolút növekményének az oddsra gyakorolt multiplikatív hatását mutatja, a többi magyarázó változó szinten tartása mellett.

Wiginton (1980) az elsők között használta a logisztikus regressziót a credit scoringban.

Napjainkban a logit modell a leggyakrabban használt klasszifikációs eljárás a credit scoring területén. Ennek legfőbb okai: könnyen interpretálható, jó a teljesítménye, nem csak klasszifikál, hanem bedőlési valószínűséget is becsül (Bázel II. előírás), valamint a módszer nem feltételezi a magyarázó változók normális eloszlását, így könnyen beépíthetők a kategóriás magyarázó változók. Nem elhanyagolható szempont, hogy sok felsőoktatási intézmény tananyagában szerepel, így többen értenek hozzá, mint például a neurális hálózatokhoz.

Az empirikus kutatás során mi is logisztikus regresszióval fogjuk építeni a scoring függvényeket.

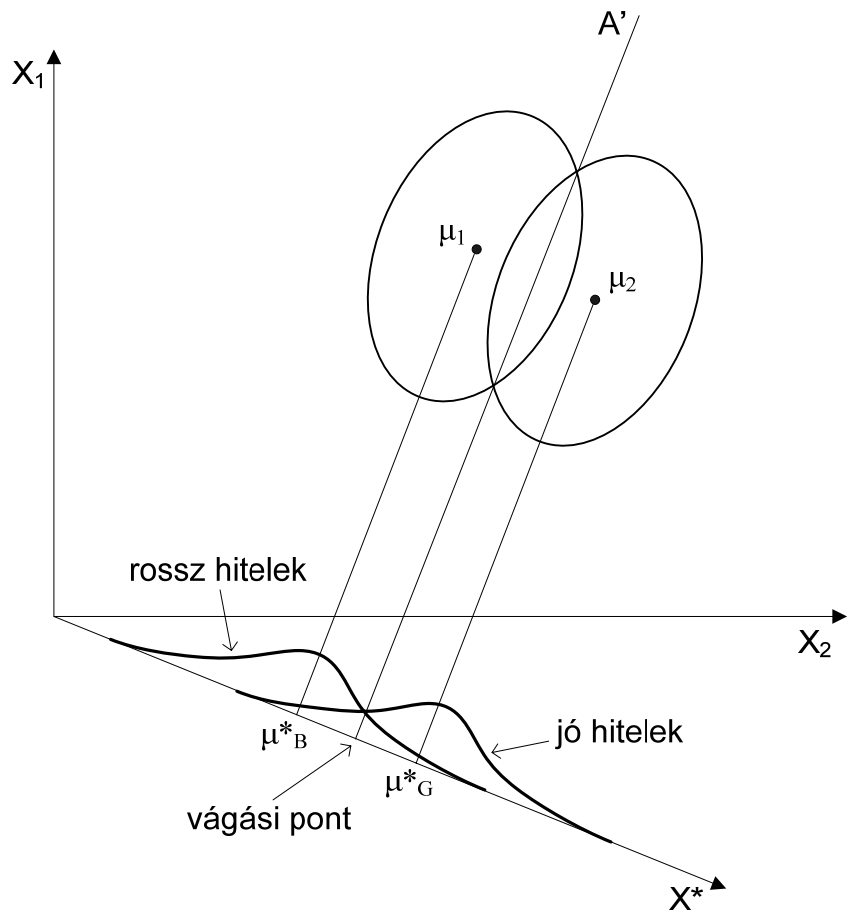
2.3 Diszkriminancia analízis

A scoring modell feladata az, hogy minden ügyfelet egyértelműen és pontosan besoroljon két csoport valamelyikébe: jó adós, rossz adós.

A diszkriminancia analízis az az eljárás, amelynek célja a megfigyelt *p* számú változó olyan lineáris kombinációinak előállítása, amelyek a lehető legjobban elkülönítik az ismert csoportokat, úgy, hogy minél kevesebb pont maradjon a nem megfelelő csoportban.

A 3. ábrán látjuk a diszkriminancia analízis grafikus modelljét kétváltozós esetre.

[12] Az odds , magyarul esélyhányados, a nemfizetési valószínűség és a komplementer esemény valószínűségének hányadosa ($\hat{p}/_i(1-\hat{p})$)



3. ábra A diszkriminancia analízis grafikus modellje kétváltozós esetre

Ha például feltételezzük, hogy $\mathbf{X_1}$ tengelyen jelöljük a havi kereset összegét, és $\mathbf{X_2}$ tengelyen pedig az adott munkahelyen eltöltött évek számát, akkor egy nagyon egyszerű, két magyarázó változós credit scoring modellt építünk fel. A redukált teret jelző egyenesen (**X***) megkapjuk a jó és rossz ügyletek eloszlását, valamint azt a *vágási ponthatárt* (cutoff score), amely kettévágja az átfedő területeket. A pontcsoportok területének átfedései azokat az eseteket jelzik, amelyeknél a legnagyobb a bizonytalanság a hitelkockázattal kapcsolatban. A konzervatív hitelezők a hitelezés során ténylegesen alkalmazott ponthatárt a metszésponttól jobbra állapítják meg, míg az agresszív hitelezési politikát folytatók balra térnek el.

Az ellipszisek lefedik a pontcsoportok egy adott hányadát (például 90-90%-át). Az **A'** egyenesre[13] merőleges **X*** tengelyen a csoportok egyváltozós eloszlásainak átfedése kisebb, mint bármilyen más egyenes esetén.

[13] Az **A'** egyenes az elválasztó hipersík.

A diszkrimináló függvények számának felső korlátja (a diszkrimináns tér lehetséges rangja) a csoportok (g) és a változók (p) számától függ, pontosan: $\min(g\text{-}1,p)$.[14] Így a példánkban két csoport és két változó esetén egydimenziós térre, az egyenesre redukálódik a modellünk.

A kiinduló helyzet tehát az, hogy adott két ügyfél csoport: G, a jó adósok és B, a rossz adósok. A feladat az, hogy egy új kérelmezőt osztályba soroljunk az őt reprezentáló $\mathbf{x=(x_1,x_2,\ldots,x_k)}$ értékek vektorának felhasználásával. A diszkriminancia analízis előállítja a $\boldsymbol{\lambda}'\mathbf{x}$, diszkrimináló függvényt, amelyben $\boldsymbol{\lambda}$ az $\mathbf{x_i}$ ismérvekhez tartozó együtthatók vektora. Ezeket az értékeket az eljárás úgy határozza meg, hogy a lehető legnagyobb különbséget hozza létre a két csoport között. (A külső eltérések maximumát - és azzal együtt a belső eltérések minimumát- keresi.)

A kérelmező jellemzőit tartalmazó $\mathbf{x}$ vektorról a módszer feltételezi, hogy a két csoportban többváltozós normális eloszlású[15], $\boldsymbol{(\mu_G, \Sigma_G)}$ illetve $\boldsymbol{(\mu_B, \Sigma_B)}$ várható értékkel és kovarianciával. Jelölje p_i annak a valószínűségét, hogy egy kérelmező az i csoporthoz tartozik, L (lost profit) azt az elvesztett profitot, amit a jók rosszként való félreklasszifikálása okoz, D (debt), pedig azt a veszteséget, amit egy rossz hitel beengedése (jóként való félreklasszifikálása) okoz. Abban az esetben, ha feltételezhetjük, hogy a két csoport kovariancia mátrixai egyenlők, azaz $\boldsymbol{\Sigma_G = \Sigma_B = \Sigma}$, az osztályba sorolási szabályt a várható téves besorolásból eredő költségek minimalizálásával kapjuk.

A szélsőérték feladat eredményeként[16] egy $\mathbf{x}$ adathalmazzal jellemzett kérelmezőt a G csoportba sorolunk, amennyiben

$$\lambda' x \geq \alpha + \ln\left(\frac{Dp_B}{Lp_G}\right),$$

Ahol: $\lambda = \Sigma^{-1}(\mu_G - \mu_B)$,

$$\alpha = \frac{\lambda'(\mu_G + \mu_B)}{2}.$$

Egyébként a kérelmezőt a B csoportba soroljuk be.

A diszkrimináló függvény, $\boldsymbol{\lambda}'\mathbf{x}$, által kapott értéket tehát az előzőleg beállított

[14] A diszkriminancia analízisről részletesen olvashatunk például Hajdu O.(2003) könyvében.

[15] Épp ez a módszer egyik gyengéje. A normalitási feltétel ugyanis sokszor nem teljesül a credit scoring területén, hiszen minőségi és diszkrét ismérvek is szerepelnek a modellekben.

[16] Az eredmények levezetését lásd például L.C. Thomas, D.B. Edelman és J.N. Crook (2002), "Credit Scoring and Its Applications", *Society for Industrial and Applied Mathematics, Philadelphia,* 42-46.o.

$$"cutoff" = \alpha + \ln\left(\frac{Dp_B}{Lp_G}\right)$$

vágási pontértékkel kell összevetni. Amennyiben a kérelmező a pontérték felett van, akkor a G csoportba kerül, egyébként a B-be.

Mivel a fenti modellben a diszkrimináló függvény **x**-re elsőrendű, így ezt az eljárást lineáris diszkriminancia analízisnek is nevezik. Ez a módszer nagyon sokáig egyeduralkodó volt a credit scoring modellek között. Ezt a módszert alkalmazta például Altman (1968) híres csődmodelljében és magyarországi vállalatokra Hajdu Ottó és Virág Miklós (1996) is. Abban az esetben, ha a csoportok kovariancia mátrixai nem egyenlők ($\mathbf{\Sigma}_G \neq \mathbf{\Sigma}_B$), az osztályozási szabály **x**-re négyzetes lesz, ezért ezt a modellt kvadratikus diszkriminancia analízisnek is szokták hívni.

2.4 Klasszifikációs fák

A klasszifikációs fák, vagy döntési fák, vagy más néven rekurzív felosztási algoritmus (Recursive Partitioning Algorithm, RPA) egy kifejezetten számítógépes alkalmazásra kifejlesztett osztályozó eljárás. Az alapötlet az, hogy a kérelemben rögzített válaszok kombinációi alapján csoportokat képzünk, és ezeket a csoportokat beazonosítjuk jó vagy rossz hitelkockázatúként, attól függően, hogy melyik van többségben az adott csoportban. Az RPA módszer eredménye egy osztályozó fa, amelynek csomópontjai és ágai egy olyan struktúrát alkotnak, amely egy adott kérelmezőt leíró jellemzőkhöz hozzárendeli a csoportot (jó adós vagy rossz adós), így jelezve, hogy egy új kérelmezőt célszerű-e meghitelezni, vagy sem.

Az eljárás először két csoportra bontja az összes kérelmezőt, ez a két alcsoport már sokkal homogénebb a hitelezési kockázat szempontjából, mint az eredeti. Aztán mindkét alcsoportot újabb két még homogénebb csoportra bontja és az eljárás így ismétlődik[17]. Ezért is nevezik rekurzív felosztásnak. Az eljárás addig folytatódik, amíg a kapott csoport meg nem felel a végpont követelménynek.

Az eljárás alkalmazásához három alapelvet kell lefektetni:

a) milyen szabályt alkalmazzunk a minta kettébontásához (szétválasztási szabály),

b) hogyan döntsük el, hogy az adott alcsoport már végpont (megállási szabály),

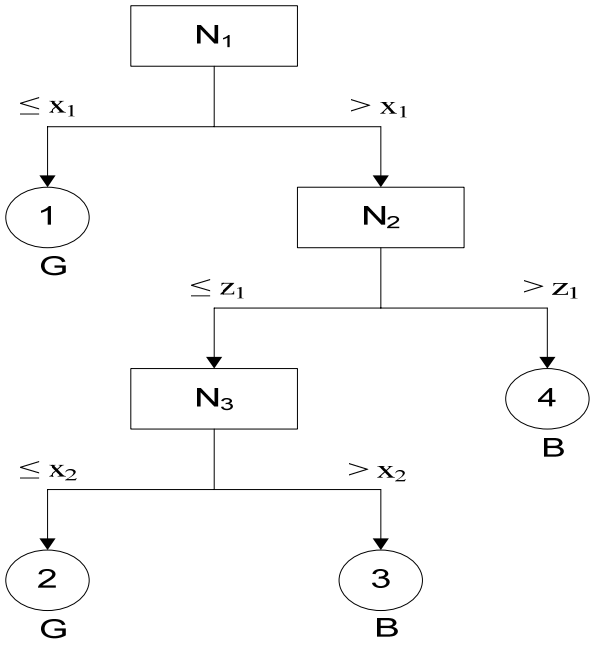
c) hogyan soroljuk be a végpontokat a jó vagy rossz kategóriákba.

[17] Vannak olyan módszerek, amelyek egy lépésben nem feltétlen csak ketté osztják az adott csoportot, hanem akár több csoportra is.

A végpontok kategóriákba való sorolása történhet egyszerűen úgy, hogy abba a kategóriába soroljuk, amelyik többségben van az adott alcsoportban. (Ez volt az alapelv.) De jobb megoldást jelent, ha figyelembe vesszük a téves besorolás eltérő költségeit is, és a besorolásnál ezt a várható költséget minimalizáljuk.

A legegyszerűbb szétválasztási szabályok minden jellemző esetében kiválasztják a legjobb szétválasztási értéket és ezek közül kiválasztják a legjobban szétválasztó jellemzőt és vágási értékét. Hogy mennyire jó a szétválasztás, arra különböző mérőszámokat lehet alkalmazni (például: Kolmogorov-Smirnov statisztika, szennyezettségi index, Gini-mutató, entrópia index, χ^2 statisztika (ez utóbbit használja például a CHAID[18]).…).

A modell egyszerű illusztrálására tegyük fel, hogy *N* kérelmezőt kell a róluk rendelkezésre álló két ismérv, **x** és **z** alapján besorolni a G (jók) és B (rosszak) csoport valamelyikébe. Egy feltételezett bináris osztályozó fa látható az alábbi ábrán:



4. ábra Egy feltételezett RPA fa

A fának négy végpontja van (a körök), amelyek közül az 1. és a 2. a G, a 3. és a 4. a B csoporthoz van hozzárendelve. Ez a hozzárendelés úgy történik, hogy a téves besorolásból eredő várható költséget minimalizálja, azaz, más megfogalmazásban:

[18] Chi-squared Automatic Interaction Detector, az eljárásról olvashatunk például Hámori, 2001 cikkében.)

csökkentse a lehető legkisebbre az osztályozó fa végpontjai és a csoportok közötti megfeleltetés megváltoztatása szükségességének kockázatát.

Annak kockázata (R: risk), hogy a *t*. osztályozó fa végpontot a G csoporthoz rendeljük, a következőképpen formalizálható:

$R_G(t)=Dp_Bp(t|B)$,

ahol továbbra is :

p_i jelöli annak valószínűségét, hogy egy objektum az *i* (*i* = G v. B) csoportba tartozik,

D(debt), pedig az a veszteség, amit a rosszak jóként való félreklasszifikálása okoz,

p(t|i) annak a feltételes valószínűsége, hogy egy *i* csoportba tartozó objektum a *t*. végponthoz kerül besorolásra.

Hasonlóképpen, annak kockázata, hogy a *t*. osztályozó fa végpontot a B csoporthoz rendeljük:

$R_B(t)=Lp_Gp(t|G)$,

ahol L(lost profit) az elvesztett profitot jelenti amit a jók rosszként való félreklasszifikálása okoz.

Ezek alapján, ha $R_G(t)<R_B(t)$, akkor a *t*. végpontot az algoritmus a G csoporthoz rendeli, egyébként pedig a B-hez.

Az RPA a kiinduló adathalmaz két részre (almintára) bontását végzi el az osztályozó fa tetőpontján. A válogatást egy jellemző vagy több jellemző lineáris kombinációjának felhasználásával hajtja végre az alábbiakban –a mintaszennyezettség (impurity) fogalmának bevezetésével – definiált „legjobb" szétválasztási szabály figyelembevételével.

A *t*. végponthoz tartozó minta szennyezettségének mértéke a következőképpen definiálható:

$I(t)= R_G(t)p(t|G)+ R_B(t)p(t|B)$,

amely úgy értelmezhető, mint a téves besorolásból származó várható költség, akkor ha a *t*. végponthoz tartozó objektumokat véletlenszerűen rendeljük a két csoporthoz.

A teljes T osztályozó fa I(T) szennyezettsége úgy definiálható, mint a végpontok szennyezettségének összege.

A fa bármely pontjában szétosztott minta szennyezettsége nagyobb, mint a belőle származtatott alminták szennyezettségének összege. Ennek alapján a *t*. csomópontban legjobb besorolási szabálynak az tekinthető, amelyiket felhasználva a legnagyobb szennyezettség-csökkenés érhető el. Ennél fogva tehát az RPA először megkeresi az adott pontban legjobb szabályt minden jellemzőre, illetve azok kombinációira, majd ez

alapján almintákra bont. A bináris osztályozási eljárás mindaddig folytatódik, amíg a további felbontás lehetetlen nem lesz, azaz amikor a szennyezettség már nem csökkenthető. Ekkor a besorolási eljárás befejeződik. Az RPA utolsó lépése a fa megfelelő komplexitásának kiválasztása.[19]

A 4. ábrán az RPA eljárást egy egyszerű, két csoportba sorolási feladaton keresztül mutatjuk be, amelynek során két (***x*** és ***z***) ismérvet használunk. Az RPA először az ***x*** változót választja osztályozó ismérvnek, és az N_1 csomópontban kettébontja az eredeti minta adathalmazt. Az eredő két alminta elemeinek meghatározásánál az $\mathbf{x_1}$ vágási értéket használja fel, mint a legjobb döntési szabályt. Tehát amennyiben valamely kérelmező ***x*** változója kisebb vagy egyenlő ezzel az értékkel, akkor a baloldali ágra kerül, és azonnal besorolódik a G csoportba a minimális kockázat mellett. Ellenkező esetben az objektum a jobboldali ágon az N_2 csomóponthoz tartozó almintába kerül, ahol egy hasonló osztályozási eljárás következik a ***z*** változón alapuló legjobb döntési szabály alkalmazásával.

Az RPA módszert sokan vizsgálták és használták fel eredményesen. Frydman, Altman és Kao (1985) a szorult helyzetben levő cégek osztályozási problémáját elemezte, összehasonlítva az RPA és a diszkriminanciaanalízisen alapuló eljárások hatékonyságát. Marais, Patell és Walfson (1984) az RPA és a probit modellek felhasználhatóságát vizsgálta kereskedelmi hiteleknél. Srinivasan és Kim (1987) az iparvállalati hitelezésben vizsgálta meg a döntési fák alkalmazhatóságát, összevetve a logit modellel és a diszkriminanciaanalízissel.

A tanulmányok egyértelműen azt mutatták, hogy az RPA a többi vizsgált eljárásnál lényegesen jobb osztályba sorolási pontosságot biztosít. A szerzők egyetértenek abban, hogy ez a tulajdonság az RPA eljárással előállított modellek nemparaméteres mivoltából fakad. A módszer további előnye, hogy automatikusan figyelembe veszi a magyarázó változók közötti interakciókat (csakúgy, mint a neurális hálók), míg a lineáris módszereknél ezeket az interakciókat előzetesen definiálni kell. A jó tulajdonságai ellenére a módszer nem annyira elterjedt, mint például a logisztikus regresszió, aminek az lehet az oka, hogy a kapott score érték nem folytonos, így nem lehet a cutoff értéket finoman beállítani.

[19] A döntési fákról és az optimális végső fa kiválasztásánál alkalmazott eljárásokról bővebben lásd például Breiman at. al. (1984) könyvét.

2.5 Lineáris programozás

A paraméteres osztályozási eljárásokkal szemben egy másik, igen ígéretes megoldásnak mutatkozik a matematikai programozás, amelynek alkalmazhatóságát először Mangasarian (1965) vizsgálta meg, két csoport és lineáris diszkrimináló függvény esetére. Freed és Glover (1981a), és Hand (1981) rámutattak, hogy a matematikai programozás akkor is használható, ha a két csoport nem feltétlen lineárisan szeparálható, olyan célok alkalmazásával, mint az abszolút hibák összegének minimalizálása (MSAE), vagy a maximális hiba minimalizálása (MME). Megmutatták, hogy egy csoportosztályozási problémát lineáris programozási feladatként is meg lehet fogalmazni, és ezáltal a modellalkotásban nagyobb szabadsághoz jutunk anélkül, hogy a paraméteres statisztikai modelleknél szokásos eloszlás feltétel korlátozná a lehetőségeket. A módszer illusztrálására vizsgáljunk meg egy egyszerű két csoportra bontási feladatot.

Legyen N darab kérelem, amelyeket két csoportba (G(jók)és B(rosszak)) akarunk besorolni. Az i-dik kérelemhez tartozó ismérvértékeket az $\mathbf{x_i}$ vektor tartalmazza. A feladat az, hogy meghatározzuk azt a $\mathbf{w}$ vektort és b küszöbértéket, amelyekre teljesül:

$\mathbf{w'x_i} \leq b$, ha $i \in G$,

és $\mathbf{w'x_i} \geq b$, ha $i \in B$.

Az $\mathbf{w'x}=b$ elválasztó hipersík elhatárolja a keresett két csoportot. Ha α_i azt jellemzi, hogy egy $\mathbf{x_i}$ adatokkal leírt kérelem mekkora mértékben sérti meg a két csoportot elválasztó határt, akkor a feladat úgy írható fel, hogy meg kell keresni az alábbi minimumot:

$$Min \sum_i c_i \alpha_i ,$$

ahol: $\mathbf{w'x_i} \leq b + \alpha_i$, ha $i \in G$,

$\mathbf{w'x_i} \geq b - \alpha_i$, ha $i \in B$.

A probléma tehát egy lineáris szeparálási feladat. A $c_i\alpha_i$ szorzat értelmezhető, mint az i. objektum téves besorolásából eredő költség, b pedig, mint egy vágási pontérték. Alkalmas módon megválasztva b és c_i értékeit a téves besorolásból fakadó várható költség minimalizálásának eredményeként megkapjuk a $\mathbf{w}$ vektort. Ismerve az optimális $\mathbf{w}$ vektort az egyes kérelmekhez számítható $\mathbf{w'x_i}$ pontérték, és az adott b vágási ponthatárral összevetve elvégezhető a besorolás.

E viszonylag egyszerű modellre alapozott elbírálási technikát Freed és Glover még ugyanabban az évben továbbfejlesztette sokkal összetettebb problémák, többek között

sokcsoportos osztályozási feladatok megoldására. Hardy és Adrian 1985-ben kimutatták, hogy a matematikai programozás legalább olyan eredményesen felhasználható a problémás hitelek besorolására, mint a tradicionálisan alkalmazott diszkriminancia analízisen alapuló modellek. Emellett kiemelték, hogy e módszer lényegesen nagyobb rugalmassággal bír a modellépítésben a kutatók számára. Például a célfüggvényben szereplő c_i súlyok alkalmas változtatásával követni lehet a hitelezési politika konzervatív vagy liberális irányú változásait[20].

A módszer rugalmasságának köszönhetően bármilyen kívánt torzítás is beépíthető a modellbe. Tegyük fel például, hogy az X_1 egy dummy változó, ami azt írja le, hogy az igénylő életkora 25 év alatti vagy sem (1,0), az X_2 szintén dummy változó pedig a 65 év feletti életkort jelzi. Ha azt szeretnénk, hogy a fiatalok alacsonyabb score-t kapjanak, mint a nyugdíjasok, egyszerűen felvesszük a $w_1 \leq w_2$ korlátot.

Az eljárás egyik hiányossága, hogy nem tesztelhető a kapott paraméterek szignifikanciája, mivel nincsenek meg az ehhez szükséges statisztikai alapok. Ziari et.al.(1997) javasolta a jackknife és a bootsrap eljárásokat a probléma orvoslására. Nath, Jackson és Jones (1992) klasszifikációs feladatra számos adatbázison (de nem hitelezési adatokon) összehasonlították a lineáris programozást a regressziós megközelítésekkel. Az ő eredményeik szerint a lineáris programozás nem olyan jól klasszifikál, mint a statisztikai módszerek.

2.6 Neurális hálózatok

A neurális hálózatokkal eredetileg megpróbálták modellezni az emberi agy kommunikációs és információ feldolgozási folyamatait. Olyan matematikai modellt akartak felépíteni, amely segítségével a természetes idegsejt (neuron) működése szimulálható. Egy valós neuron számunkra fontos részei a dentritek, amelyeken keresztül jelzés juthat a neuronba, és az axonok, amelyek segítségével a feldolgozott információ továbbjut a többi neuronhoz. Az információk feldolgozásában fontos szerepet játszanak a szinapszisok. Az axonok ezeken keresztül kapcsolódnak más neuronok dentritjeihez.

A matematikai neuron egy adott függvénnyel feldolgozza a dentritektől kapott információt, és ha a bemenő jel meghaladta az úgynevezett ingerküszöb értéket, akkor az axon közvetítésével továbbítja az információt. Az idegsejt legfontosabb

[20] Ha például a c_i-t csökkentjük a jók esetében és növeljük a rosszak esetében, akkor konzervatívabb irányba mozdulunk.

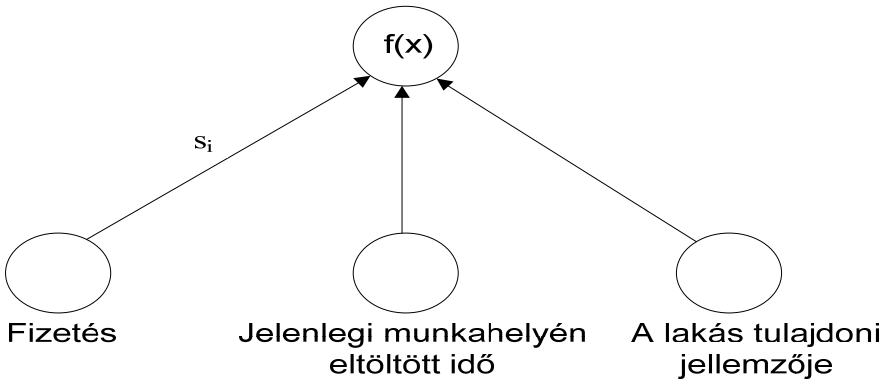
tulajdonsága az, hogy állandóan változtatja működését (azaz a belső függvényét) a kapott információk alapján – „tanul”. E tanulási folyamatban jelentős szerepet játszanak a szinapszisok, ugyanis képesek erősíteni vagy gyengíteni a többi neuronból érkező jelet. A tanulás folyamán megváltoznak a jelerősítési tényezők (másnéven súlyok) a szinapszisokon. A neurális hálózat sok ilyen neuronból áll.

A neuron klasszikus modellje sok bemenetből (amelyek a dentriteknek felelnek meg) és egy kimenetből (axon) áll. A jelerősítési tényezőket súlyszámokként ábrázolják, a küszöbértéket pedig egy alkalmasan megválasztott átviteli (aktivizálási) függvény állítja elő.

A neuronok a hálózaton belül elfoglalt helyük alapján rétegekbe sorolhatók. A bemeneti jeleket fogadók alkotják az úgynevezett bemeneti réteget, a kimenetet adók pedig a kimeneti réteget. E két rétegen keresztül kommunikál a hálózat, ezért ezeket látható rétegeknek is nevezik. A többi neuron, amely a hálózat belsejében helyezkedik el, úgynevezett rejtett rétegekben foglal helyet.

A hálózat felépítése tetszőleges lehet, pusztán attól függ, hogy mire akarjuk felhasználni. Az 5. ábra egy egyszerű, két réteget tartalmazó hálózatot mutat. Három bemenő változó értékéből dolgozik, és működése az *f(x)* átviteli függvény definíciójától függ:

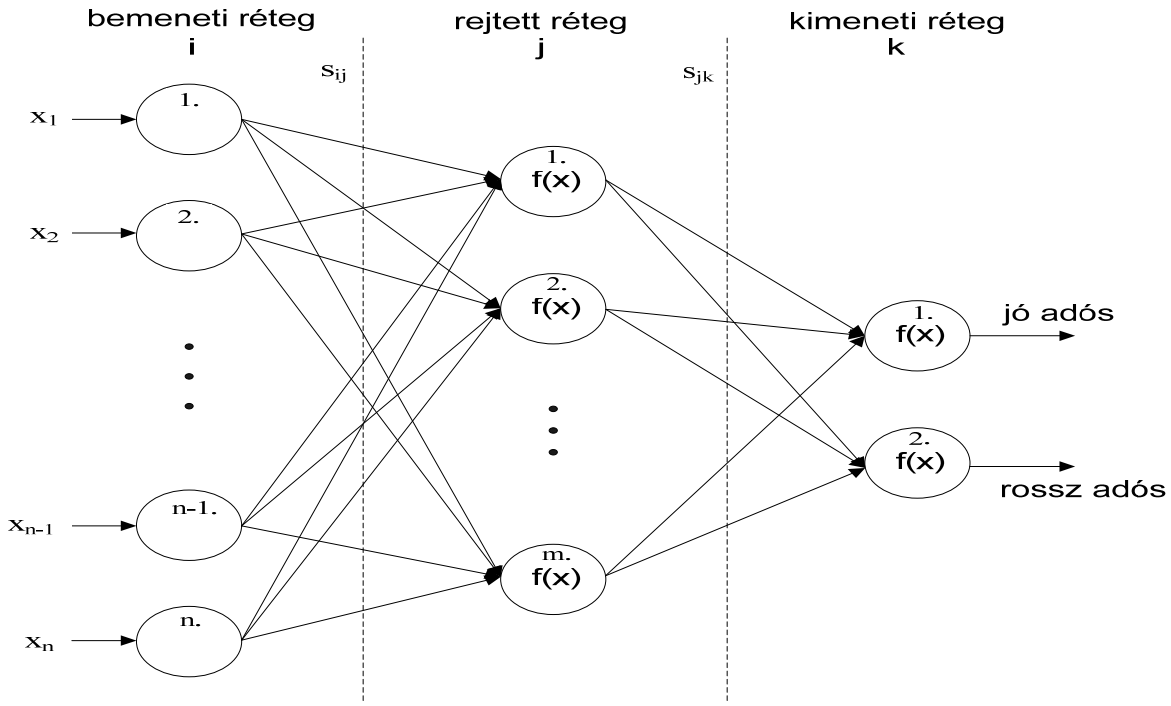
- ha f(x) lineáris függvény, akkor a hálózat lineáris regressziós modellként viselkedik;
- ha f(x) logisztikus függvény, akkor a hálózat logisztikus regressziós modellként működik.



5. ábra Egyszerű, két rétegből álló neuron hálózat

Bizonyítható, hogy csak egy minimálisan háromrétegű hálózattal lehet tetszőleges modellt előállítani (két réteggel a kizáró vagy kapcsolat nem valósítható meg). A

bemeneti és a kimeneti rétegek között tetszőleges számú, úgynevezett rejtett réteg helyezkedhet el, azonban leggyakrabban az egy (6. ábra), vagy néhány rejtett réteggel rendelkezőt alkalmazzák (ún. Multi-Layer Preceptron, MLP).



6. ábra Neurális hálózat egy rejtett réteggel

A neurális hálózat helyes működése szempontjából kritikus kérdés a tanítási illetve tanulási algoritmus helyes megválasztása is. E gyakorlási folyamatban alakul ki a modell súlyrendszere a hálózatnak megmutatott tapasztalati adathalmaz feldolgozása közben. Fontos, hogy a tanító adathalmaz megfelelően legyen kiválasztva, különben rendszerünk nem fog jól működni. A tanulás időigénye jelentős mértékben függ a kezdeti súlyrendszertől és az előírt konvergencia-hibától, ezért ezek alkalmas megválasztása fontos feladat a modellépítés során.

A neurális hálózatokon alapuló credit scoring modellek fejlesztésével és ezek alkalmazásaival számos cikk foglalkozik. Például Tam (1991, 1992) és Virág, Kristóf (2006) összevetette e modellek hatékonyságát a klasszikus módszerekkel egy vállalati csőd-előrejelzési feladat kapcsán. McLeod et al. (1993) a hitelezési alkalmazásokat vizsgálták meg, West (2000) több különböző neurális hálózati paradigmára épülő modellt vetett össze hagyományos eljárásokkal. Egybehangzóan állítják, hogy a neurális hálózat a korábbi pontozási rendszereknél pontosabb előrejelzéseket képes

adni. Ráadásul a neurális hálózaton alapuló döntési modell kifejlesztése, karbantartása olcsóbb és gyorsabb, mint amit a korábbi rendszereknél tapasztaltak.
A módszer az előnyei ellenére valószínűleg azért nem annyira elterjedt, mert a döntéshozók nem értik a működését, számukra fekete dobozként üzemel.

2.7 Szakértői rendszerek

A szakértői rendszer általánosságban olyan eljárások gyűjteménye, amelyek egy szakértő döntéshozási viselkedését próbálják utánozni.

A szakértői rendszereknek négy alapvető ismérve van:

- a rendszer tudásbázison alapul;
- vannak eszközei a tudásbázis karbantartására és bővítésére;
- következtetni tud;
- döntéseit képes magyarázni.

Az alapot adó tudásbázis nem csak adatokat, tényeket, hanem szabályokat is tartalmaz arra vonatkozóan, hogy miképpen kell a tudást feldolgozni. A szabályok matematikai logikai összefüggések formájában adhatók meg. Mind a rendszerbe beépített logika, mind a szabályok szintaxisa sokféle lehet. Ezek megfelelő megválasztásától nagymértékben függ a rendszer hatékonysága és a leírható ismeretek mélysége.
A szabályok általában „HA…., AKKOR…" formájúak. Az alábbiakban felsorolunk néhány lehetséges szabályt, amelyek hitelbíráló szakértői rendszerekben előfordulhatnak:

- ha az illető korábban már kért hitelt, de bebizonyosodott, hogy nem valós adatokat szolgáltatott, akkor automatikusan utasítsd vissza;
- ha az éves törlesztőrészlet meghaladja az éves fizetés 50%-át, akkor utasítsd vissza;
- ha a kérelmezőnek korábban volt hitelügylete, és nem volt vele semmiféle fizetési probléma, akkor növeld 10 ponttal az összpontszámát[21];
- ha a kérelmező összpontszáma kisebb, mint $\mathbf{c}_{\text{alsó}}$, akkor utasítsd vissza a kérelmet, ha $\mathbf{c}_{\text{felső}}$ fölé esik, add meg automatikusan, egyébként töltesd ki a kiegészítő űrlapot (ahol $\mathbf{c}_{\text{alsó}}$ és $\mathbf{c}_{\text{felső}}$ értékét előzetes számítások illetve becslések adják);

[21] Itt az eddigiekkel ellentétben a magasabb pontszám jelöli a jobb ügyfeleket (kisebb bedőlési valószínűséget), azért, mert a szakértői rendszereknél ez az elterjedt megoldás.

A szakértői rendszerek létrehozásánál több akadállyal is szembe kell nézni az alkotóknak. Az egyik az, hogy nem minden tudást lehet szabályokkal, vagy egyéb formális módszerekkel reprezentálni. Például az egyszerű hétköznapi logika, „a józan paraszti ész" nem írható le ilyen módon, mert túlságosan általános és sokrétű, holott majdnem minden ember rendelkezik vele. Ezért a szakértői rendszerek csak azokon a területeken tudnak eredményesen működni, amelyek eléggé szűkek ahhoz, hogy jól le lehessen írni őket, ugyanakkor elegendően bonyolultak ahhoz, hogy ilyen eszközre szükség legyen. Emellett a rendszer létrehozásához és karbantartásához szükség van az adott terület szakértőire, akiknek a tudásából ki lehet indulni. Fontos követelmény, hogy ezek a specialisták olyanok legyenek, akik között a téma alapvető kérdéseiben egyetértés van.

A hitelkérelmek elbírálása, az ügyfélminősítés például ilyen terület. Pau (1986) és Holsapple et al.(1988) vizsgálta a szakértői rendszerek pénzügyi menedzsmentben, mindenek előtt a hitelezésben történő alkalmazhatóságát. Brooks (1989) összegezte az első szélesebb körű alkalmazási tapasztalatokat, amelyeket a személyi és a jelzáloghitelek elbírálásánál szereztek.

2.8 A módszerek hiányosságai

A jelenleg használatos credit scoring modellek több korláttal is rendelkeznek. Az egyik legfontosabb probléma az, hogy e modellek *elsősorban a hitelügylet nemfizetési valószínűségén* alapulnak, amely csak az egyik, bár tény, hogy a legfontosabb dimenziója az ennél sokkal összetettebb profitmaximalizálási döntési feladatnak. A másik komoly fogyatékossága e modelleknek, hogy legtöbbjük statikus, azaz csak az adott ügyletre koncentrál, *nem képes figyelembe venni a jövőben várható, a jelen ügylet által indukált újabb termékértékesítési lehetőségeket*[22], és így nem képes valóban maximalizálni a profitot. Ezt az „egyperiódusú" szemléletet többször próbálták már javítani, de még nem sikerült kielégítően megoldani a jövőben várható újabb ügyletek előrejelzésének problémáját.

Egy másik problémája a jelenlegi scoring rendszereknek, hogy azt a kérdést akarják megválaszolni, hogy „Mi a valószínűsége, hogy a kérelmező csődbe megy (vagy nem fizet) a jövőben egy adott időpontig?" Pedig, ha a credit scoring-tól szeretnénk a profit

[22] Vagy a mostani elutasítás miatti esetleges veszteségeket, hiszen lehet, hogy ennek hatására a már meglévő ügyfél más bankhoz viszi a jövedelmező hiteleit, megtakarításait is.

scoring irányába mozdulni, akkor az is nagyon fontos kérdés, hogy, ez *a nemfizetés mikor fog bekövetkezni*. Erre a kérdésre a túlélési függvények alkalmazásával épített modellek adhatnak választ. Hasonlóképpen kevesen vizsgálják a hitel életciklusában rendszeresen felmerülő revíziók során, hogy mekkora az a törlesztési késedelem, amely mellett egy adott ügylet még nyereséges tud maradni. Amennyiben a minősítő rendszer kínálna ilyen jellegű információkat a döntéshez, sok, majdnem sikertelen ügyletet lehetne szerényebb profit mellett, de veszteség nélkül zárni.

Fontos hiányossága még a jelenlegi hitelelbíráló rendszereknek, hogy többségük döntési javaslataik meghozatalakor az adott ügyletet kiragadva vizsgálja. Pedig kívánatos lenne a hitelnyújtó teljes portfoliójának figyelembe vétele a pénzintézet jobb eredő teljesítményének érdekében. *Portfolió szintű credit scoring* esetén a teljes portfolió kockázati helyzetét, és a kockázat diverzifikálására rendelkezésre álló lehetőségeket is figyelembe kell venni egy döntés meghozatalakor.

A negyedik, a fentiekkel összefüggő kérdés az, hogy az ügyfél *gazdasági környezetének jövőbeli várható változását*[23] is előre kellene jelezni a pontosabb értékelés érdekében. Ebbe az irányba tett kísérletek közé tartozik a Markov-láncok alkalmazása.

A mesterséges intelligencián alapuló rendszerek és a hagyományos (paraméteres és nemparaméteres) eljárások kombinációjával – a döntéstámogatásnak a jelenleginél minőségileg magasabb szintje érhető el. Az emberi gondolkodást szimuláló, a hitelügyintézők sokéves munkája során felhalmozott, szubjektívnek vélt, de valójában csak képletekkel ki nem fejezhető objektív tapasztalatok felhasználásával működő rendszerek a mainál lényegesen nagyobb biztonsággal fogják tudni a döntéshozatalt támogatni.

3. A klasszifikációs eljárások teljesítményének mérése

A scoring modellek építése során felmerül a kérdés, hogy mennyire jó az adott modell. Mivel a modellt a jó és rossz ügyfelek beazonosítására szeretnénk használni, jóságon azt a tulajdonságot értjük, hogy a modell mennyire képes megkülönböztetni a két csoportot.

[23] gazdasági ciklusok, demográfiai változások (elöregedő társadalom)

Amennyiben a modell besorolási pontosságát ugyanazon a mintán ellenőrizzük, amin megépítettük, akkor számíthatunk arra, hogy modellünk magyarázó erejét (illeszkedését) kedvezőbbnek fogjuk megítélni, mintha az ellenőrzést egy másik, az előzőtől független mintán hajtottuk volna végre, azaz alulbecsüljük a téves besorolás valószínűségét. Ez azért valószínű, mert a klasszifikációs modellbe beépülnek az adathalmaz olyan sajátosságai is, amelyek más adathalmazokban nincsenek. A mintát ezért érdemes kettébontani, az egyik minta szolgál a modellépítés céljaira (training), míg a másik minta (holdout sample) a modell prediktív erejének ellenőrzésére (validation). A módszer hátránya, hogy a kisebb minta következtében mintainformációt veszítünk. Ez a hitelezés területén általában nem jelent problémát, mert nagy adatbázisok érhetők el a múltbeli ügyfelekről. Lehet ellenőrizni a modellt egy későbbi időszak adatain is (out-of-time sample).

Előfordul azonban, hogy kevés a rendelkezésre álló adat, például egy új termék, vagy új ügyfélcsoport scoring modelljének építésénél. Erre az esetre vannak olyan módszerek, amelyek ugyanazon a mintán mérik a modell teljesítményét, amelyiken azt építették, de anélkül, hogy a téves besorolás valószínűségét alulbecsülnék. Ilyen pédául a Jackknife módszer, melynek lényege, hogy az n elemű mintából újabb $n-1$ elemű mintákat vesznek visszatevés nélkül. Azaz a paraméterek becslésénél mindig egy esetet elhagynak a mintából és a megmaradt adatokat használva becslik a paramétereket, majd az így elkészült modellt ellenőrzik a kihagyott eseten, azt vizsgálva, hogy vajon jól sorolja-e be azt. Az eljárás így folytatódik mindaddig, míg az összes eset sorra nem kerül. A Bootstrap módszer is hasonlóan ismételt paraméterbecslések sorozata, de itt az n elemű mintából újabb n elemű mintát vesznek visszatevéssel. A módszerek igen számításigényesek, de ez napjainkban már nem jelent igazi problémát.

A scorecardok teljesítményének mérésére szolgáló eszközöket három csoportba oszthatjuk (Mays, 2004): (1) szeparációs statisztikák, (2) rangsorolási statisztikák (3) előrejelzési hiba statisztikák. Az alábbiakban áttekintjük a szakirodalom által leginkább ajánlott -, illetve a credit scoring területén leggyakrabban használt módszereket, mérőszámokat.

A továbbiakban feltételezzük, hogy a modell által generált score értékek egyenesen arányosak a becsült bedőlési valószínűséggel, azaz a magasabb score nagyobb bedőlési valószínűséget, így rosszabb hitelkockázatú ügyletet jelöl[24].

3.1 Szeparációs statisztikák

Kolmogorov – Smirnov statisztika

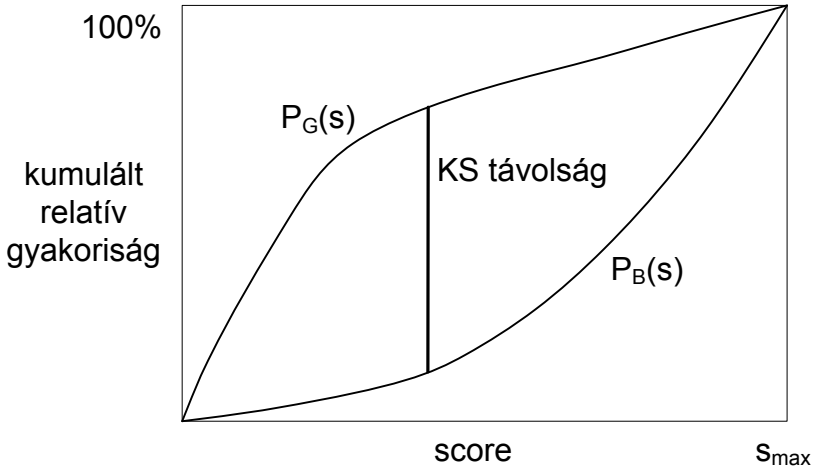
Azt méri, hogy a jók és a rosszak score-jainak eloszlásai milyen messze vannak egymástól. Formálisan:

$$KS = \max_s |P_G(s) - P_B(s)|.,\ \text{ahol}$$

$P_G(s) = \Sigma_{x \le s} p_G(x)$, *és* $P_B(s) = \Sigma_{x \le s} p_B(x)$

*(*folytonos score esetén integráljel értendő a szumma helyett)

A KS statisztika tehát a score szerint sorbarendezett sokaságban a jók és a rosszak eloszlásának kumulált relatív gyakorisága közötti maximális különbség.



7. ábra Kolmogorov-Smirnov távolság

A Kolmogorov – Smirnov statisztika a *szeparációs statisztikák* közé tartozik, amelyek csak egy pontban nézik a két csoport közötti távolságot, tehát nem ragadnak meg minden fontos tulajdonságot az eloszlások alakjával kapcsolatban. A KS statisztika is

[24] A gyakorlatban a bedőlési valószínűséggel mind az egyenesen, mint a fordítottan arányos score-ok alkalmazása is elterjedt, ezért tisztázzuk, hogy mi az előbbi megoldást választottuk.

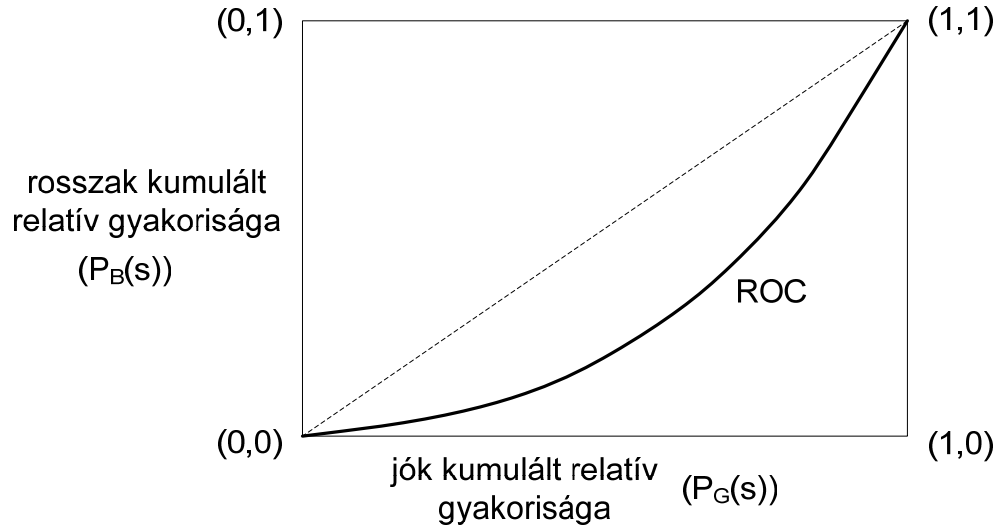
a scorecard egy általános tulajdonságát jeleníti meg, egyetlenegy pontban nézi a jó és rossz eloszlások távolságát (ott ahol ez a távolság maximális), míg a gyakorlatban inkább az a fontos, hogyan teljesít a scorecard a választott cutoff mellett.

3.2 Rangsorolási statisztikák

A rangsorolási statisztikák már a jó és rossz eloszlások teljes tartományára vonatkozó információt használnak:

AUROC (area under the ROC curve) a ROC görbe alatti terület

Az előző ábra a Kolmogorov – Smirnov statisztikát ábrázolta két görbe segítségével. Ugyanez az információ egy görbével is megjeleníthető, ha $P_B(s)$ –t ábrázoljuk $P_G(s)$ függvényében. Ezt mutatja a következő ábra. A görbe minden pontja valamilyen s score-hoz tartozik, a vízszintes koordinátája $P_G(s)$, a függőleges koordinátája $P_B(s)$.



8. ábra ROC-görbe

Ez a ROC (receiver operating characteristic) görbe[25], ami tulajdonképpen egy olyan Lorenz diagram, ami két kumulált relatív gyakorisági sort ábrázol egy görbével. Leírja a scorecard klasszifikációs tulajdonságát a cutoff score változása esetén. Az

[25] A ROC görbe elnevezése utal eredetére, mert eredetileg az üzenetek adása és vétele során előforduló klasszifikációs hibák becslésére használták.

elméletileg létező lehető legjobb scorecard ROC görbéje a négyzet oldalán fut (a vízszintes tengelyen az (1,0) pontig, majd a függőlegesen az (1,1) pontig). Ekkor az (1,0) pont egy olyan *s** score-hoz tartozik, ahol $P_B(s^*)=0$ *és* $P_G(s^*)=1$, azaz minden rossz *s**–nál nagyobb score-ral rendelkezik és minden jó annál kisebbel. A négyzet átlóján fekvő ROC görbe azt jelzi, hogy minden score esetén $P_G(s)= P_B(s)$, azaz a jók és rosszak aránya minden score esetén állandó. Ez pedig nem jobb, mint a véletlenszerű klasszifikálás, úgy, hogy ismert a sokaságon belüli jó – rossz arány. Tehát minél távolabb van a görbe az átlótól, annál jobb a scorecard. Ha egy scorecardnak olyan ROC görbéje van, amely mindig távolabb van a diagonálistól, mint egy másik scorecard görbéje, akkor az első jobban klasszifikál, mint a második, minden cutoff esetén. Általában azonban a ROC görbék metszik egymást, így az egyik scorecard jobban klasszifikál a cutoff értékek egy tartományában, míg a másik a cutoff értékek egy másik tartományában jobb.

Minél nagyobb tehát a *négyzet átlója és a ROC görbe által bezárt terület*, annál pontosabb a klasszifikáció. A *Gini mutató* ennek a területnek a kétszerese. A mutatóról elmondható, hogy: $0 \leq G \leq 1$. A mutató értéke tökéletes klasszifikáció esetén 1, véletlenszerű klasszifikálás esetén 0. A SAS által generált *C* és az SPSS által generált Area mutató értéke a (terület + 0,5).[26] A két mutató közötti összefüggés tehát: $G = 2(C - 0{,}5)$. Ezek a mutatók egy számban összesítik a scorecard teljesítményét a cutoff értékek teljes tartományára. Ez egyrészt frappáns megoldás, másrészt viszont félrevezető lehet, mert számunkra általában a lehetséges cutoff értékek csak egy szűk tartományán érdekes a scorecard teljesítménye.

3.3 Előrejelzésihiba-statisztikák

Hosmer-Lemeshow-statisztika

A Hosmer-Lemeshow statisztika az illeszkedés jóságát méri logisztikus regressziós modellek esetén. A HL statisztika azt teszteli, hogy mennyire képes a modell az adott score tartományokra megbecsülni a rosszak tényleges számát. A HL teszthez kapcsolódó táblázat értékei úgy keletkeznek, hogy az egyes esetek becsült bedőlési valószínűségeit növekvő sorba rendezzük, és az így keletkezett rangsort *k* (általában

[26] Ha az ábrán a tengelyeket felcseréljük, akkor valóban ez jelzi a görbe alatti területet. (Az átló és a görbe által bezárt terület + 0,5 (az alsó háromszög területe).)

10) egyenlő elemszámú csoportra bontjuk (kvantilisekbe (általában decilisekbe) rendezzük). Ezek után megvizsgáljuk, hogy az egyes csoportokba az egyes kategóriákból (1,0) hány megfigyelt (observed), és hány a regressziós becslés által várt (expected) eset tartozik. A Hosmer-Lemeshow statisztika értéke nem más, mint az erre a táblázatra alkalmazott Pearson féle *(k-2)* szabadságfokú χ^2 statisztika.

Ellentétben a fenti két mutatóval, a HL statisztika kisebb értékei jelentik a jobb klasszifikációt (nagyobb besorolási pontosságot).

A HL statisztika hibája, hogy nagyon érzékeny a kialakított csoportok számára. Ha túl sok kategóriát alakítunk ki, akkor kevés lehet az egyes kategóriába eső rosszak száma, így a modell nehezen tudja pontosan megbecsülni, a HL statisztika pedig gyenge teljesítményt fog jelezni. Ha túl kevés a kategória, akkor a modell könnyen ad jó becslést, így a HL statisztika szerint mindig jó a modell előrejelző képessége.

Klasszifikációs tábla (vagy konfúziós mátrix)

A klasszifikációs tábla, a helyes és a téves besorolásokat összefoglalóan, egy 2x2-es táblázatban jeleníti meg. A mátrix általános alakja:

klasszifikációs tábla		**valóságos kategória**		
		jó (G)	rossz (B)	
a modell által besorolt kategória	jó (elfogadás) (A)	n_{AG}	n_{AB}	n_A
	rossz (elutasítás) (R)	n_{RG}	n_{RB}	n_R
		n_G	n_B	n

3. táblázat Konfúziós mátrix

▪ Ekkor a ***hiba arány***: $(n_{AB} + n_{RG})/ n$.

A credit scoring területén előfordul, hogy ezt a hiba arányt úgy lehet minimálisra csökkenteni, ha mindenkit jónak klasszifikálnak (mivel a rosszak aránya általában kicsi). Természetesen nagy badarság lenne valóban így cselekedni.

Az itt elkövethető kétféle hiba, a jók rosszként való félreklasszifikálása, azaz elutasítása, illetve a rosszak jóként való félreklasszifikálása, azaz meghitelezése megfeleltethető a hipotézisvizsgálatban általánosan használt elsőfajú -, illetve másodfajú hibának. A kétféle hiba elkövetésének valószínűsége (α és β) csak egymás

rovására csökkenthető[27], ezért figyelembe kell venni a kétféle hiba elkövetésének igencsak eltérő költségeit.

Ha *D* (debt) jelenti azt a veszteséget, amit egy rossz ügyfél jóként való félreklasszifikálása (elfogadása) okoz, és *L* (lost profit) jelenti az elvesztett profitot, amit egy jó ügyfél rosszként való félreklasszifikálása (elutasítása) okoz, akkor

▪ a klasszifikálási hiba által okozott egy ügyfélre eső várható veszteség: $(L\ n_{RG} + D\ n_{AB})/\ n$.

A konfúziós mátrix nemcsak különböző modellek teljesítményének összehasonlítására használható, hanem a megfelelő cutoff érték kiválasztásához is.

Itt azonban meg kell állnunk egy kicsit. A *költség*ek figyelembevétele természetesen jobb, mintha csak a hibák számát minimalizálnánk, viszont a célunk a profit maximalizálás, nem pedig a költség minimalizálás. Sokkal jobb lenne a különböző esetek hasznait figyelembe venni. A *haszon*nak ugyanis van egy természetes viszonyítási alapja, amelyhez képest mérhetjük, legyen az pozitív, vagy negatív. Ez a viszonyítási alap a döntés meghozatala előtti helyzet. Ha helyes a döntés, akkor pozitív a haszna, ha nem helyes, akkor negatív.

Ha viszont (úgy mint fentebb) a költség fogalmában gondolkodunk, akkor könnyen készülhet olyan költségmátrix, amely logikailag ellentmondásos, mert nem minden értékének azonos a viszonyítási pontja. Gyakran láthatunk például ilyen költségmátrixot:

költség		**valóságos kategória**	
		jó (G)	rossz (B)
a modell által besorolt kategória	jó (elfogadás) (A)	0	D=10
	rossz (elutasítás) (R)	L=1	0

4. táblázat Hibás költségmátrix

A tényleges cashflow elutasítás esetén ugyanaz lesz, függetlenül attól, hogy valójában jó, vagy rossz ügyfélről van szó. Ezért minden észszerű költségmátrixban az *elutasítás* sorban lévő két értéknek egyeznie kell.

[27] Lásd például Hunyadi L. és Vita L.(2002)

A költségeket, vagy hasznokat mérhetjük bármilyen viszonyítási szinthez, de ennek a viszonyítási szintnek állandónak kell lennie. Az L elvesztett profit valójában egy elszalasztott lehetőség költsége, nem tényleges pénzkiáramlás. Könnyű elkövetni azt a hibát, hogy a különböző alternatív költségeket (opportunity cost) különböző szintekhez mérjük. A fenti hibás költségmátrix például magyarázható így: „ A jó ügyfelek meghitelezésének és a rosszak elutasításának nincs költsége, mert mindkét esetben jó döntést hoztunk. Ha egy jó ügyfelet elutasítunk, akkor a költség az 1 egységnyi elvesztett profit (L=1). Ha egy rossz ügyfelet meghitelezünk, akkor a veszteség a nyújtott hitel összege (D=10)"

Nézzük, miért hibás a fenti „gondolatmenet". Elsőként tegyük fel, hogy a banknak mind a négy típusból lesz egy ügyfele. Ekkor a bank eszközeiben beálló nettó változás -9. Most tételezzük fel, hogy jön négy jó ügyfél, akit meg is hiteleznek, ők pedig rendben visszafizetik a hitelt. Ebben az esetben az eszközökben beálló nettó változás +4. A két eset közötti különbség 13. Bármilyen viszonyítási alapot használunk is, ennyinek kell lenni a különbségnek. Ha viszont a fenti hibás költségmátrixot használjuk, akkor az első esetben a teljes költség 11, a másodikban pedig 0 (ezek különbsége pedig nem 13).

A hasznok megállapításához természetesen a jövedelmek és költségek nettó jelenértékét (NPV-jét) kell kifejezni.

Egy a fentieknek megfelelő haszonmátrix lehet a következő táblázatban szereplő. A fix 10 egységnyi hitelösszeg helyett beépítettük a hitelösszeg nagyságát is. Itt az x a hitelösszeg értéke eFt-ban. Feltételezzük, hogy a jó hiteleken a bank haszna 10%, nemfizetés esetén a várható veszteség 80%[28].

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

5. táblázat Haszonmátrix

Sajnos ezzel a megoldással kapcsolatban is vannak problémák:

- Azt látjuk, hogy elutasítás esetén nincs változás, márpedig a hitelelbírálásnak költsége van, akkor is, ha elutasítják a kérelmet.

[28] A nemfizetés esetén várható veszteséget (LGD- loss given default) valamilyen módon becsülni kell, értéke termékenként, ügyfélcsoportonként más-más lehet.

Ez igazából *nem probléma*, mert mind a négy értékből kivonhatjuk ezt a hitel-elbírálási költséget, de egy konstans hozzáadása minden értékhez nem változtatja az optimumhelyet.[29]

- A másik probléma, hogy a rosszak elutasítása egy helyes döntés, ezért valamilyen pozitív érték kéne, hogy társuljon hozzá, a jók elutasítása pedig hibás döntés, ezt negatív értékkel kellene kifejezni.

Elutasítás esetén végül is a bank haszna (bevétele) nem változik, ezért nem találtam megfelelő negatív, illetve pozitív értékeket.

A cutoff kiválasztása:

Elméletileg 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.

Ez akkor lenne igaz, ha a cél ezen az egy szegmensen elérhető profit maximalizálása, azaz, ha a bank döntési tere csak annyi lenne, hogy ide kihelyezi-e a pénzét vagy sem. Általában több helyre is kihelyezheti a tőkéjét, ezért inkább akkor érdemes befogadni a kérelmet, ha a befogadás hozama nagyobb, mint az alternatív befektetés(ek) várható hozama. Esetenként tehát inkább hozamban (nem profitban) érdemes gondolkodni, és azt maximalizálni. Ekkor viszont már felmerül a fix költségek allokálásának problémája (nem mindegy, hogy hány hitelre kell elosztani őket). A fix költségek elosztására nem létezik egyetlen legjobb eljárás. Az alkalmazandó (alkalmazható) megoldás függ a bank számviteli folyamataitól és konkrét üzleti céljaitól is.

A *gyakorlatban* általában a cutoff értékek lehetséges tartományán megvizsgálják a modellépítési vagy tesztelési mintán a különböző cutoff (score vagy becsült bedőlési valószínűség) értékekhez tartozó profit (vagy hozam) értékeket és azt a cutoff értéket választják, amely mellett a mintán maximális a profit (vagy hozam).[30]

Brier score:

$$\mathrm{BS} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{f}(1|x_i))^2$$

[29] Ha az elbírálási költségek eltérőek az elutasítottak és az elfogadottak között, akkor természetesen azokat be kell írni a táblázatba.

[30] A gyakorlatban sokszor az új scoring modell bevezetésénél az első időszakra olyan cutoff értéket választanak, ami hasonló beengedési rátát eredményez, mint a meglévő régi scorecard.

Itt y_i = 1, ha rossz hitelről van szó, és y_i = 0 ha jó, $\hat{f}(1|x_i)$ pedig annak a becsült valószínűsége, hogy az *i* objektum a rosszak közé tartozik. A mutató értékének elméleti minimuma 0 (tökéletes modell esetén, ami a jók esetén 0 bedőlési valószínűséget, a rosszak esetén 1-et becsül), maximuma 1 (épp ellentétes besorolás esetén).
Ezen mutatóval egyező elven működik a logaritmikus score:

Logaritmikus score:

$$\text{LS} = -\frac{1}{N}\sum_{i=1}^{N}(y_i \ln \hat{f}(1|x_i) + (1-y_i)\ln(1-\hat{f}(1|x_i))) = -\frac{1}{N}\sum_{i=1}^{N}\ln(|y_i + \hat{f}(1|x_i) - 1|)$$

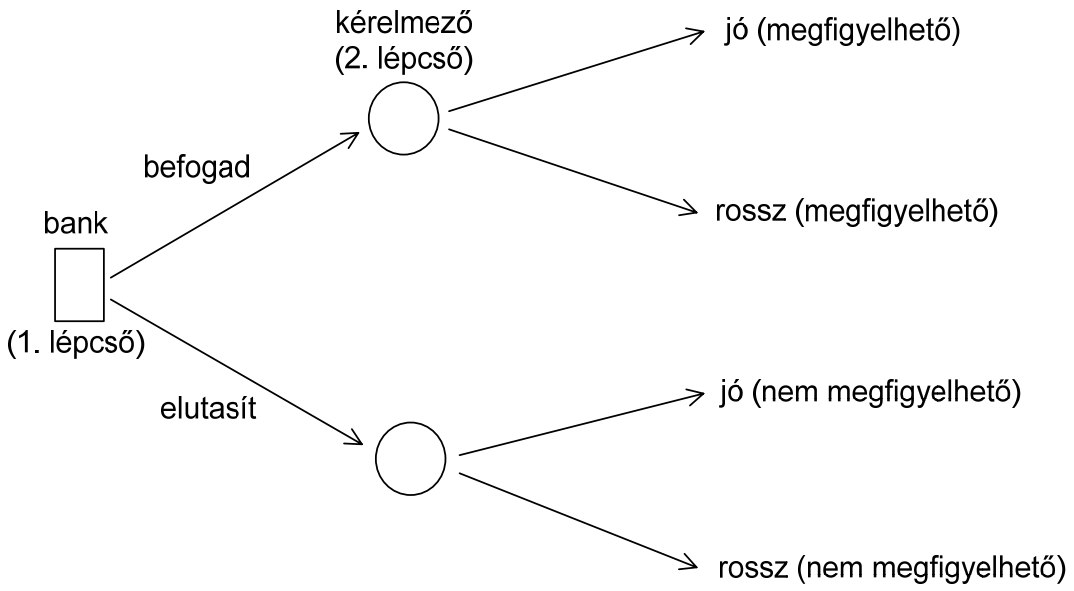
A mutató 0 közeli értékei jelzik a modell jó teljesítményét.
A Brier és a logaritmikus score is az előrejelzési hibát mérik, ezért a 0 közeli értékek jelentik a nagyobb besorolási pontosságot.

Ebben a fejezetben áttekintettük a leggyakrabban alkalmazott credit scoring modelleket és a modellszelekciós kritériumokat, így a következő fejezetben rátérhetünk a modellekben jelentkező szelekciós torzításra.

III. Módszerek a szelekciós torzítás csökkentésére

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 scoring modellek fejlesztésének egyik fő problémája éppen az, hogy csak azon ügyfelek teljesítéséről van múltbéli tapasztalati adat, 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 klasszifikációs eljárás esetén, amit ezen a mintán építenek, 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:



9. ábra A credit scoring esetén fellépő szelekciós torzítás kétlépcsős folyamata

Az első lépcsőben a bank eldönti, hogy meghitelezi-e az ügyfelet, vagy sem (A,R). A második lépcsőben megfigyelhető, hogy az ügyfél a jó vagy rossz kockázati csoportba (G,B) tartozik-e, de csak azoknál az ügyfeleknél, akiket meghiteleztek[31].
A scoring modellek építésénél fellépő szelekciós torzítás kiküszöbölésére számos módszert kipróbáltak és ajánlottak az utóbbi időkben, ezen technikák összefoglaló neve lett a *reject inference* (következtetés az elutasítottak felhasználásával). Ez tulajdonképpen azt jelenti, hogy megpróbáljuk megbecsülni, hogyan viselkedtek volna az elutasítottak, ha meghiteleztük volna őket, és az elutasítottakat is felhasználjuk a modellépítéshez, vagy az elfogadottakon épített modell kiigazításához.

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

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 adatokkal való statisztikai következtetés egy példájaként is. Ekkor a fenti csoportosítást megfeleltethetjük a Little és Rubin -féle

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

adathiány mechanizmusok három típusának: 1. teljesen véletlen adathiány (MCAR), 2. véletlen adathiány (MAR) és 3. nem véletlen adathiány (NMAR).

Nézzük meg ezt egy kicsit közelebbről!

Először is különböztessük meg a *szelekciós mechanizmust*, ami meghatározza, hogy egy ügyfelet beenged vagy elutasít a hitelező, és az *eredmény mechanizmust*, ami meghatározza, hogy az ügyfél jó vagy rossz. A szelekció tulajdonképpen az adathiány mechanizmus, hiszen meghatározza, hogy az ügyfélnél megfigyelhető-e az eredményváltozó (jó-rossz) értéke. A credit scoringban az elsődleges cél az eredmény mechanizmus modellezése. A hitelező szeretne egy javított, frissített befogadási/elutasítási szabályt, amit az új ügyfeleknél alkalmazhat. A továbbiakban tegyük fel, hogy $\mathbf{x} = (x_1, \ldots, x_k)$ a magyarázó változók vektora, ami 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, ha befogadnak egy ügyfelet és $a = 0$, ha elutasítanak. 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!

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 $\mathbf{x}$, sem az y értékétől. Azaz: $P(A|\mathbf{x},y) = P(A)$[32]

Ez azt jelenti, hogy a kérelmek elfogadása és elutasítása véletlenszerű (például pénzfeldobással döntik el ki kapjon hitelt). /Ez akkor történhet meg, ha így próbálnak

[32] Furcsának tűnhet, hogy feltételként szerepel a hitelkockázatot leíró változó (y), aminek értéke (jó vagy rossz) csak időben 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.

meg információt vásárolni az egyébként elutasítandókról./ A legtöbb hitelintézet azonban ettől szofisztikáltabb elbírálási rendszerrel rendelkezik.

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 cutoff érték.[33]

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-k eloszlása megegyezik. Ez a MAR feltételből következő fontos tulajdonság, amit 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 befogadás/elutasítás döntés meghozatalakor (override) azaz előfordul „kivétel ág”-on való beengedés vagy ügyintézői elutasítás is.

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

[33] Ez az s függvény a scoring modell, ami bármilyen lehet a korábban ismertetett eljárásoknak megfelelően (lineáris, logisztikus, klasszifikációs fa…).

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.

Vegyük észre, hogy amint elmozdulunk a MCAR esettől a MAR-on keresztül a NMAR felé, az *y* megfigyelhető értékeivel rendelkező meghitelezettek csoportja egy egyre inkább szelektált és nem jellemző csoport lesz a sokaságon belül, így a mintaszelekció problémája felerősödik (Schafer és Graham, 2002).

2. Szelekciós torzítást csökkentő technikák

Nézzük tehát, hogyan lehet felhasználni az elutasítottakról rendelkezésre álló részleges információkat a scoring rendszer 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 scorecard 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 egy 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 annak 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[34] kisebb. Így egy rétegzett mintát kapunk egyfajta költségoptimális mintaelosztással. Végül átsúlyozással kaphatunk egy a sokaságot valóban reprezentáló mintát anélkül, hogy vállalni 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, amin a scoring függvény épül, (MAR és NMAR adathiány mechanizmus), akkor a következő megoldásokat javasolja a szakirodalom:

[34] A várható veszteség egyenesen arányos a bedőlés valószínűségével és a hitelösszeg nagyságával és fordítottan arányos a fedezet nagyságával.

2.2. Módszerek MAR esetén

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

Ha az $\mathbf{x}_{\text{régi}}$ részhalmaza az $\mathbf{x}_{\text{új}}$ -nak, azaz minden jellemző ami alapján elfogadtak vagy elutasítottak egy kérelmet az most is elérhető, 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 érték kombinációi 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 befogadás/elutasítás döntés meghozatalánál, de nem rögzítették őket, így most nem elérhetők. (Ez már a NMAR eset)

2. 2.1 Augmentáció (vagy átsúlyozás)

Az augmentáció módszerét először Hsia (1978) vázolta fel. A módszer tulajdonképpen nem más, mint a hiányzó adatok kezelésénél ismertetett 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

$P(y = 1|\boldsymbol{x},A)$, 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 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. És az A_j beengedett ügyfélből G_j jó eset volt és B_j rossz. (Lásd az alábbi táblázatot.)

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	$(R_k + A_k) /A_k$

6. táblázat Átsúlyozás

Hsia ezután alkalmazza azt a feltételezést, hogy

$$P(B|\, s, R) = P(B|\, s, A) \qquad (1)$$

azaz a csőd valószínűsége adott *s* befogadás-elutasítás score mellett a befogadott és az elutasított kérelmeknél megegyezik[35]. Ahol

$$P(B|\, s, A) = \sum_{x;s(x)=s} p(B|\mathbf{x},A)\, P(\mathbf{x}|s(\mathbf{x})=s). \qquad (2)$$

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.

Így a j kategóriába eső A_j befogadott ügyfél akkora súlyt kap, hogy reprezentálja az A_j és R_j eseteket is, ez 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

[35] Ez azt is jelenti, hogy $P(G|\, s, R) = P(G|\, s, A)$,mert $P(G) = 1\text{-}P(B)$

(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, amiben 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, 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.

Hand és Henley (1993) is azt fogalmazták meg kritikájukban, hogy ha az $\mathbf{x}_{\text{régi}}$ nem részhalmaza az $\mathbf{x}_{\text{új}}$-nak, akkor az azonos valószínűség feltételezése torzítást okoz a csődvalószínűségek becslésében. Nézzük meg ezt egy kicsit részletesebben és formalizálva! Alkalmazott jelölésünk továbbra is:

y = (1,0) vagy (B,G) jelöli, hogy az ügylet rossz (1 vagy B), vagy jó volt (0 vagy G),

a = (1,0) vagy (*A,R*) jelöli, hogy az y megfigyelhető-e (1 vagy A a korábban beengedett ügyfeleknél) vagy hiányzik (0 vagy R a korábban elutasított ügyfeleknél).

Elmondhatjuk, hogy

$$P(B|\ \mathbf{x}_{új}) = P(B|\ A,\ \mathbf{x}_{új}) \cdot P(A|\ \mathbf{x}_{új}) + P(B|\ R,\ \mathbf{x}_{új}) \cdot P(R|\ \mathbf{x}_{új}) \qquad (3)$$

Ha $P(a\ )$ csak az $\mathbf{x}_{\text{új}}$ –tól függ, akkor $P(B|\ A,\ \mathbf{x}_{új}) = P(B|\ R,\ \mathbf{x}_{új})$, és a (3) egyenlet átírható

$$P(B|\ \mathbf{x}_{új}) = P(B|\ A,\ \mathbf{x}_{új}) \qquad (4)$$

De ha az $\mathbf{x}_{régi}$ nem részhalmaza az $\mathbf{x}_{új}$-nak, hanem tartalmaz még egyéb látens változókat, akkor $P(B|\ A,\ \mathbf{x}_{új}) \neq P(B|\ A,\ \mathbf{x}_{új})$, így az azonos valószínűség feltétel nem teljesül és torzított becsléseket kapunk.

Hand és Henley (1993) és Banasik et al (2001) is megmutatták, hogy ez az eset a Little és Rubin-féle nem véletlen adathiány (NMAR) egy példája. Ugyanis, ha az $\mathbf{x}_{régi}$ nem részhalmaza az $\mathbf{x}_{új}$-nak, hanem tartalmaz még látens változókat ($\mathbf{x}_{látens}$), akkor az A függ az $\mathbf{x}_{új}$ –tól és az $\mathbf{x}_{látens}$ –től. Ha ezek a változók nem teljesen irrelevánsak, akkor az $\mathbf{x}_{látens}$ hat a y-ra is. Így az a függ a $\mathbf{x}_{új}$ –tól és az y-tól is, azaz az adathiány függ

magától a hiányzó adatokat tartalmazó változótól is, ami a nem véletlen adathiány (NMAR) egzakt definíciója.

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 milyen típusú számláról van szó, mikor nyitották a számlát).
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) tanulmányukban azt találták, 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 illesztünk egy modellt a bedőlési valószínűségre 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 cutoff 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. (Lásd Ash és Meester, 2002.)

Természetesen az elutasítási tartomány[36] 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. Ekkor, ha 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 építhető az új scoring modell.

2.2.2.1 Az extrapoláció két lehetséges megközelítése

A célunk az eredmény mechanizmus megismerése, azaz annak 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)[37]. 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*[38]. Nézzük meg röviden mindkét megközelítést, mert reject inference esetén teljesen más következményeik lesznek. (Hand és Henley, 1993)

[36] Elutasítási tartományon itt azt a score (vagy becsült bedőlési valószínűség) tarományt értem, amely esetén az ügyfeleket elutasítják (azaz nem hitelezik).
[37] Ezt az elnevezést használja pédául Friedman (1997).
[38] Ezt az elnevezést használja például Elkan (2001)és Smith (2005).

A) 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 $\sigma_y^2(x) = f(\mathbf{x})(1 - f(\mathbf{x}))$.

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

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

ahol $\Lambda(.)$ a logisztikus eloszlás függvény. 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. Ezt láthatjuk, ha meggondoljuk, hogy a különböző megfigyelések hogyan járulnak hozzá a likelihood függvényhez. Egy n elemű FAE minta esetén a likelihood : $L = \Pi L_j$, ahol

$L_j =$ $p(y = i|\mathbf{x}_j)$, ha $y_j = i$ $(i = 0,1)$

$\sum_{i=0}^{1} p(y = i|\mathbf{x}_j)$ ha y_j hiányzik.

Ha az y_j hiányzik, akkor a nem informatív 1 szorzófaktorral járul a likelihood értékéhez, azaz nem változtatja. / $\sum_{i=0}^{1} p(y = i|\mathbf{x}_j) = 1$, mert $p(y = 1) = 1 - p(y = 0)$ / Tehát az elutasítottak modellbe építése ugyanolyan likelihoodot, így ugyanolyan becslést eredményez, mintha egyáltalán nem foglalkoznánk velük.

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.

B) Közvetett becslés (Density estimation)

Az f (**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 $\hat{p}_i$(x | Ti) feltételes eloszlást és a $\hat{\pi}_i$ prior valószínűséget. Aztán ezekből a Bayes tétel alapján kaphatjuk meg az $\hat{f}$(**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 (Avery, 1981).

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 komponens eloszlásokat.

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 diszkriminancia analízis). 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 diszkriminancia analízis) 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 torzítás komponense nem csökken, a variancia komponense 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.

Ha azonban az $\mathbf{x}_{új}$ -on kívül vannak egyéb változók is, amelyek hatnak a befogadás és a bedőlés valószínűségére akkor újfent elmondhatjuk, hogy nemvéletlen adathiánnyal (NMAR) van dolgunk, így az extrapoláció a posterior valószínűségek torzított becslését eredményezi.

Ha az $\mathbf{x}_{régi}$ részhalmaza az $\mathbf{x}_{új}$-nak és a (3) egyenlet teljesül (a Little és Rubin -féle véletlen adathiány esete (MAR)), még mindig számolnunk kell egy másik hibaforrással a valószínűségek becslésénél. Nevezetesen azzal, hogy a mintabeli (elfogadottakon belüli) jó/rossz arány nem egyezik a sokasági (elutasítottakat is tartalmazó) jó/rossz aránnyal. Ekkor, ha olyan cutoff valószínűséget választunk, amely egyenlővé teszi a prediktált rosszak számát a mintában ténylegesen előforduló rosszak számával, akkor ez eltér attól a cutoff valószínűségtől, ami egyenlővé teszi a tényleges és a becsült rosszak számát a teljes (ajtón bejövő) sokaságban. Így, bár a modell torzítatlanul becsli a posterior valószínűségeket, de az ügyfeleket félreklasszifikálhatjuk a nem megfelelő cutoff miatt.

2.2.3 Keverék eloszlások

Elméleti háttér

A keverék eloszlások olyan eloszlások, amelyek kifejezhetők más eloszlások „súlyozott átlagaként”. (McLachlan és Basford,1988), (Everitt és Hand, 1981)

Egy véges keverék általános felírása:

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

ahol c a komponensek számát, π_i a keverési súlyokat és θ_i a komponens paraméter vektorokat jelöli.

Itt feltételezzük, hogy a komponens eloszlások 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.

A megfigyelt y-nal (a komponens ismert) és a hiányzó y-nal (a komponens ismeretlen) rendelkező esetek hozzájárulása a likelihoodhoz:

$$\begin{aligned} L_j = \quad & \pi_i p_i(\mathbf{x}_j) && \text{ha } y_j = i \ \ (i = 0,1) \\ & p(\mathbf{x}_j) = \Sigma\, \pi_i p_i(\mathbf{x}_j) && \text{ha } y_j \text{ hiányzik.} \quad i=0,1 \end{aligned}$$

Ha m elutasított és n befogadott hitelünk van, akkor a megfigyelt adatok likelihoodja felírható:

$$L(\Psi) = \prod_{j=1}^{m} \left\{ \sum_{i=0}^{1} \pi_i p_i(x_j;\theta_i) \right\} \prod_{j=m+1}^{m+n} \left\{ \sum_{i=0}^{1} z_{ij} \pi_i p_i(x_j;\theta_i) \right\}$$

ahol $\boldsymbol{\psi} = (\pi',\theta')'$ jelöli a nem ismert paraméterek vektorát és $z_{ij} = 1$, ha a j megfigyelés esetén

$y_j = i$ egyébként 0.

A számítások megkönnyítése érdekében gyakran kifejezzük a loglikelihoodot:

$$\log L(\Psi) = \sum_{j=1}^{m} \log \left\{ \sum_{i=0}^{1} \pi_i p_i(x_j;\theta_i) \right\} + \sum_{j=m+1}^{m+n} \left\{ \sum_{i=0}^{1} z_{ij} \log(\pi_i p_i(x_j;\theta_i)) \right\}$$

Ez általában nagyon bonyolult függvénye **ψ**-nek, és a maximum likelihood becslés megtalálásához speciális számítási algoritmusra lehet szükség. Ilyen lehet például az EM algoritmus, melynek menetét korábban (II.rész) ismertettük.

Alkalmazások, feltételek, eredmények

A keverék eloszlások megközelítés szerint tehát feltételezhetjük, hogy a sokaság két eloszlás keverékéből származik – a jók és a rosszak eloszlásából – és hogy ezen eloszlások típusa ismert. Ezt a megközelítést alkalmazta Feelders (1999). Ha például az **x** jellemzőkkel rendelkezők aránya $p(\mathbf{x})$, 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 $p(\mathbf{x}|G)$ és $p(\mathbf{x}|B)$ paraméterei teljesen specifikálják a jobb oldalt. Ezek után olyan paramétereket kell választani, amelyek minimalizálják a két oldal közötti különbséget.

A $p(\mathbf{x}|G)$ és a $p(\mathbf{x}|B)$ paramétereinek becsléséhez használhatók az elfogadottak és -az EM algoritmus segítségével- az elutasítottak is.

Szokásos feltételezés, hogy $p(\mathbf{x}|G)$ és a $p(\mathbf{x}|B)$ 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 *feltételezések*kel, amelyek ha helyesek, akkor 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, ami alapján korábban elutasították a kérelmezőt. Ez azonban egy 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 tényleg 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 őket a modellépítésbe. A legrosszabbnak tűnő eseteket kiválaszthatjuk az eddigi scoring függvény alapján (legmagasabb pontszámú egyedek), vagy egyéb negatív információ alapján (KHR[39] lista, 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 $P(y=1|\mathbf{x})=1$ a mintatér egy jelentős részére, amiről tudjuk, hogy nem igaz, és eltorzíthatja a modellt az elfogadott kérelmekre is.

[39] Központi Hitelinformációs Rendszer, a korábbi Bankközi Adós- és Hitelinformációs Rendszer (BAR) új elnevezése

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, 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, ami 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 tartották a legcélravezetőbbnek, ők ezt „kalibráló mintának" nevezték.

Ha tudjuk, hogy a kalibráló minta véletlen kiválasztás eredménye, akkor egyszerűen kombináljuk az elfogadottakkal és egy olyan statisztikai technikát használunk, ami 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) 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 hitel termékekről, 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). Aztán, ha megvizsgáljuk ezen jellemzők 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 és az elfogadottak és az elutasítottak értékei közötti kapcsolatnak az ismeretében, megbecsülhetjük az elutasítottak jellemző értékét.

b) A módszerhez csak egy kalibráló mintára van szükség. Első lépésként építünk egy scorecard-ot (1) az új mintán, ami csak az elfogadottakat tartalmazza. Aztán a kalibráló mintából csak elfogadottakon építünk egy modellt (2), majd a teljes kalibráló mintán is építünk egy scorecard-ot (3). Így a két kalibráló modell (3-2) eltérése felhasználható az új scorecard (1) kiigazításához.

Egy egyszerű példaként tegyük fel, hogy lineáris regresszióval készítünk scorecard-ot. Legyenek az új elfogadottakra készített regresszió (1) paraméter vektora: $\boldsymbol{\alpha}$: $[\alpha_1,\ldots,\alpha_n]$, 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-dik 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.

c) Szintén egy kalibráló mintára van szükség a keverék eloszlások megközelítésű modell javításához. A kalibráló mintából ismerjük az ügyfelek egy részének valóságos jó-rossz osztályát (nem csak az elfogadottakét). Ezt az információt felhasználhatjuk, amikor $p(\boldsymbol{x}|G)$ 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 a szimplán a befogadottakon épített modell.

2.3.3 Speciális logit

Néhány reject inference módszert speciálisan a credit scoring-nál alkalmazott logisztikus regresszióhoz ajánlottak. Joanes (1993) olyan kiigazított posterior valószínűségeket származtatott, amelyek figyelembe veszik a jó-rossz csoportba tartozás a priori valószínűségét és a félreklasszifikálás különböző költségeit is. Aztán iteratív újraklasszifikálás következik, amely módosított paraméterbecsléseket eredményez (az elutasítottak különböző arányú szétosztása következtében). Bár a módszer az augmentációs eljáráson alapul, ami használja azt a feltételezést, miszerint

a jók aránya az elutasítottak és az elfogadottak között megegyezik, az iteráción keresztül ez a megszorítás bizonyos mértékig lazítható. Itt sem tudjuk azonban, hogy mi a relatív előnye ennek a korrekciós eljárásnak, mivel egy nem tesztelt feltételezésen alapszik.

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.

2.3.4 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 (csőd modell) leírható egy kétlépcsős modellel, részleges megfigyelhetőséggel.
Ezzel a megoldással foglalkozott Poirier (1980), Van de Ven és Van Pragg (1981) pedig formalizálták. Meng és Schmidt (1985) foglalkoztak a modellben lévő részleges megfigyelhetőség költségével. Copas és Li (1997) további kutatásokat végeztek a nem véletlen mintákból való következtetésekkel kapcsolatban az eljárás kiterjesztésével. Más kutatók (Boyes et al. 1989, Greene 1998, Jacobson és Roszbach 1999) alkalmazták ezt a modellt. A tanulmányok megmutatták, hogy szignifikáns szelekciós torzítás jelentkezett a csak a meghitelezetteken épített scoring modelleknél. A Heckman modell alkalmazhatósága nagyon erősen támaszkodik a két egyenlet (hiteldöntés és csődmodell) teljes specifikálására.

Nézzük meg ezt a modellt!

A credit scoring esetén fellépő szelekciós torzítás modellezhető egy kétlépcsős folyamatként, amint azt már korábban láthattuk (62. o.)

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 beazonosítsa a várhatóan jó ügyfeleket.

Alkalmazzuk a *kétváltozós probit modellt minta szelekcióval.* 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 (csőd egyenlet), ahol

y_i = 1 csőd (rossz hitel) esetén ($y_i^* \geq 0$)

0 nem csőd (jó hitel) esetén ($y_i^* < 0$)

A szelekciós egyenlet:

$$a_i^* = \mathbf{z}_i \alpha + \varepsilon_i$$ [40]

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 csőd egyenlet 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 megmutatja, 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 logit modellek direkt alkalmazása a csődegyenletben torzított paraméterbecslésekhez vezet. Meng és Schmidt (1985) megállapították, hogy a részleges megfigyelhetőség költsége a kétváltozós probit modellnél igen magas, ezért ha lehetséges, érdemes pótlólagos információkat is beszerezni. A credit scoring

[40] A szelekciós egyenletben azért jelöltem a magyarázó változókat **z**-vel, mert nem feltétlen egyeznek meg a csődegyenletben szereplő **x**-ekkel.

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 (Poirier, 1980). 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:

$$\begin{aligned} \ln L(\alpha,\beta,\rho) = \; & \Sigma(1 - a_i) \ln(1 - \Phi(\mathbf{z}_i\alpha)) + \\ & \Sigma\, a_i\,(1 - y_i) \ln(\Phi(\mathbf{z}_i\alpha) - \Phi_2(\mathbf{z}_i\alpha, \mathbf{x}_i\beta; \rho)) + \\ & \Sigma\, a_i\, y_i \ln(\Phi_2(\mathbf{z}_i\alpha, \mathbf{x}_i\beta; \rho)) \end{aligned}$$

Ezt maximálva kapjuk meg a modellek paramétereinek ML becslését.

Ha az egyenleteket sikerült jól specifikálni (ez fontos feltétel és nem biztos, hogy teljesül!) és 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.

Másrész, 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

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

azaz 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 találtak a két hibatag között. (A talált ρ értékek rendre: + 0,9234; + 0, 353 és + 0,1178)[41]. Jacobson és Roszbach (1998) arra a következtetésre jutottak, hogy a vizsgálatba bevont bankok nem akarták minimalizálni a bedőlési kockázatot. Ezt nemcsak a pozitív korreláció alapján gondolták, véleményüket alátámasztotta az a tény is, 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 tudták 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, mert 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.

[41] Az alkalmazandó legjobb reject inference módszer esetenként más-más lehet. A különböző vizsgálatokban nagyon eltérő eredmények adódtak, ez is azt jelzi, hogy az adatbázisok nagyon 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.

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 (2002) 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. Chen és Astebro (2001) azt találták, hogy nagyon költséges lehet a szelekciós torzítás figyelmen kívül hagyása, de azt is megállapították, hogy a Heckman eljárás -bár elméletileg egy jól hangzó technika- nem képes megfelelően kontrolálni ezt a torzítást, megbízhatatlan és nagyon érzékeny, ráadásul támaszkodik a normalitásra, ami gyakran nem teljesül. Az ő eredményeik is megerősítették Hand és Henley (1993) véleményét, miszerint egy megbízható modellhez valóban érdemes pótlólagos információkat beszerezni az elutasítottak teljesítményéről.

2.3.5 Három csoport

A mintát három csoportra oszthatjuk: jók, rosszak és elutasítottak. Ebbe a három csoportba való klasszifikálást javasolta például Reichert, Cho és Wagner(1983). 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 diszkriminancia analízis esetén, ha három csoportba klasszifikálunk, akkor feltételezzük, hogy mindhárom csoportnak közös kovariancia mátrixa van. Így ez egy módja lehet annak, hogy felhasználjuk az elutasítottakat is a kovariancia mátrix becslésének javítására. (Igazából ez az előny is kérdéses credit scoring esetén, mert ez a közös kovariancia feltétel nem valószínű, hogy tartható, hiszen az elutasítási döntés, ami a csoportokat képzi, korrelál az ügyfelek megfelelő jellemzőivel.)

2.3.6 Bayes-i határ és összecsukás (Bound and Collapse)

Legyen továbbra is $y = j$ a hitel kockázat kimenetele (j = 0:jó ; j = 1:rossz), $s = i$ ($i=1,...,r$) pedig a credit score. (Vannak olyan scoring alkalmazások, 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 ($i=1,...,r$).)

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, a 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, amelyet *Határ és Összecsukás (Bound and Collapse)* -ként neveztek el. 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.[42] 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 Bayes-i alapú[43] 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.

[42] 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.)

[43] A Bayes-i gondolat az adathiány mechanizmusról meglévő információk, vagy feltételezések beépítése a modellbe.

Nézzük Chen és Astebro alkalmazását!

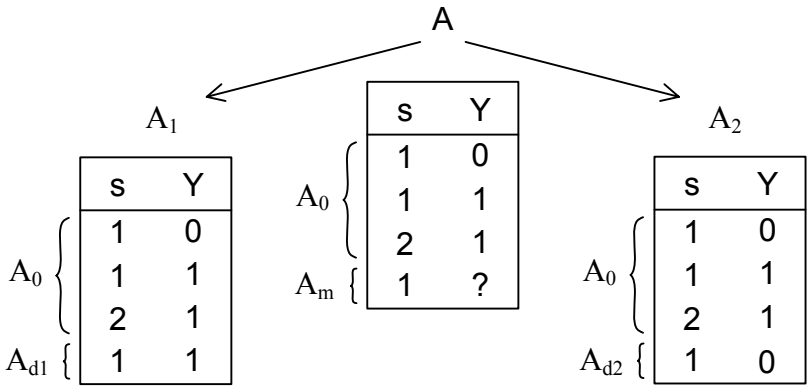
Feltételezték, hogy *(s,y)* multinomiális eloszlású, $\theta_{ij} = P(s = i, y = j|\theta)$ valószínűségekkel, ahol $\theta =(\theta_{10} , \theta_{11} , \theta_{20} , \theta_{21} ,..., \theta_{r0} , \theta_{r1}) = (\theta_{ij}), (\theta_{ij} \geq 0$, *minden i,j-re és* $\Sigma_{ij}\theta_{ij} = 1$) parametrizálja az együttes eloszlást.

/A standard konjugált prior θ- ra egy Dirichlet eloszlás $D(\alpha), \alpha =(\alpha_{10} , \alpha_{11} , \alpha_{20} , \alpha_{21} ,..., \alpha_{r0} , \alpha_{r1})$, eloszlása:

$$p(\theta) = \prod_{i=1}^{r} \prod_{j0,1} \frac{\Gamma(\alpha)}{\Gamma(\alpha_{ij})} \theta_{ij}^{\alpha_{ij}-1} \qquad (1)$$

ahol $\alpha_{ij} \geq 0$, minden *i,j*-re és $\alpha = \Sigma_{ij}\alpha_{ij}$ /

Ha nincs adathiány, akkor a Bayesi analízis nem okoz nehézséget. Credit scoring esetén azonban, *y* értéke nem ismert a korábban elutasított ügyfeleknél (egyszerű esetben az olyan ügyfeleknél, ahol az *s* credit score értéke nagyobb, mint egy *h* (cutoff) érték). Legyen $A = (A_0, A_m)$, ahol A_0 jelenti a minta teljes megfigyeléseket tartalmazó részét (meghitelezettek), A_m pedig a hiányzó y-nal rendelkező eseteket (elutasítottak). Legyen A_k az A_0 lehetséges kiegészítése. $A_k = (A_0, A_{dk})$, ahol A_{dk} az A_m-ben lévő hiányos esetek egy lehetséges eloszlása. A következő ábra mutatja egy megfigyelt minta két lehetséges kiegészítését.



10. ábra A megfigyelt minta lehetséges kiegészítései

Legyen n_{ij} a teljesen megfigyelt, adathiányt nem tartalmazó ($s =i, y = j$) esetek gyakorisága, m_i az ($s = i, y = ?$) esetek száma, és $n = \Sigma_{ij}n_{ij}$ a teljesen megfigyelt esetek száma, $m = \Sigma_i m_i$ a részlegesen megfigyelt, adathiányt tartalmazó esetek száma, ($n + m$) pedig a mintanagyság.

Little és Rubin (1987) nyomán megjeleníthetjük ezt a mintát egy *r* x ($c + 1$) –es kontingencia táblában, ahol $c = 2$. A ($c + 1$)-dik oszlop tartalmazza az adathiányos esetek gyakoriságát minden $s = i$ kategóriára.

s	*y*		
	0	*1*	*?*
1	n_{10}	n_{11}	m_1
2	n_{20}	n_{21}	m_2
:	:	:	:
h	n_{h0}	n_{h1}	m_h
:	:	:	:
r	n_{r0}	n_{r1}	m_r

7. táblázat Az adathiány megjelenítése

Sebastiani és Ramoni (2000) szerint θ posterior eloszlása Dirichlet eloszlások keveréke, ahol a komponens eloszlások súlya *A* lehetséges kiegészítéseinek valószínűsége. Megmutatták, hogy még az adathiány mechanizmusra vonatkozó külső információ nélkül is megadhatók a lehetséges becslések határai, amelyek konzisztensek a mintából elérhető információkkal. Aztán ha lesznek információink az adathiány mechanizmusról, akkor a lehetséges becslési értékekből kiválasztható egyetlen érték.

Legyen φ az adathiány mechanizmus. A $p(y = j|\ s = i)$-re vonatkozó lehetséges becslések határai, adott *A* minta esetén:

$$p_{\min}(j|i) = \frac{\alpha_{ij} + n_{ij}}{\alpha_{i+} + n_{i+} + m_i} \le p(y = j|s = i|A) \le \frac{\alpha_{ij} + n_{ij} + m_i}{\alpha_{i+} + n_{i+} + m_i} = p_{\max}(j|i) \qquad (2)$$

ahol $\alpha_{i+} = \Sigma_j \alpha_{ij}$, és $n_{i+} = \Sigma_j n_{ij}$.

Tegyük fel, hogy van az adathiány mechanizmusra vonatkozó külső információnk, amelyből tudunk készíteni a nemválaszolásra egy valószínűségi modellt:

$P(y = j|\ y = ?, s = i, \varphi, \theta) = \varphi_{j|i}$, (3)

ahol $\sum_j \varphi_{j|i} = 1$, minden i-re.

Ez az információ aztán használható egy pontbecslés kiválasztására a $[p_{min}(j|i), p_{max}(j|i)]$ valószínűségi intervallumon belül, az extrém valószínűségek egy konvex kombinációjával:

$$\hat{p}_{j|i} = \varphi_{j|i} p_{\max}(j|i) + (1 - \varphi_{j|i}) p_{\min}(j|i) = \frac{\alpha_{ij} + n_{ij} + \varphi_{j|i} m_i}{\alpha_{i+} + n_{i+} + m_i} \qquad (4)$$

Ahhoz, hogy használni tudjuk ezt a módszert a credit scoring esetén fellépő szelekciós torzítás csökkentésére, explicit módon kell kifejeznünk a nemvéletlen adathiány

mechanizmust leíró (3) egyenletet. Ez a feladat azonban nagyon adatbázis függő, ezért ezzel itt nem foglalkozunk. Csak egy általános módszert mutatunk a nemvéletlen adathiány leírására reject inference esetére.

A módszer egy egyszerű alkalmazásánál a hitelek beengedése csak az eredeti scoring modellen (*s* függvény, *h* cutoff értékkel) alapul, és nincs más beengedési szabály. Ekkor az előző táblázat így néz ki:

s	*y*		
	0	*1*	*?*
1	n_{10}	n_{11}	*0*
2	n_{20}	n_{21}	*0*
:	:	:	:
h-1	n_{h-10}	n_{h-11}	*0*
h	*0*	*0*	m_h
h+1	*0*	*0*	m_{h+1}
:	:	:	:
r	*0*	*0*	m_r

8. táblázat Az adathiány speciális megjelenése

és a (4) egyenlet speciálisan:

$$\hat{p}_{j|i} = \varphi_{j|i} p_{\max}(j|i) + (1-\varphi_{j|i}) p_{\min}(j|i) = \frac{\alpha_{ij} + \varphi_{j|i} m_i}{\alpha_{i+} + m_i} \qquad (5)$$

lesz azokra a score kategóriákra, ahol nincs az eredményváltozónak megfigyelt értéke. Az (5) egyenletből látható, hogy a minta adatok nem szolgáltatnak érvényes információt az y hiányzó adatainak imputációjához, ha $h \le s \le r$. Ennek oka, hogy a hiányzó *y* becsült valószínűségét az *y* prior eloszlása és az adathiány mechanizmus is meghatározza.

A modell alkalmazásához kapcsolódó kérdések

A módszer alkalmazásához látnunk kell, hogyan lehet becsülni az adathiány mechanizmust és hogyan kell kiválasztani a prior eloszlást.

A NMAR adathiánnyal foglalkozó kutatások fundamentálisan eltérő megközelítésük alapján két csoportra bonthatók: *szelekciós modellek* és *mintázat-keverék (pattern-mixture) modellek*.[44] Ezek a modellek az együttes valószínűséget eltérő módon bontják fel. A szelekciós modellek a $P(y_{hiányzó}, y_{megfigyelt}) = P(y_{hiányzó} | y_{megfigyelt})$

[44] Amint azt már az I. fejezetben a modell alapú eljárásoknál említettük.

$P(y_{megfigyelt})$ felbontást használják. Ennek egyik példája a Heckman-féle kétváltozós kétlépcsős modell. A mintázat keverék modellek pedig a $P(y_{hiányzó}, y_{megfigyelt}) = P(y_{megfigyelt}| y_{hiányzó}) P(y_{hiányzó})$ felbontást alkalmazzák. A mintázat keverék modellek osztályozzák a hiányzó adatokat a hiányzás alapján és minden hiányzási osztályon belül leírják a megfigyelt adatokat. Az $y_{hiányzó}$-ra vonatkozó információ nélkül az eloszlásnak semmilyen jellemzőjét nem tudnánk, és a modell nem lenne identifikálható. Ezért ezekhez a modellekhez szükség van a hiányzó adatokra vonatkozó információra.

Ez a *BC modell* szorosan kapcsolódik a mintázat-keverék modellekhez, hiszen a hiányzó adatokat tartalmazó megfigyelések csoportosítva vannak a különböző hiányzási mintákat tartalmazó osztályokba az *s* credit score-on keresztül. Ez a modell leírja a megfigyelt adatok és a hiányzás eloszlását minden hiányzási csoportra és ennek a viselkedésnek az aspektusait extrapolálja a nem megfigyelt adatokra.

Chen és Astebro (2003) az adathiány mechanizmus becslésére[45] 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 modell alkalmazásához kapcsolódó másik kérdés a prior eloszlás kiválasztása. A modell a Béta eloszlás többváltozós általánosítását a Dirichlet eloszlást használja konjugált prior eloszlásként (lásd (1) egyenlet). Számos elfogadható nem informatív Dirichlet prior létezik. Ha $\alpha_{ij} = 1$ minden *i,j*-re, akkor egyenletes eloszlást kapunk. Az $\alpha_{ij} = 0$, minden *j*-re beállítás alkalmatlan prior eloszlást eredményez, ami egyenletes $\log(\theta_{ij})$-ben. A posterior eloszlás alkalmas, ha minden *i,j* kategóriában van legalább

[45] azaz a hiányzás valószínűségének leírására

egy megfigyelés. Kellően nagy adatbázis esetén a fenti két prior alkalmazásával nyert eredmények között nincs nagy különbség. Chen és Astebro (2003) az $\alpha_{ij} = 0$ beállítást alkalmazták tanulmányukban.

A módszer előnye a kidolgozott elméleti háttér, a relatíve egyszerű alkalmazás és hogy könnyen kiterjeszthető többszörös imputáció alkalmazására is.
Chen és Astebro (2003) 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, és *külső információ:* a régi scoring építésnél felhasznált teljes adatokból, megfelelően súlyozva). Azt találták, hogy NMAR esetén (a tréning adatokon) ez a Bayes-i módszer jobb, mint a Heckman-féle kétváltozós kétlépcsős modell. Ha viszont az adathiány inkább MAR jellegű (a teszt adatokon), akkor *a modell gyengébb teljesítményű, mint a pusztán csak 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.

2.3.7 Maximum likelihood alapú módszer

A Chen és Astebro (2005) által javasolt modell a hagyományos maximum likelihood megközelítésen alapul. Bár ők logit modellnél használták, a módszer alkalmazható minden maximum likelihood alapú eljárás esetén. Ez a reject inference technika beépíti az adathiány mechanizmus okozta bizonytalanságot a modellépítésbe.
A logit modellek 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, β a paraméterek vektora, u_i a hibatényező és y^* nem megfigyelhető. Csak az y dummy változót 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:

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

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 logit modell feltételezi, hogy az u_i eloszlásfüggvénye logisztikus. Ekkor:

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

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

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:

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

$$\sum_{y_i=hiányzó} ((1-\lambda_i)\ln(F(-\beta' x_i)) + \lambda_i \ln(1 - F(-\beta' x_i))) \tag{4}$$

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, a hiányzás valószínűsége becsülhető csak az elfogadottakból.

Azonban, ha 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.

Paik, Sacco és Lin (2000) is egy ehhez a modellhez hasonló számítási módszert javasoltak. Az ő eredeti modelljük célja a kétváltozós bináris adatok kezelése volt NMAR adathiány esetén. A hiányzó y_i kezelése úgy történik, hogy kicserélik a

feltételes várható értékével, adott megfigyelések mellett $E(y|x_i,a_i,s_i)$, ahol s_i a credit score.

Az $E(y|x_i,a_i,s_i)$ nem más, mint a hiányzás várható értéke (λ) a (4) egyenletben. Feltéve, hogy az $E(y|x_i,a_i,s_i)$ elérhető, Paik et al.(2000) a Newton-Raphson algoritmust használva kapja meg a becsléseket, és jackknife variancia becsléssel a varianciát.

Chen és Astebro szerint az ő modelljük legalább olyan jó, mint a Paik és szerzőtársai által javasolt, egyrészt a maximum likelihood alapú becslések nagymintás kedvező tulajdonságai miatt, másrészt mivel a módszer számításigényessége a napjainkban elérhető modern statisztikai szoftverek segítségével már nem jelent valódi hátrányt. A módszer sikeres alkalmazásának kulcsa egyedül a λ hiányzási valószínűség megfelelő becslése.

Chen és Astebro (2005) 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" információ forrásnak 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. Összefoglalás

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) azaz $p(y=1|s,R) > p(y=1|s,A)$, még akkor is, ha nem tudjuk korrekten számszerűsíteni, hogy mennyivel nagyobb. Az elutasítottak alkalmazásának sikere nagymértékben függ

attól, mennyire sikerül a sokasági jó-rossz arány becslése. Ezt az arányt azonban nehéz megbecsülni.

Az elutasítottak tényleges és imputált adatainak alkalmazásának haszna függ az elutasítási aránytól, a mintabeli és sokasági eloszlásoktól és az alkalmazott statisztikai feltételek teljesülésétől. Van néhány portfolió, ahol nagyon alacsony az elutasítottak aránya (ilyen például a jelzáloghitelek piaca). Ilyen esetekben felesleges lehet az elutasítottakkal foglalkozni, hiszen elhanyagolható az arányuk a populáción belül, így az általuk okozott torzítás sem igényel korrekciót. Másrészt a nagyobb kockázatú portfoliók esetén, 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 adatbázis-függő.

Néhány statisztikus szerint az elutasítottak hiányzó bedőlési adatainak megfelelő imputációjával megoldható a nem véletlen mintából való következtetés problémája (Joanes 1993/4, Donald 1995, Copas és Li 1997, Greene 1998). Külföldön a scorecard fejlesztők már alkalmaznak reject inference technikákat, amelyben statisztikai szoftver csomagok (például SAS) is segítik őket. Ezek azonban sokszor fekete dobozként üzemelnek, mert a mögöttük lévő alapelvek és feltételezések nem világosak a felhasználók számára.

Ha elfogadhatónak tartunk bizonyos feltételezéseket és valamilyen imputációs eljárással felhasználjuk az elutasítottakat, akkor felmerül a kérdés, hogy hogyan validálhatjuk a modellünket és hogyan mérhetjük az általa elért javulást. Ebben a témában kevés releváns tanulmány született, mert a tesztelésre használt adatbázisok többsége nem teljes, vagy szimulált volt (Donald 1995, Feelders 1999, Manning et al. 1987).

Hand és Henley (1993/4) megmutatták, hogy az üzleti életben alkalmazott megoldások problematikusak, mert általában igen kétséges feltételezéseken alapulnak.

A szakirodalom áttanulmányozása után ugyanarra a következtetésre jutottam, mint ők, azaz:

A torzítás kiküszöbölésének egyetlen robusztus és megbízható módja, ha az elutasítottak egy részét ténylegesen meghitelezik és így figyelik meg viselkedésüket és 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 csökkentésének egy lehetséges módja a véletlen időszakokban való (résnyire) nyitott kapu alkalmazása egyfajta költségoptimális mintaelosztással.

Gyakorlati szempontból jó lenne elkerülni a nemvéletlen 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 kivétel ágon való beengedés, mind az ügyintézői elutasítás esetén – megérné a fáradságot a scoring függvény 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 kisösszegű és nagyszámosságú) 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 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. Igen ám, 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, teljes banki adatbázison megvizsgálom a pótlólagos információszerzés által elérhető modell-javulást, annak költségeit és várható hasznait. A pótlólagos információkat a résnyire nyitott kapu módszerrel, költségoptimális mintaelosztással szerezzük.

IV. Az empirikus kutatás és eredményei

Nagyon nehéz a reject inference vizsgálatához megfelelő adatbázist találni. Elméletileg szükség van egy olyan adatbázisra, amelyben senkit nem utasítanak el. Az ezen a teljes mintán épített scoring modell az *etalon modell*. 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 scoring modell lesz a *kiinduló modell*. Az etalon modell 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, azaz hogy mennyire sikerült közelíteni a kiinduló modellt az etalonhoz. A modellek jóságát a II/3. fejezetben ismertetett mutatók és mérőszámok segítségével vizsgáljuk. Végül megvizsgáljuk a módszer költségeit és várható hasznát. A következő hipotéziseket fogjuk vizsgálni.

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 modell javulás által elérhető többlethaszon egy bizonyos üzemméret (portfolió-volumen) fölött meghaladja az információszerzés költségeit.

2. Adatbázis

A kutatáshoz egy magyarországi bank bocsátott rendelkezésemre egy a fenti elvárásoknak eleget tevő adatbázist. Az adatbázis egy olyan lakossági hiteltermék (hitelkártya) fogyasztóiról tartalmaz adatokat, amelynél egy adott időszakban majdnem mindenkit beengedtek (éppen scoring építési 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.[46] (A kategóriás változók a modellekben dummy változókkal vannak szerepeltetve, a kategóriák felsorolásának sorrendjében, mindig az utolsó kategória a referencia csoport.)

APPLICATION_ID	Azonosító
CSALADI_ALLAPOT	Családi állapot (egyedülálló/ élettársi kapcs /elvált/ házas/özvegy)
FOGLALKOZAS	Foglalkozás (alkalmazott vezető/ fizikai alkalmazott/ közalkalmazott, köztisztviselő/ vállalkozás tulajdonosa/ nyugdíjas/ szellemi alkalmazott)
FSZLA_VEZETO_BANK	Számlavezető bank (0:ez a bank/ 1:másik bank)
ISKOLAI_VEGZETTSEG	Iskolai végzettség (8 általános vagy kevesebb/ érettségi/ felsőfokú/ szakképesítés)
LAKAS_JOGCIM	Lakásjogcím (bérlő/ családtag/ egyéb/ tulajdonos)
NEM	Nem (0:nő/ 1: férfi)
UGYFELTIPUS	Kártyatípus (0: A/ 1: B)
BUDGET_JOVEDELEM	Jövedelem kategorizálva (kvintilisek)
ELETKOR	Életkor kategorizálva (kvintilisek)
DEFAULT_	Visszafizetés (1:rossz adós / 0: jó adós)

9. táblázat Az adatbázisban szereplő változók

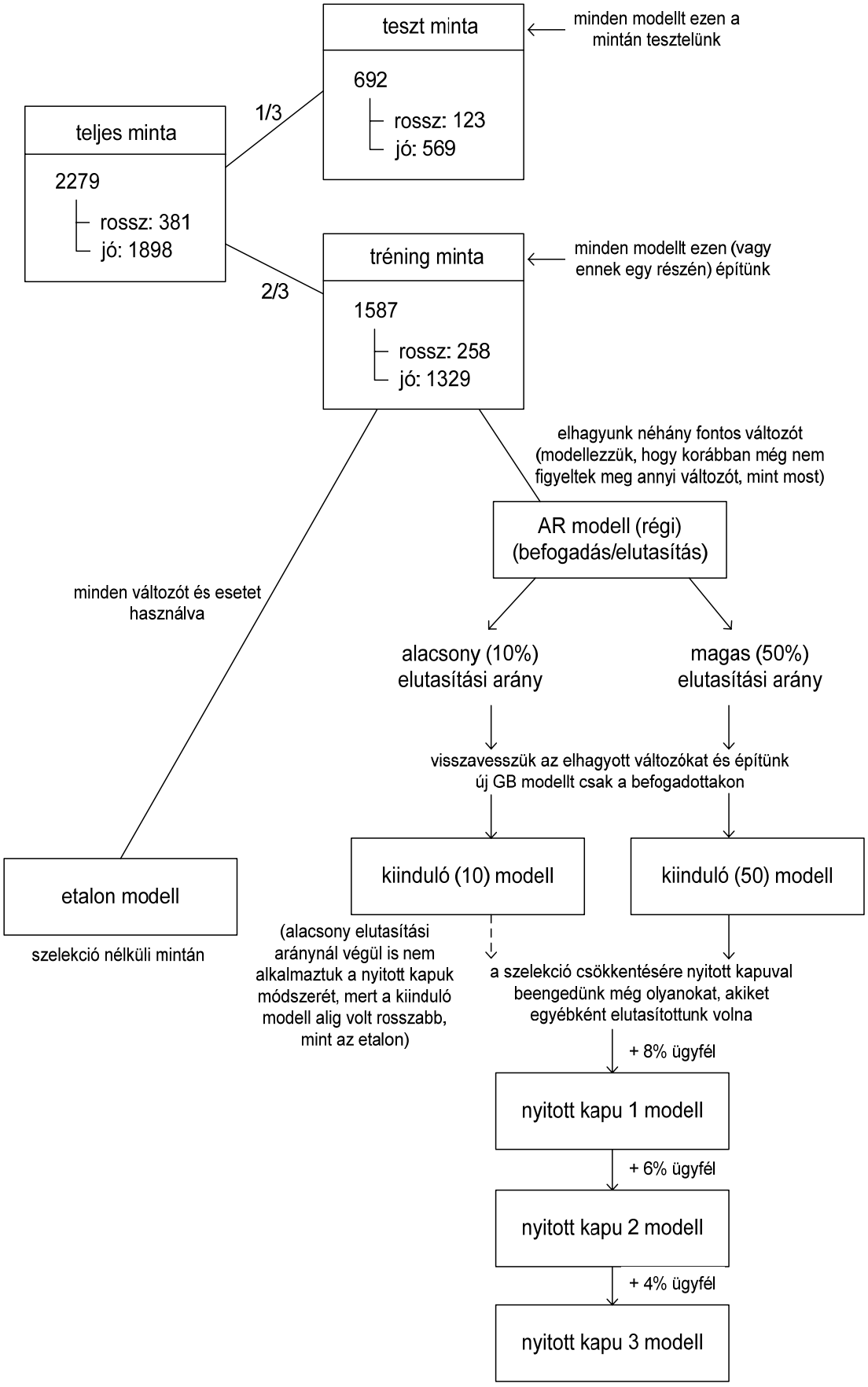
3. A modellezés folyamata

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 nagyobbnak értékeljük. Minden modellt a tréning adatbázison (vagy annak egy részén) építünk, de a modellek teljesítményét a teszt adatokon mérjük.

A modelleket SPSS programcsomag segítségével, logisztikus regresszióval 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 szignifikancia szint beállítással), így a modellek közötti különbségek csak a minta különbözőségének tudhatók be.

A következő ábra mutatja a modellezés folyamatát:

[46] A dolgozatban csak azokat a változókat használhattam, amelyeket szinte minden bank alkalmaz és azokat is csak kategorizálva.



11. ábra A modellezés folyamata

4. Eredmények

A tréning mintán megépítettük az *etalon modellt*, ez most számunkra a létező legjobb modell, mert ez egy teljesen véletlen adatbázison épült.

Az etalon modell paramétereit és illeszkedési mutatóit láthatjuk az alábbi táblázatokban.[47]

Variables in the Equation (etalon)

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 5(a)	NEM(1)	,801	,159	25,366	1	,000	2,228
	Eletkor			61,212	4	,000	
	Eletkor(1)	1,803	,298	36,563	1	,000	6,071
	Eletkor(2)	1,378	,305	20,359	1	,000	3,967
	Eletkor(3)	,605	,328	3,395	1	,065	1,831
	Eletkor(4)	,569	,338	2,835	1	,092	1,766
	iskvégzettség			50,598	3	,000	
	iskvégzettség(1)	-,990	1,048	,893	1	,345	,372
	iskvégzettség(2)	-,436	,165	6,956	1	,008	,647
	iskvégzettség(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
	kartyatipus(1)	,580	,181	10,286	1	,001	1,786
	Constant	-2,354	,375	39,332	1	,000	,095

10. táblázat Az etalon modell paraméterei

Tehát szignifikáns magyarázó változók lettek: a nem, életkor, iskolai végzettség, jövedelem (BUDGET), és a kártyatípus változók. Például az életkor(1) dummy változó B=1,803 –as paramétere így értelmezhető: Exp(B)= 6,071, ami azt jelenti, hogy a legfiatalabbak esetén a p/(1-p) odds értéke 6,071-szeresére nő a legöregebbekéhez képest, minden más magyarázó változó változatlansága esetén.[48]

[47] Csak erre az egy modellre vesszük részletesen végig az outputokat, a többi modellnél csak egy összefoglaló táblázatot közlünk.

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

Model Summary (etalon)

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.

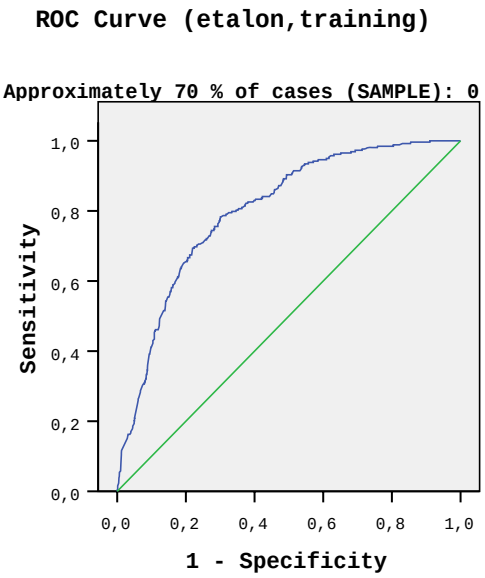
11. táblázat Az etalon modell illeszkedési mutatói a tréning adatbázison

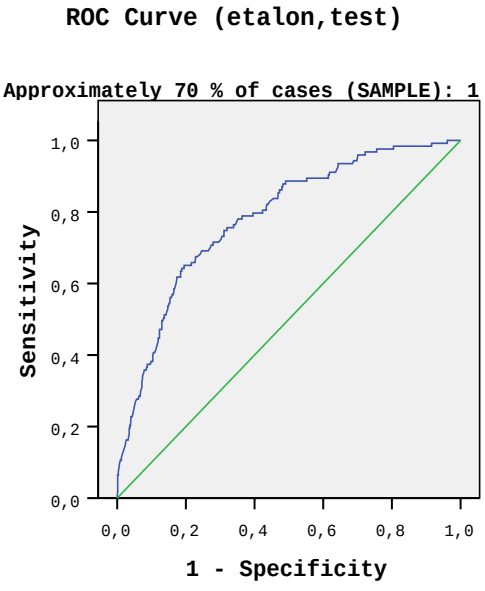
A logit modell paramétereinek becslése maximum likelihood módszerrel történik. A likelihood maximalizálása ekvivalens a -2 Log likelihood minimalizálásával, tehát minél kisebb a -2 Log likelihood, annál jobb a modell. A lilelihood értékét az üres modelléhez kell hasonlítani. Ezt teszi a Cox-Snell R^2, 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.[49] Az etalon modellünk magyarázóereje tehát 25,4%-os[50].

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

[50] Mindig az utolsó lépésbeli mutatókat kell néznünk. Mivel backward eljárással építettük 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.

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





12. ábra ROC görbék az etalon modellre a tréning és a teszt adatokon

Area Under the Curve (etalon)[c,d]

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.

12. táblázat AUROC értéke az etalon modellnél a tréning és a teszt adatokon[51]

Az AUROC értékének minimuma 0,5, maximuma 1, minél nagyobb, annál jobb a modell. Az etalon modell esetében a modellépítésre szánt mintán a mutató értéke 0,8 a tesztelésre szánt mintán kicsit kisebb 0,782. Azt a nullhipotézist, miszerint az aktuális modellünk nem különbözik szignifikánsan a véletlenszerű besorolást jelentő (üres) modelltől, minden szokásos szignifikancia szinten elvethetjük (p=0,000).

A modellek teljesítményét a teszt adatbázison fogjuk összehasonlítani, ezért a leggyakrabban alkalmazott AUROC-on kívül kiszámítottuk a szakirodalom által leginkább ajánlott (például Hand, 1997) Brier-score és Logaritmikus score értékét is a teszt-adatokon.

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 etalon modellhez fogjuk hasonlítani.

Ezek után *szimuláljuk a szelekciót*, azaz úgy csinálunk, mintha a bank alkalmazott volna valamilyen szűrőt (beengedés/elutasítás -, AR modellt) az ügyfelek beengedésénél. Ezt a szűrőt úgy készítjük el, hogy a (tréning) adatbázison építünk egy logit modellt, de úgy, hogy ne szerepeljenek benne a *jövedelem* és a *kártyatípus változók* . Ezen változók elhagyásával azt szeretnénk szimulálni, hogy a múltban még

[51] 0: tréning, 1: teszt

nem figyeltek meg annyi változót, mint most[52]. Az így létrejött AR modell outputjait már nem mutatjuk be külön részletesen, hanem az alábbi táblázatban összefoglaljuk az összes modell jellemzőit:

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

13. táblázat A modellek és jellemzőik összefoglaló táblázata

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

14. táblázat A modellek magyarázó változói

[52] A valóságban nem csak 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 etalon modell között, de ennek vizsgálata nem célja a kutatásnak és nem is tudnánk beépíteni a modellezésbe, mert csak egy időszakból vannak adataink.

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)[53] *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.[54] Ezen a mintán tehát építünk egy új modellt, amelyhez már minden elérhető magyarázó változót felhasználunk. (Konkrétan 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 teszt adatbázison hasonló az etalon modell értékeihez, tehát a modell teljesítményén nincs mit javítani. (Az AUROC értéke (0,786) még jobb is mint az etalon modell 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. Az AUROC értékekre számítottunk konfidencia intervallumot (lásd függelék), ebből azt látjuk, hogy a nagynak tűnő különbség 5%-on nem szignifikáns, de 10%-on már igen.

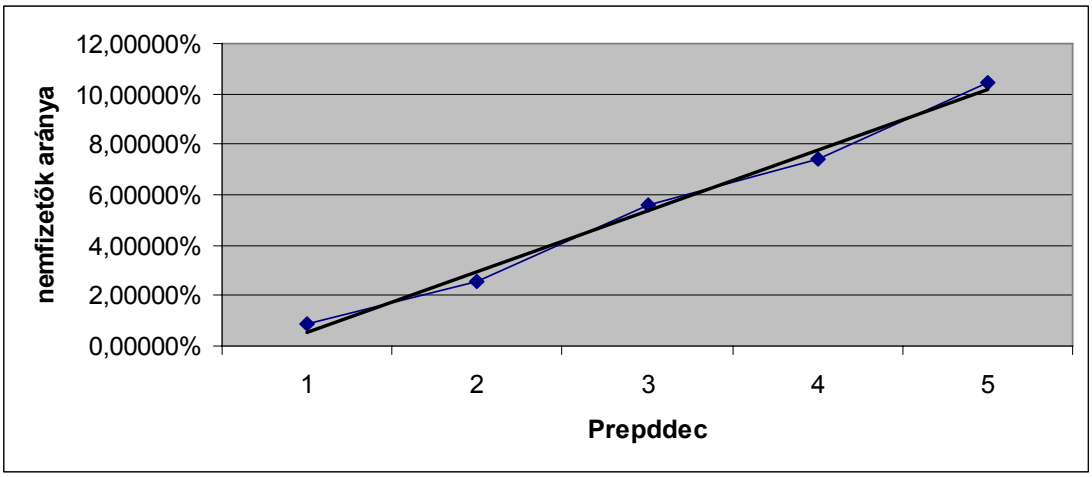
[53] Nem tudjuk pontosan 10 és 50%-ra beállítani az elutasítási arányt, mert kategóriás változóink vannak és sok az egyforma eset.

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

Nézzük mi lehet az oka ennek. A múltban fontos volt a nem, életkor, foglalkozás és iskolai végzettség (ezek az AR modell szignifikáns változói). Tudjuk, hogy ezek most is fontosak + az újonnan megfigyelt változók is (jövedelem, kártyatípus) (ezek az etalon modell 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 lett szignifikáns. Ez lehet pusztán amiatt, 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 (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 megpróbálni javítani rajta.[55]

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.[56]
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.

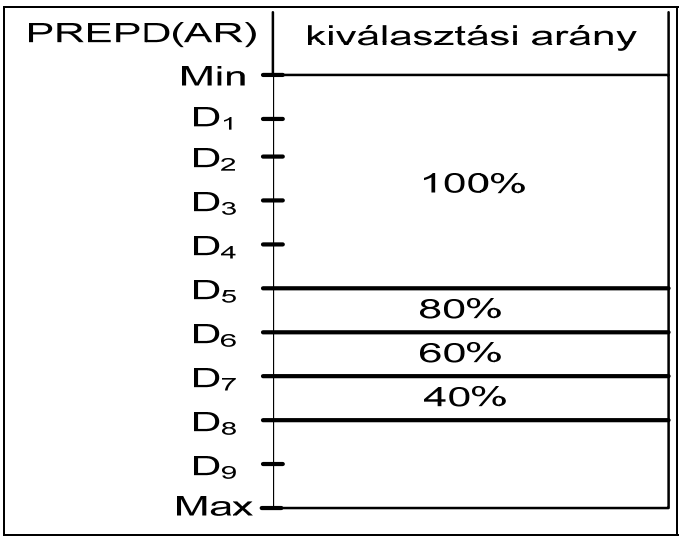
[55] A valóságban az etalon modellt soha nem ismerjük. Most azért készítettük el, hogy lássuk, hogy egyáltalán mekkora az elérhető maximális javulás. És ennek tükrében értékeljük majd a javított modelljeinket.
[56] Az adatbázis nem tartalmazza, hogy a hitelkeretből mekkora rész került felhasználásra, ezért feltételezzük, hogy mindig a teljes hitelkeretet kihasználják.



13. ábra Empirikus nemfizetési arány

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).[57]

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.



14. ábra A mintába kerülés valószínűsége a prediktált bedőlési valószínűség függvényében

[57] A valóságban csak feltételezhetjük ezt a tendenciát, most viszont, mivel az egyébként elutasítandókról is van adatunk ellenőrizhetjük a feltételezés helytállóságát. (A lineáris tendencia folytatódott.)
Valójában tehát itt mégis alkalmazunk egy feltételezést, aminek 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 ettől kisebb kiválasztási arányokat kell beállítanunk, ha nagyobb, akkor lehet nagyobbakat.

A résnyire nyitott kapuval való beengedést három fokozatban hajtjuk végre.

Az első esetben csak + 8% ügyfelet engedünk be, az eddig beengedettekhez közel álló esetek[58] 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.

A NYK 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 függelék tartalmazza a modelleknél használt súlyokat.)

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

[58] Az eddig beengedett legrosszabbaknak tűnőktől csak kicsit rosszabbak.

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 etalon modell, 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 szignifikancia szinten nem szignifikáns (összeérnek a konfidencia intervallumok), 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 csökkent, 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 etalon modell 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 (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[59], de a különbség igen kicsi, nem szignifikáns.

[59] 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.

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 többlet információra 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 modell javulás által elérhető többlethaszon egy bizonyos üzemméret (portfolió-volumen) 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:

	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:

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 most 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 cutoff értékét, azaz, hogy milyen 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 cutoff értékét[60].
Mint azt már a II/3. fejezetben tárgyaltuk a *gyakorlatban* általában a cutoff értékek lehetséges tartományán megvizsgálják a modellépítési mintán a különböző cutoff értékekhez tartozó profit (vagy hozam) értékeket és azt a cutoff értéket választják, amely mellett a mintán maximális a profit.
Elméletileg 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 0-100% -ig[61] terjedő cutoff értékek mellett elérhető profitot. A fenti haszonmátrix mellett a vizsgálandó modelleknél kiválasztottam a profitmaximalizáló cutoff értéket a tréning adatbázison. (A profitgörbék megtalálhatók a függelékben, az optimális cutoff értékek pedig a korábbi összefoglaló táblázatban és az alábbi táblázatban is.) Ha több maximumhelye volt a profit görbének, akkor (a nagyobb piaci részesedés miatt) a nagyobbat választottam.
Az elméleti megoldás szerint a cutoff 0,0909 ami azt jelenti, hogy a 9%-os bedőlési valószínűségű ügyfeleket még be kell fogadni, a 10%-ost el kell utasítani. Mivel 1%-os lépésközű profitszámítást készítettem, így itt az elméleti cutoff 10% (0,1).
Az így meghatározott optimális cutoff értékek mellett kiszámítottam a teszt adatbázison elérhető profitot. (A teszt adatbázishoz tartozó profitgörbék is a függelékben találhatók.)
Költség és haszon eredmények:

[60] Most elkülönülten csak ezen az egy terméken elérhető profitot akarjuk maximalizálni.
[61] 1%-os lépésközzel

	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 nagy mértékben. Az eredményekből látható, hogy *az elméleti* 10%-os *cutoff alkalmazásával* minden esetben[62] *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:

	jó	rossz	Összes
K50	766	37	803
NYK1	877	60	937
NYK2	957	78	1035
NYK3	996	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 finanszírozni, hanem a költséget megosztva, más bankoktól, vagy

[62] Kivéve a K50 modellnél, mert itt a kétféle cutoff egyezik.

hitelinformációs 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 portfolión is megtérül, hiszen a 692 elemű teszt adatbázison a modelljavulás következtében 3,3-ról 14,9-re nőtt a profit.[63]

A NYK2 modellhez +98 ügyfelet engedtünk be, ami 6,4 egység költséget jelentett és a modell javulás hatására további 1 egységgel nőtt a profit a 692 elemű teszt adatbázison. Hogy a költségek megtérüljenek egy 6,4-szer ekkora jövőbeli portfolióra van szükség. Az 1035 fős modellezési mintához képest tehát 4,3-szor akkora jövőbeli várható portfolió 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.

[63] Igazából ez egy jövőbeli profit, tehát diszkontálnunk kellene, mert a költségeket viszont most kell vállalni.

5. Összefoglalás

A dolgozatban a credit scoring modelleknél fellépő szelekciós torzítást és annak csökkentésére szolgáló módszereket vizsgáltuk.

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.

A nemhitelezett (elutasított) ügyfelekről vannak ugyanis adataink, de ezeknél hiányzik a hitelvisszafizetést leíró eredményváltozó értéke. Ezért az értekezésben az adósminősítési modelleknél fellépő szelekciós torzítást adathiányból eredő problémaként kezeltem.

Áttekintettem a hiányzó adatok típusait és kezelésük lehetséges módjait, kiemelve az egyes módszerek előnyeit, hátrányait, alkalmazásuk feltételeit. A hiányzó adatok kezelésére nem létezik egyetemesen legjobb megoldás. Lényeges szempont, hogy a választott imputációs eljárás összhangban legyen a később elvégzendő elemzésekkel, és az imputált adatbázisok esetében a felhasználók is láthassák az imputáló által alkalmazott módszert.

A credit scoring modellek elméleti áttekintése mellett összegeztem az eddigi gyakorlati tapasztalatokat, és kiemeltem e modellek legfőbb hiányosságait. A modellek teljesítményének mérésével kapcsolatban viszonylag nagyobb figyelmet szenteltem a konfúziós mátrixhoz tartozó haszon- (vagy költség-) mátrix összeállításával -, illetve az optimális cutoff kiválasztásával kapcsolatos kérdéseknek.

Az elméleti felvezetés után bemutattam a szakirodalomban fellelhető módszereket, amelyek a scoring modelleknél fellépő szelekciós torzítás csökkentését szolgálják. Mindegyik módszer valamilyen módon felhasználja az elutasítottakról meglévő információkat.

Az elutasítottak tényleges visszafizetési adatát nem ismerjük, ezért – mivel a semmiből nem keletkezhet új információ-, ha fel akarjuk használni őket a

modellépítéshez, akkor vagy *feltételezések*kel kell élnünk, vagy *pótlólagos*an *információ*t kell szerezni a visszafizetési viselkedésükről.

Megmutattam ezen (reject inference) technikák elméleti hátterét, kiemelve az alkalmazott feltételezéseket vagy a pótlólagos információ szerzésének és felhasználásának módját és összegeztem 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 mennyivel nagyobb. Az elutasítottak tényleges és imputált adatainak alkalmazásának haszna függ az elutasítási aránytól, a mintabeli és sokasági eloszlásoktól és az alkalmazott statisztikai feltételek teljesülésétől. Van néhány portfolió, 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 portfoliók 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 adatbázis-függő.

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

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, egyfajta költségoptimális mintaelosztással.
Ez azt jelenti, hogy 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 kisebb. Így egy rétegzett mintát kapunk egyfajta költségoptimális mintaelosztással. Végül átsúlyozással kaphatunk egy a sokaságot valóban reprezentáló mintát anélkül, hogy vállalni kellett volna a mindenki beengedésével járó hatalmas költségeket.
Az utolsó részben empirikus kutatás keretében egy valós banki adatbázison (lakossági hitelkártya adatokon) vizsgáltam az ezzel a módszerrel elérhető javulást, annak költségeit és várható hasznait.

Az empirikus kutatás során azt tapasztaltuk, hogy *magas elutasítási arány (erőteljes és nem teljesen véletlenszerű szelekció) esetén gyengébb teljesítményű modellek építhetők,* mint kisebb arányú elutasítás esetén. Ennek egyik oka, hogy ekkor kevés rossz ügyfél kerül a portfolióba, 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 információ szerzés egyik módja, 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 terméken elérhető profit is nőtt.*

Azt tapasztaltuk, hogy ha a profitmaximalizálás a cél, akkor *jobb, ha az elméleti úton meghatározott cutoff 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, azoktól csak kicsit rosszabbnak tűnő ügyfelekből érdemes résnyire nyitott kapuval beengedni még ügyfeleket.*

Ez az elsőlépcsős nagymértékű modelljavulás és profitnövekedés valószínű 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 plusz minta költségei tervezhetőbbek, és kisebbek is, mintha egy távoli tartományból vennénk mintát.

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

FÜGGELÉK

1. A modellek outputjai

(sorrendben: Etalon, AR, K10, K50, NYK1, NYK2, NYK3)

Variables in the Equation (etalon)

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 5(a)	NEM(1)	,801	,159	25,366	1	,000	2,228
	Eletkor			61,212	4	,000	
	Eletkor(1)	1,803	,298	36,563	1	,000	6,071
	Eletkor(2)	1,378	,305	20,359	1	,000	3,967
	Eletkor(3)	,605	,328	3,395	1	,065	1,831
	Eletkor(4)	,569	,338	2,835	1	,092	1,766
	iskvégzettség			50,598	3	,000	
	iskvégzettség(1)	-,990	1,048	,893	1	,345	,372
	iskvégzettség(2)	-,436	,165	6,956	1	,008	,647
	iskvégzettség(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
	kartyatipus(1)	,580	,181	10,286	1	,001	1,786
	Constant	-2,354	,375	39,332	1	,000	,095

Model Summary (etalon)

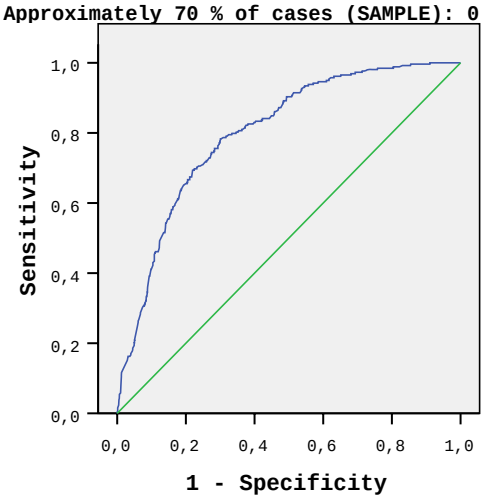
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1132,707[a]	,158	,269
2	1132,812[a]	,158	,269
3	1136,158[a]	,156	,266
4	1141,521[b]	,153	,261
5	1148,883[b]	,149	,254

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.

ROC Curve (etalon,training)

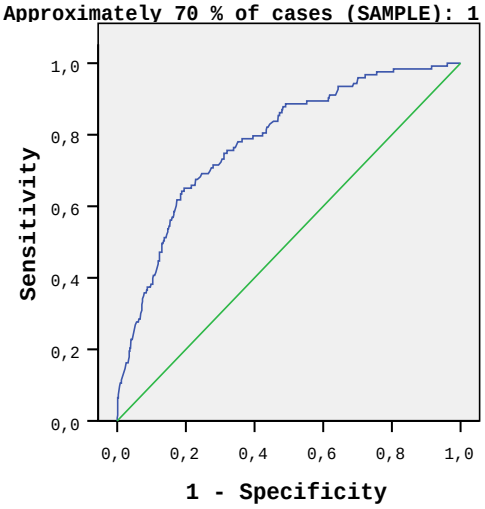
Approximately 70 % of cases (SAMPLE): 0



Diagonal segments are produced by ties.

ROC Curve (etalon,test)

Approximately 70 % of cases (SAMPLE): 1



Diagonal segments are produced by ties.

Area Under the Curve (etalon)[c,d]

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.

Variables in the Equation(AR)

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 4(a)	NEM(1)	,833	,159	27,589	1	,000	2,300
	Eletkor			46,575	4	,000	
	Eletkor(1)	1,704	,345	24,410	1	,000	5,495
	Eletkor(2)	1,367	,349	15,381	1	,000	3,924
	Eletkor(3)	,692	,364	3,610	1	,057	1,997
	Eletkor(4)	,560	,370	2,297	1	,130	1,751
	foglalkozás			12,030	5	,034	
	foglalkozás(1)	,328	,309	1,124	1	,289	1,388
	foglalkozás(2)	-,009	,213	,002	1	,966	,991
	foglalkozás(3)	-,831	,317	6,861	1	,009	,436
	foglalkozás(4)	-,112	1,122	,010	1	,920	,894
	foglalkozás(5)	-,532	,490	1,178	1	,278	,588
	iskvégzettség			30,894	3	,000	
	iskvégzettség(1)	-,758	1,050	,521	1	,470	,469
	iskvégzettség(2)	-,257	,167	2,346	1	,126	,774
	iskvégzettség(3)	-1,841	,334	30,452	1	,000	,159
	Constant	-2,627	,379	48,165	1	,000	,072

Model Summary (AR)

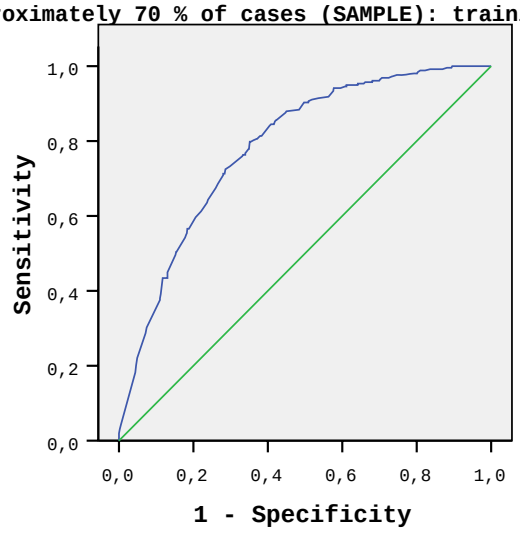
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1163,181[a]	,143	,244
2	1163,705[a]	,143	,243
3	1168,294[a]	,141	,239
4	1173,431[b]	,138	,234

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.

ROC Curve (AR)

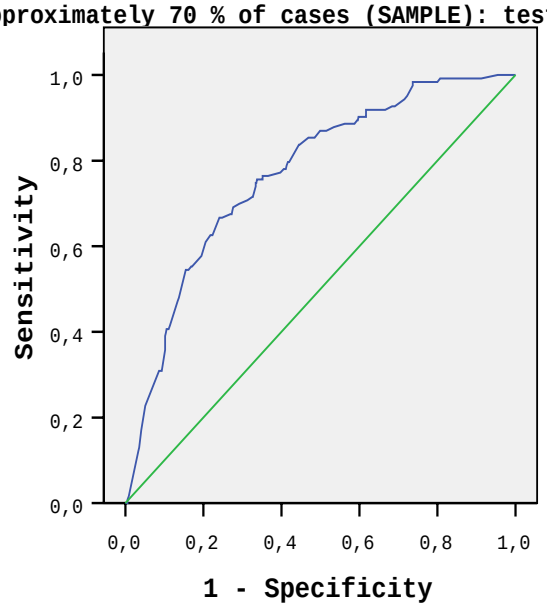
Approximately 70 % of cases (SAMPLE): training



Diagonal segments are produced by ties.

ROC Curve (AR)

Approximately 70 % of cases (SAMPLE): test



Diagonal segments are produced by ties.

Area Under the Curve (AR)[c,d]

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	,785	,014	,000	,757	,813
1	,769	,022	,000	,725	,813

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.

Statistics (AR)

		bsar	lsar
N	Valid	692	692
	Missing	0	0
Sum		,127298	,402803

Variables in the Equation (K10)

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 5(a)	NEM(1)	,798	,158	25,412	1	,000	2,221
	Eletkor			60,370	4	,000	
	Eletkor(1)	1,864	,305	37,301	1	,000	6,447
	Eletkor(2)	1,479	,312	22,523	1	,000	4,388
	Eletkor(3)	,752	,332	5,139	1	,023	2,122
	Eletkor(4)	,640	,343	3,476	1	,062	1,897
	iskvégzettség			48,037	3	,000	
	iskvégzettség(1)	-,913	1,048	,759	1	,384	,401
	iskvégzettség(2)	-,402	,164	5,975	1	,015	,669
	iskvégzettség(3)	-2,270	,329	47,496	1	,000	,103
	BUDGET			14,770	4	,005	
	BUDGET(1)	-,290	,274	1,124	1	,289	,748
	BUDGET(2)	-,523	,251	4,343	1	,037	,593
	BUDGET(3)	-,878	,259	11,481	1	,001	,416
	BUDGET(4)	-,647	,257	6,323	1	,012	,523
	kartyatipus(1)	,526	,180	8,568	1	,003	1,692
	Constant	-2,448	,380	41,470	1	,000	,086

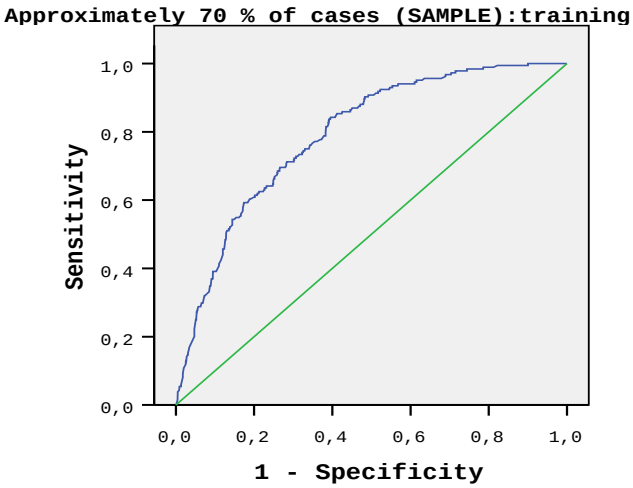
Model Summary (K10)

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1142,494[a]	,155	,263
2	1142,507[a]	,155	,263
3	1146,955[a]	,152	,259
4	1151,059[b]	,150	,255
5	1160,200[b]	,145	,247

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

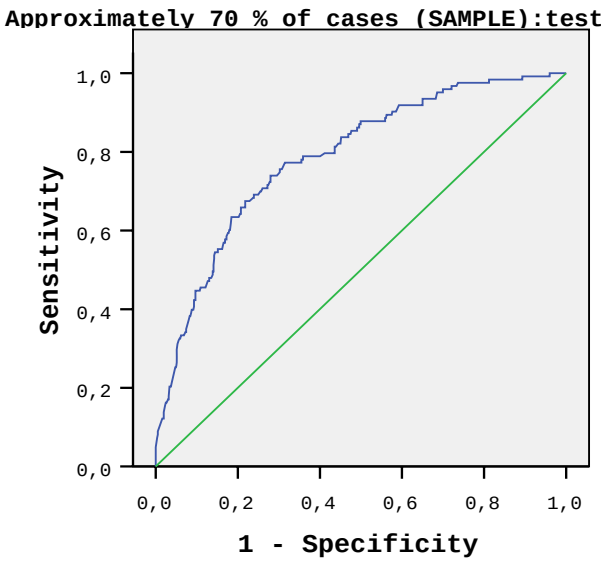
b. Estimation terminated at iteration number 6 because parameter estimates changed by less than ,001.

ROC Curve (k10)



Diagonal segments are produced by ties.

ROC Curve (k10)



Diagonal segments are produced by ties.

Area Under the Curve (k10)[c,d]

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	,790	,016	,000	,759	,822
1	,786	,022	,000	,742	,830

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.

Variables in the Equation (k50)

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 7(a)	Eletkor			13,252	4	,010	
	Eletkor(1)	2,299	,695	10,925	1	,001	9,963
	Eletkor(2)	,929	,694	1,793	1	,181	2,532
	Eletkor(3)	,486	,517	,885	1	,347	1,626
	Eletkor(4)	,047	,480	,010	1	,921	1,049
	foglalkozás			25,378	5	,000	
	foglalkozás(1)	-,767	1,082	,502	1	,478	,464
	foglalkozás(2)	1,691	,470	12,939	1	,000	5,426
	foglalkozás(3)	-,388	,529	,539	1	,463	,678
	foglalkozás(4)	,773	1,188	,423	1	,516	2,166
	foglalkozás(5)	,449	,651	,477	1	,490	1,567
	kartyatipus(1)	-,677	,403	2,820	1	,093	,508
	Constant	-3,477	,538	41,730	1	,000	,031

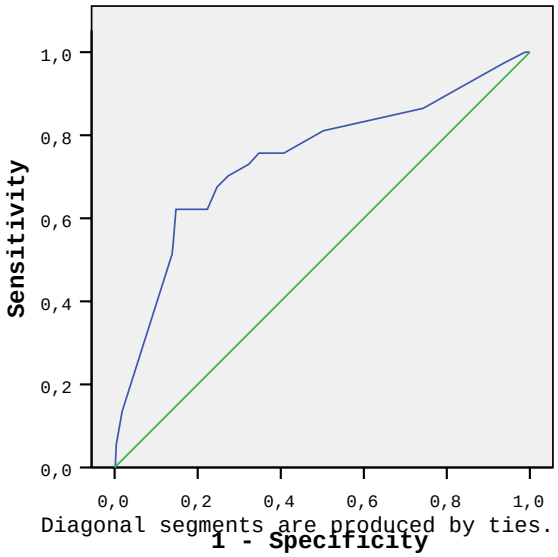
Model Summary (k50)

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	284,259[a]	,062	,181
2	284,446[a]	,061	,180
3	286,358[a]	,059	,173
4	289,210[a]	,056	,164
5	292,519[a]	,052	,152
6	295,069[a]	,049	,143
7	297,634[a]	,046	,135

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

ROC Curve (k50)

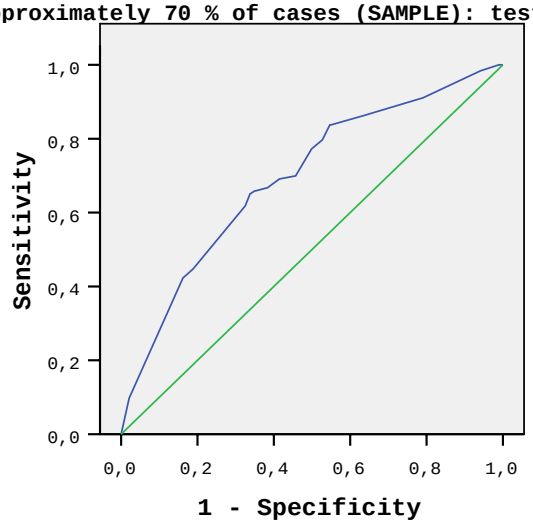
Approximately 70 % of cases (SAMPLE): training



Diagonal segments are produced by ties.

ROC Curve (k50)

Approximately 70 % of cases (SAMPLE): test



Diagonal segments are produced by ties.

Area Under the Curve (k50)[c,d]

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	,742	,048	,000	,648	,836
1	,694	,026	,000	,642	,746

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.

Statistics (k50)

		bs	ls
N	Valid	692	692
	Missing	0	0
Sum		,140566	,483118

Variables in the Equation (nyk1)

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 6(a)	NEM(1)	1,029	,372	7,664	1	,006	2,798
	Eletkor			16,946	4	,002	
	Eletkor(1)	2,402	,607	15,676	1	,000	11,048
	Eletkor(2)	1,107	,549	4,065	1	,044	3,025
	Eletkor(3)	,528	,458	1,325	1	,250	1,695
	Eletkor(4)	,461	,442	1,088	1	,297	1,586
	foglalkozás			14,844	5	,011	
	foglalkozás(1)	,254	,639	,158	1	,691	1,289
	foglalkozás(2)	1,246	,460	7,332	1	,007	3,478
	foglalkozás(3)	,086	,483	,032	1	,859	1,090
	foglalkozás(4)	,551	1,182	,217	1	,641	1,734
	foglalkozás(5)	-,015	,633	,001	1	,981	,985
	iskvégzettség			11,282	3	,010	
	iskvégzettség(1)	-1,079	1,066	1,025	1	,311	,340
	iskvégzettség(2)	-,637	,347	3,368	1	,066	,529
	iskvégzettség(3)	-1,649	,514	10,283	1	,001	,192
	Constant	-3,242	,575	31,800	1	,000	,039

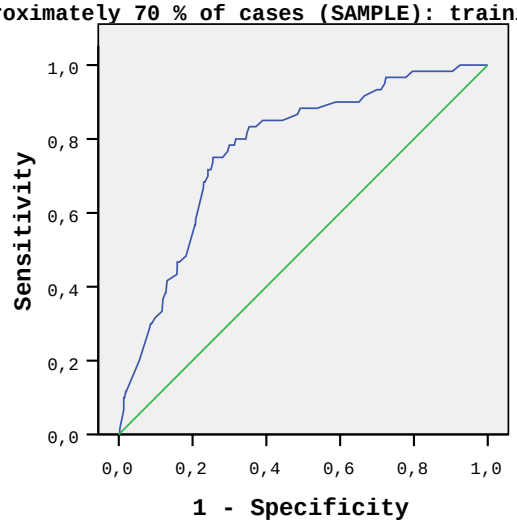
Model Summary (nyk1)

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	378,151[a]	,070	,184
2	378,851[a]	,069	,182
3	379,343[a]	,069	,181
4	383,218[a]	,065	,171
5	388,467[a]	,059	,157
6	390,782[a]	,057	,151

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

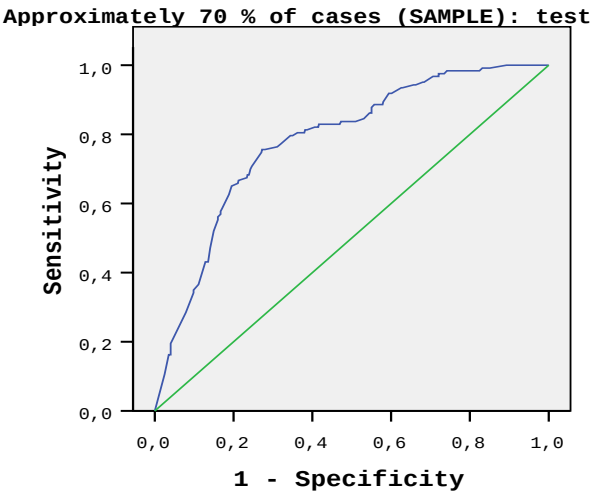
ROC Curve (nyk1)

Approximately 70 % of cases (SAMPLE): training



Diagonal segments are produced by ties.

ROC Curve (nyk1)



Diagonal segments are produced by ties.

Area Under the Curve (nyk1)[c,d]

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	,773	,029	,000	,716	,830
1	,782	,022	,000	,739	,824

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.

Statistics (nyk1)

		bs	ls
N	Valid	692	692
	Missing	0	0
Sum		,130885	,411004

Variables in the Equation (NYK2)

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 4(a)	NEM(1)	1,013	,265	14,595	1	,000	2,753
	Eletkor			22,138	4	,000	
	Eletkor(1)	2,871	,645	19,793	1	,000	17,657
	Eletkor(2)	1,374	,425	10,473	1	,001	3,952
	Eletkor(3)	,871	,382	5,189	1	,023	2,388
	Eletkor(4)	,593	,363	2,667	1	,102	1,810
	iskvégzettség			35,681	3	,000	
	iskvégzettség(1)	-1,234	1,105	1,247	1	,264	,291
	iskvégzettség(2)	-,949	,280	11,479	1	,001	,387
	iskvégzettség(3)	-2,629	,451	33,932	1	,000	,072
	lakjogcim			7,072	3	,070	
	lakjogcim(1)	-,081	,562	,021	1	,885	,922
	lakjogcim(2)	-,924	,402	5,295	1	,021	,397
	lakjogcim(3)	,899	,779	1,333	1	,248	2,457
	BUDGET			9,497	4	,050	
	BUDGET(1)	-,030	,428	,005	1	,944	,970
	BUDGET(2)	-,600	,403	2,212	1	,137	,549
	BUDGET(3)	-,992	,405	6,003	1	,014	,371
	BUDGET(4)	-,625	,395	2,500	1	,114	,535
	kartyatipus(1)	,691	,286	5,823	1	,016	1,996
	Constant	-2,300	,493	21,735	1	,000	,100

Model Summary (nyk2)

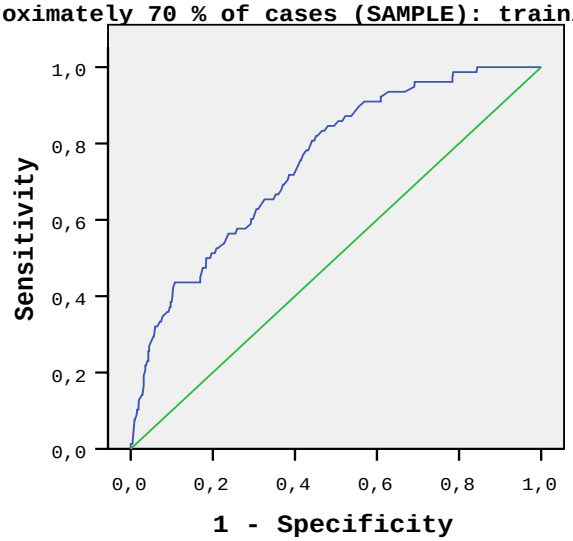
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	500,880[a]	,086	,197
2	501,051[a]	,086	,196
3	506,068[a]	,081	,186
4	514,207[b]	,074	,170

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.

ROC Curve (nyk2)

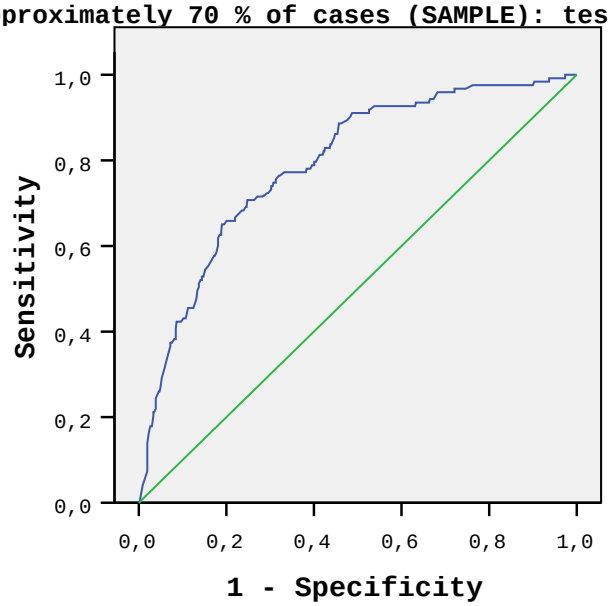
Approximately 70 % of cases (SAMPLE): training



Diagonal segments are produced by ties.

ROC Curve (nyk2)

Approximately 70 % of cases (SAMPLE): test



Diagonal segments are produced by ties.

Area Under the Curve (nyk2)[c,d]

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	,751	,027	,000	,699	,803
1	,791	,022	,000	,748	,834

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.

Statistics (nyk2)

		bs	ls
N	Valid	692	692
	Missing	0	0
Sum		,121808	,41304

Variables in the Equation (NYK3)

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 4(a)	NEM(1)	,944	,238	15,766	1	,000	2,570
	Eletkor			32,888	4	,000	
	Eletkor(1)	2,519	,470	28,713	1	,000	12,412
	Eletkor(2)	,951	,401	5,617	1	,018	2,589
	Eletkor(3)	,410	,342	1,432	1	,231	1,507
	Eletkor(4)	,456	,357	1,630	1	,202	1,578
	csaladiallapot			10,012	4	,040	
	csaladiallapot(1)	,919	1,160	,627	1	,428	2,507
	csaladiallapot(2)	2,056	1,138	3,265	1	,071	7,816
	csaladiallapot(3)	1,734	1,191	2,121	1	,145	5,664
	csaladiallapot(4)	1,418	1,100	1,660	1	,198	4,127
	iskvégzettség			42,712	3	,000	
	iskvégzettség(1)	-1,086	1,121	,938	1	,333	,338
	iskvégzettség(2)	-,874	,254	11,807	1	,001	,417
	iskvégzettség(3)	-2,603	,402	41,931	1	,000	,074
	BUDGET			18,303	4	,001	
	BUDGET(1)	-,041	,373	,012	1	,912	,959
	BUDGET(2)	-,905	,364	6,173	1	,013	,405
	BUDGET(3)	-1,058	,353	9,002	1	,003	,347
	BUDGET(4)	-,827	,355	5,443	1	,020	,437
	kartyatipus(1)	,659	,262	6,319	1	,012	1,934
	Constant	-3,509	1,149	9,319	1	,002	,030

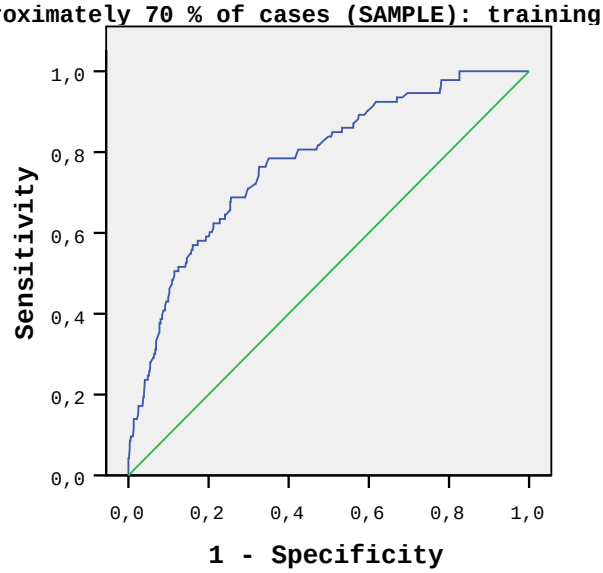
Model Summary (nyk3)

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	621,000[a]	,116	,238
2	623,456[a]	,115	,234
3	623,820[a]	,114	,234
4	629,119[a]	,110	,225

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

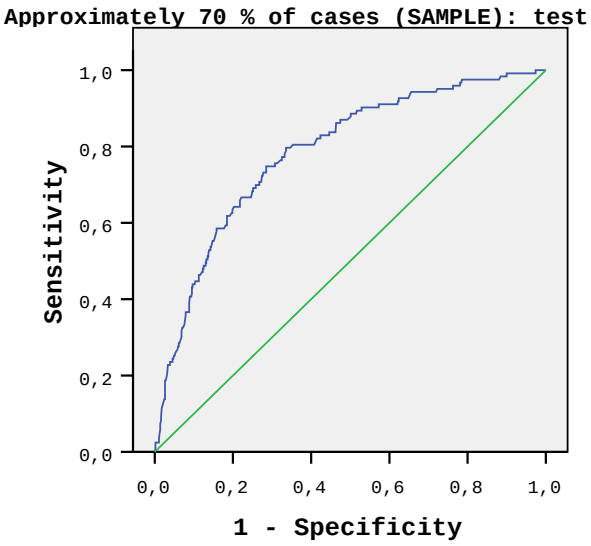
ROC Curve (nyk3)

Approximately 70 % of cases (SAMPLE): training



Diagonal segments are produced by ties.

ROC Curve (nyk3)



Diagonal segments are produced by ties.

Area Under the Curve (nyk3)[c,d]

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	,776	,025	,000	,727	,825
1	,786	,022	,000	,742	,830

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.

Statistics (nyk3)

		bs	ls
N	Valid	692	692
	Missing	0	0
Sum		,123599	,420510

2. A nyitott kapu modelleknél alkalmazott súlyok

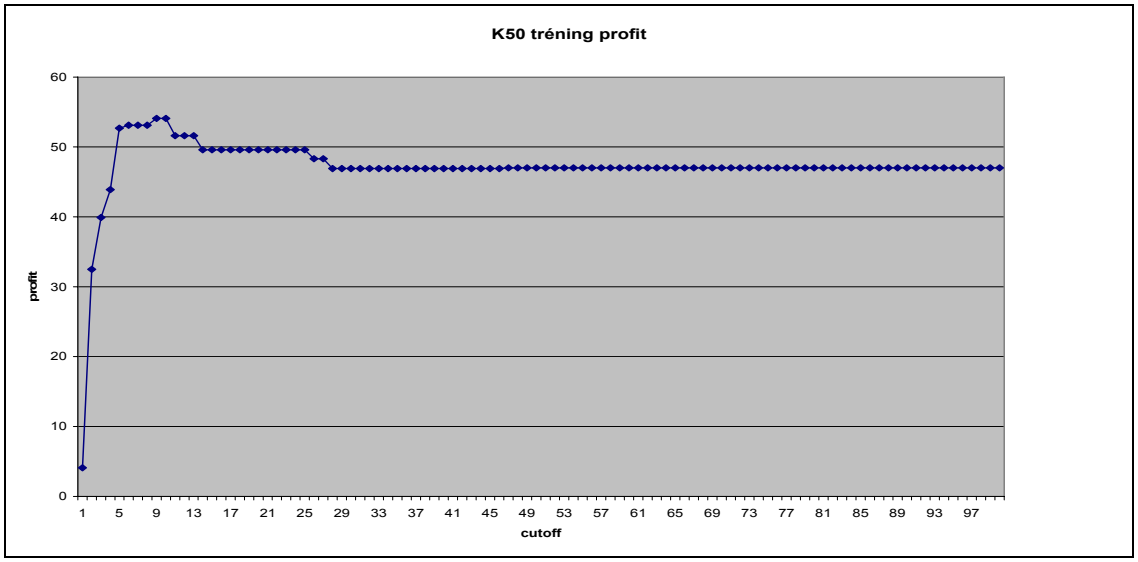
(sorrendben: NYK1, NYK2 és NYK3)

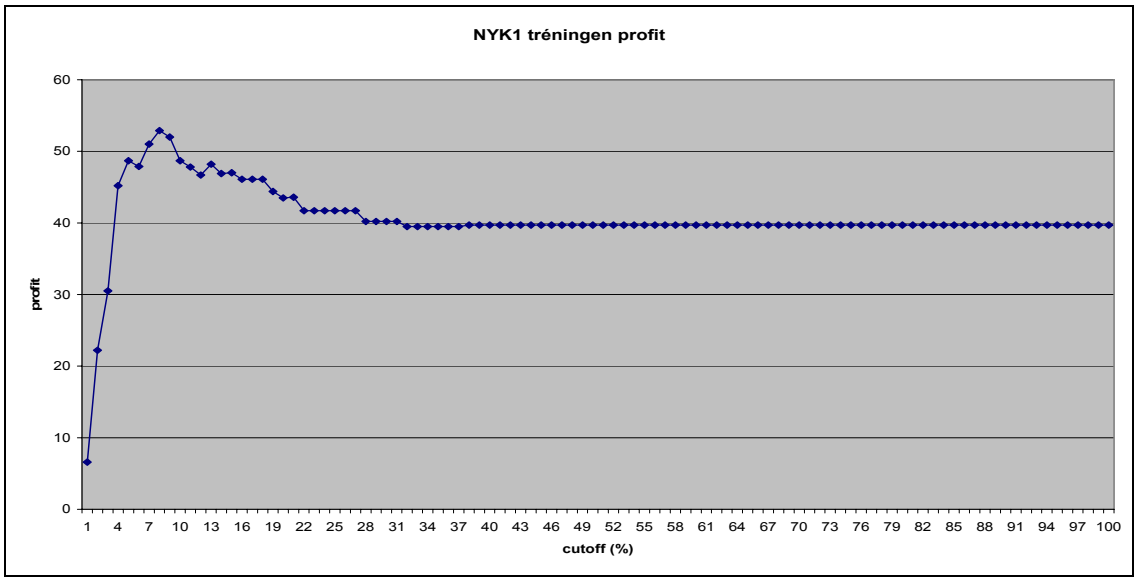
PREPD(AR)	kiválasztási arány	tényleges esetszám	súly (1)	súly végleges
Min				
D_1				
D_2	100%	803	1	$1 \bullet \frac{937}{970,5} = 0,9655$
D_3				
D_4				
D_5	80%	134	1 / 0,8 = 1,25	$1,25 \bullet \frac{937}{970,5} = 1,2069$
D_6	Σ	937	970,5	937
D_7				
D_8				
D_9				
Max				

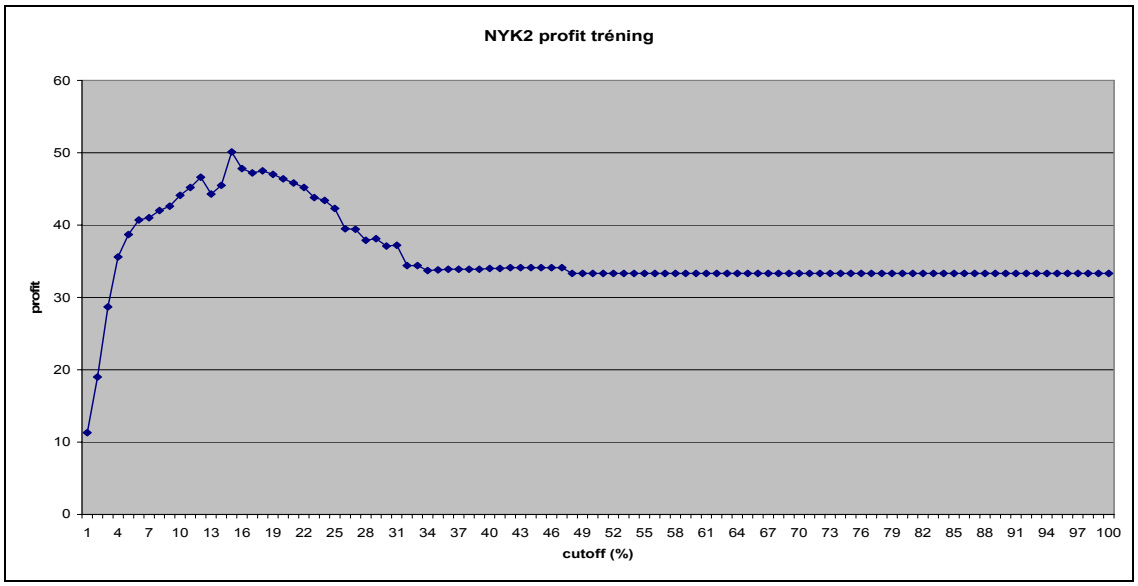
PREPD(AR)	kiválasztási arány	tényleges esetszám	súly (1)	súly végleges
Min				
D_1				
D_2	100%	803	1	$1 \bullet \frac{1035}{1133,83} = 0,9128$
D_3				
D_4				
D_5	80%	134	1 / 0,8 = 1,25	$1,25 \bullet \frac{1035}{1133,83} = 1,14104$
D_6	60%	98	1 / 0,6 = 1,66	$1,66 \bullet \frac{1035}{1133,83} = 1,52138$
D_7	Σ	1035	1133,833	1035
D_8				
D_9				
Max				

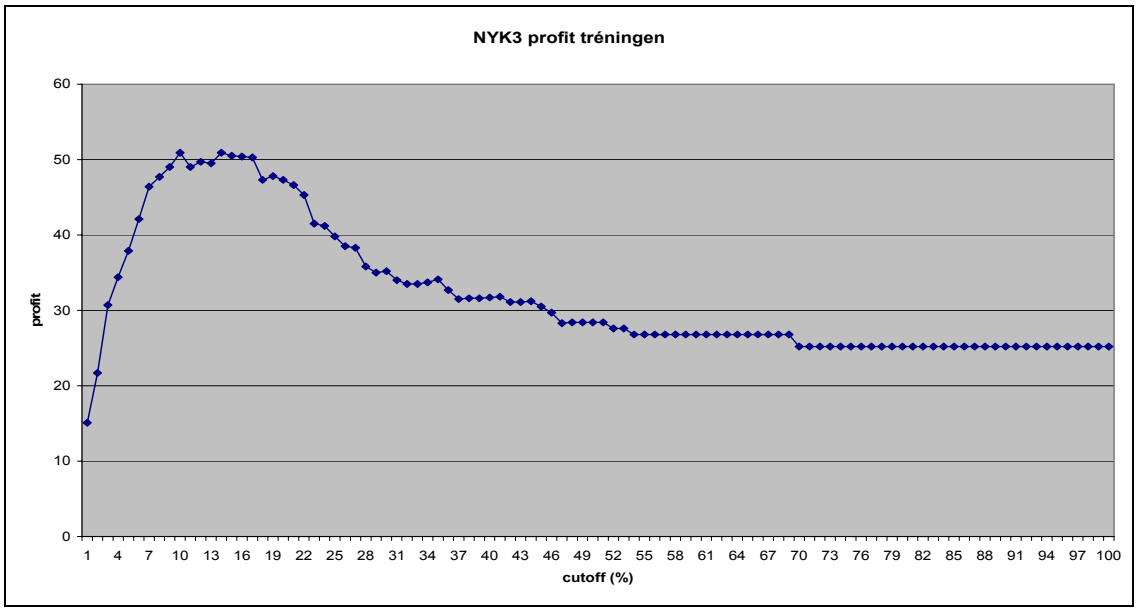
PREPD(AR)	kiválasztási arány	tényleges esetszám	súly (1)	súly végleges
Min				
D_1				
D_2	100%	803	1	$1 \bullet \frac{1133}{1268,83} = 0,8929$
D_3				
D_4				
D_5	80%	134	1 / 0,8 = 1,25	1,25 • 0,8929 = 1,1162
D_6	60%	98	1 / 0,6 = 1,66	1,66 • 0,8929 = 1,4882
D_7	40%	54	1 / 0,4 = 2,5	2,5 • 0,8929 = 2,2324
D_8	Σ	1133	1268,833	1133
D_9				
Max				

3. Profitgörbék a trénig adatokon

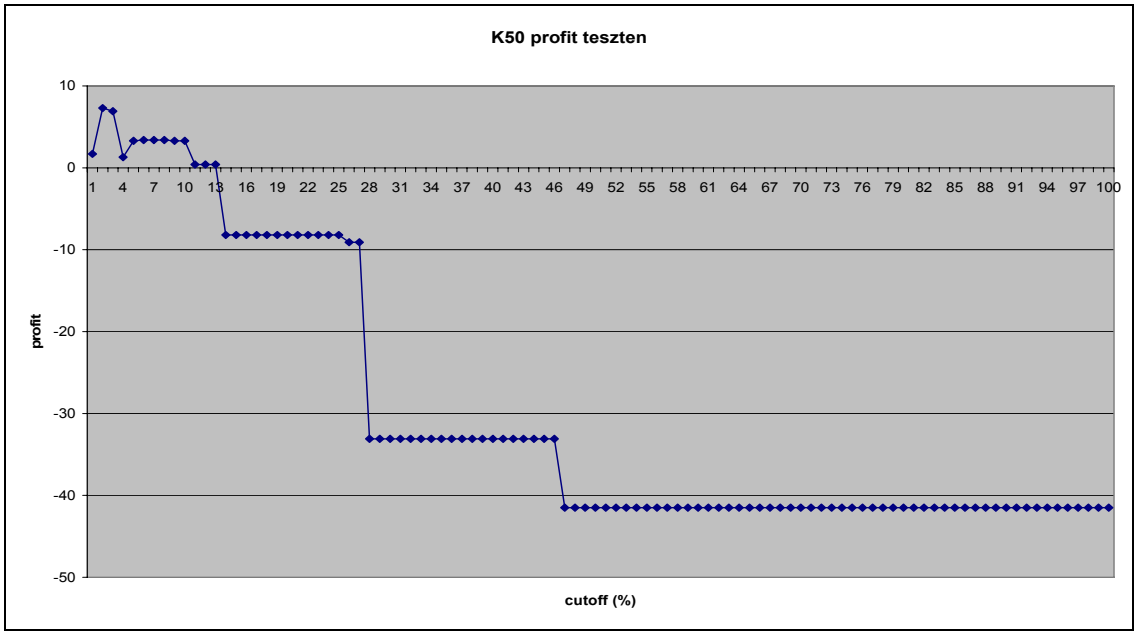


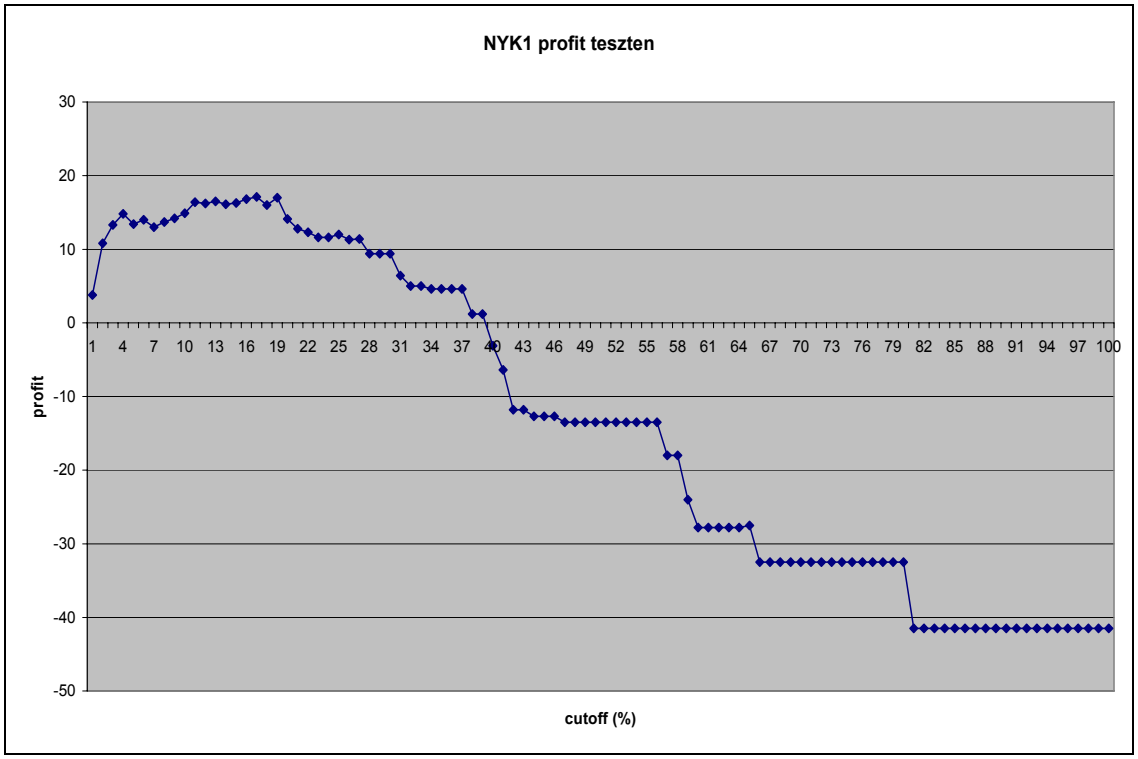

NYK1 tréningen profit
profit
cutoff (%)

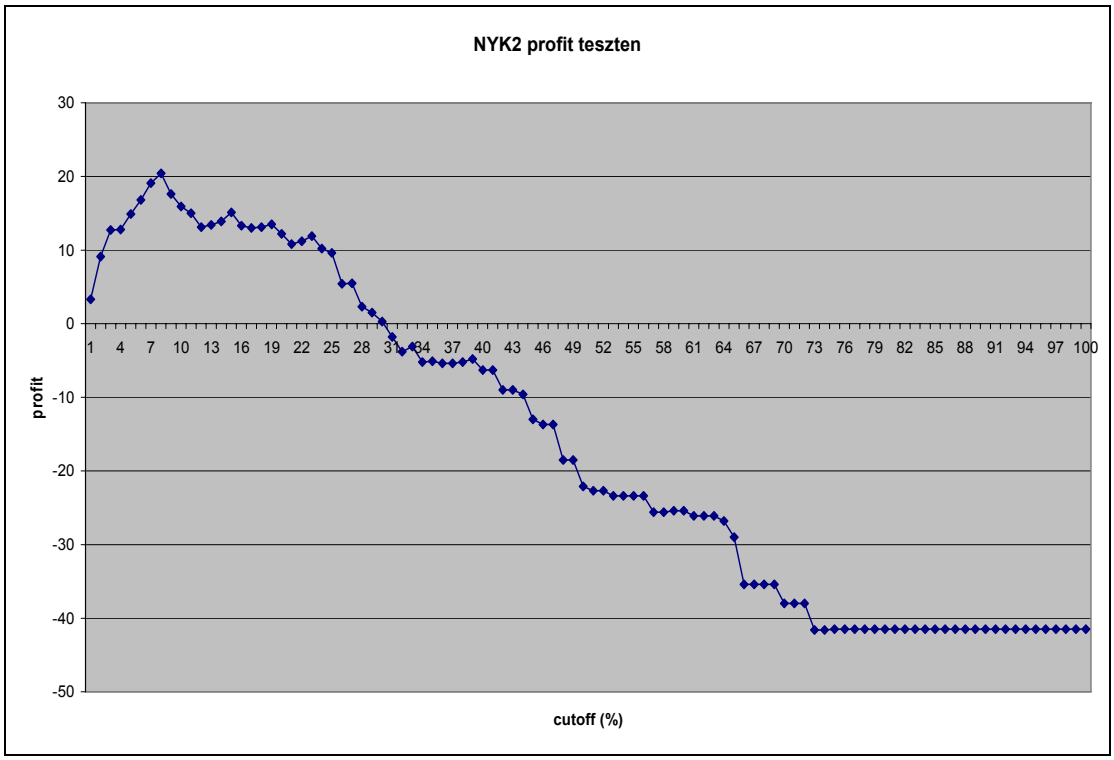

NYK2 profit tréning
profit
cutoff (%)

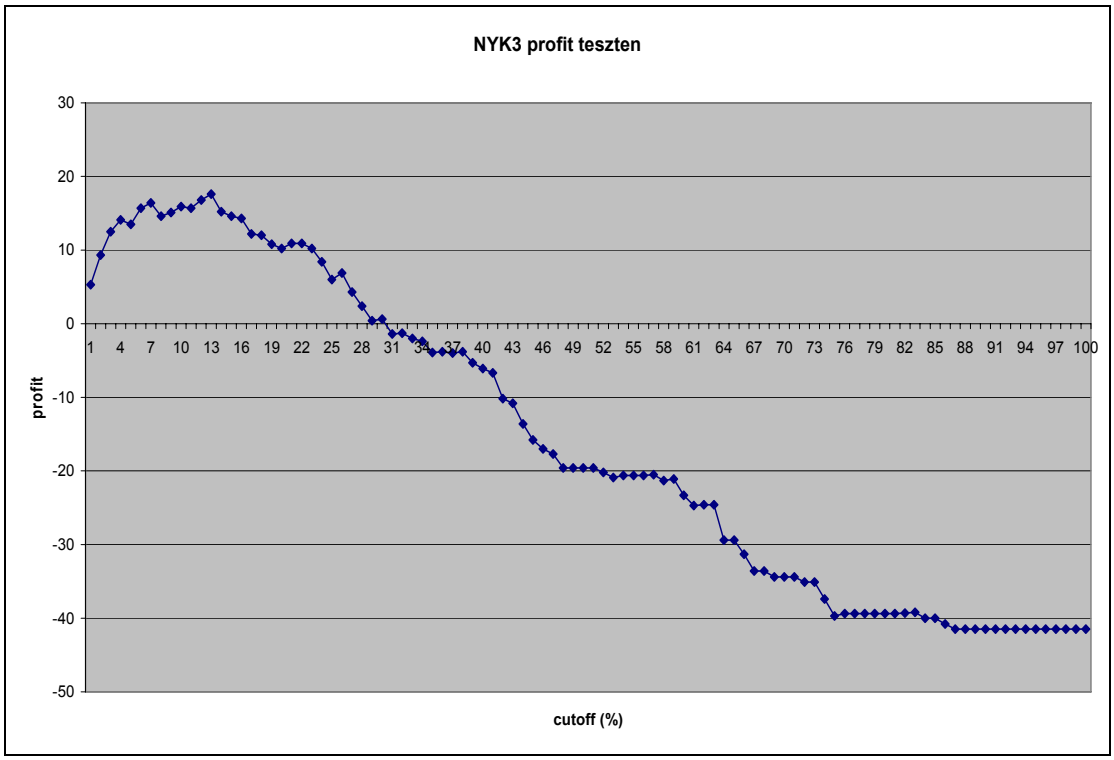

NYK3 profit tréningen
profit
cutoff (%)

4. Profitgörbék a tesztadatokon






NYK2 profit teszten
30
20
10
0
-10
-20
-30
-40
-50
profit
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64 67 70 73 76 79 82 85 88 91 94 97 100
cutoff (%)


NYK3 profit teszten
30
20
10
0
-10
-20
-30
-40
-50
profit
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64 67 70 73 76 79 82 85 88 91 94 97 100
cutoff (%)

Irodalomjegyzék

E.I. Altman [1968], "Financial Ratios, Discriminant Analysis, and the Prediction of Corporate Bankruptcy", *Journal of Finance, 1968 szeptember, Vol.23. 589-609.old.*

A. B. Anderson, A. Basilevsky és D.P.J. Hum [1983], " Missing data: A review of literature", In: P.H. Rossi, J.D. Wright és A.B. Anderson (Editors), Handbook of Survey Research Academic Press, San Diego, 415-494.o.

L. Asch [1995], "How the RMA/Fair, Isaac Credit-Scoring Model Was Built," *Journal of Commercial Lending*, 10-16.o., 1995. június.

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

T. Astebro és I. Bernhardt [2001], "Bank Loans as Predictors of Small Start-up Business Survival". Journal of Economics and Business 55 (4), 303-320.o.

T. Astebro és G. Chen [2000], "Missing Data Analysis for Single Choice and Multiple Choice Survey Questions When Data are Sparse," *Presented at 2000 Academy of Management Conference*, Symposium entitled "Much ado about missing data," Kanada, Toronto.

S. Azen és M. Van Guilder [1981], "Conclusions regarding algorithms for handling incomplete data" *Proceedings Statistical Computing Section, American Statistical Association.,*53-56.o

K. Baker, P. Harris, J. O'Brien [1989] Data Fusion: "An Appraisal and Experimental Evaluation", *Journal of Market Research Society, vol 31*., 153-212 o.

J.B. Banasik, J.N. Crook és L.C. Thomas [2001], "Sample Selection Bias in Credit Scoring Models," *Working paper 01/5*, Credit Research Centre, University of Edinburg, Anglia.

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

J. Barnard,és D.B. Rubin [1999], "Small-sample degrees of freedom with multiple imputation", *Biometrica,* 86. évf, 949-955.o.

H. Bierman és W. H. Hausman [1970], „The Credit Granting Decision," *Management Sciences*, 16. évf., 519-532.o.

Bódy Sándor, Sulyok Pap Márta [1997], "Cégminősítés", *Nemzetközi Bankárképző Központ,* Budapest

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

A.P. Bradley [1996], "ROC Curves and the χ^2 Test," *Pattern Recognition Letters*, 17. évf. 3. szám 287-294.o., 1996. március.

L. Breiman, J. H. Friedman, R.A. Olshen, C.J. Stone [1984], "Classification and Regression Trees", *Wadsworth International Group*, Belmont

N. A. L. Brooks [1989], "Expert Systems", *Bank Administration*, 1989. augusztus, 65.évf. 8.szám, 36-37.o.

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

G. Chen és T. Astebro [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, Anglia, 2001. szeptember.

G. Chen és T. Astebro [2003], "How to Deal with Missing Categorical Data: Test of a Simple Bayesian Method," *Organizational Research Methods*, 6. évf. 3. szám 309-321.o.

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

G. Chen és T. Astebro [2006], "A Maximum Likelihood Approach for Reject Inference in Credit Scoring", November 25, 2006, kézirat Available at SSRN: http://ssrn.com/abstract=872541

R.K. Chhikara [1989], "The State of the Art in Credit Evaluation", *American Journal of Agricultural Economics*, 71.évf, 5.szám, 1138-1144.o.

J.B. Copas és H.G. Li [1997], "Inference for Non-random Samples with discussion," *Journal of the Royal Statistical Society*, B, 59. évf. 55-95.o.

R. Cressy [1996], "Are Business Startups Debt-Rationed?" *The Economic Journal*, 106. évf. 1253-1270.o., 1996. szeptember.

J. Crook és J. Banasik [2002], “Does Reject Inference Really Improve the Performance of Application Scoring Models?” *Working paper 02/3*, Credit Research Centre, University of Edinburg, Anglia.

A.P. Dempster, N.M. Laird és D.B. Rubin [1977], “Maximum Likelihood Estimation From Incomplete Data Via the EM Algorithm,” *Journal of the Royal Statistical Society B*, 39. évf. 1-38.o.

S.G. Donald [1995], “Two-step Estimation of Heteroskedastic Sample Selection Models,” *Journal of Econometrics*, 65. évf. 347-380.o.

A. Donner, B. Rosner [1982],”Missing Value Problems in Multiple Linear Regression With Two Independent Variables”, *Communication in Statistics*, vol. 11, 127-140.o

H.L. Dunham [1938], “A Simple Credit Rating for Small Loans”, *Bankers Monthly*

D. Durand [1941], “Risk Elements in Consumer Instalment Lending”, *National Bureau of Economic Research, New York, Vol. study 8.*

B.S. Everitt és D.J. Hand [1981], “Finite Mixture Distributions,” *Chapman and Hall*, London.

A.J. Feelders, S. Chang és G.J. McLachlan [1998], “Mining in the Presence of Selectivity Bias and its Application to Reject Inference,” *Proceedings of the fourth international conference on knowledge discovery and data mining [KDD-98]*, AAAI Press, 199-203.o.

A.J. Feelders [1999], “Credit Scoring and Reject Inference with Mixture Models,” *International Journal of Intelligent System in Accounting, Finance and Management*, 8 évf. 271-279.o.

A.J. Feelders [2000], “Credit Scoring and Reject Inference with Mixture Models,” *International Journal of Intelligent System in Accounting, Finance and Management*, 9. évf. 1-8.o.

A.J. Feelders [2001], “An Overview of Model Based Reject Inference for Credit Scoring,” *Working paper*, Institute for Information and Computing Sciences, Utrecht University, Hollandia.

D.J. Forgarty [2005], “Multiple Imputation as a Missing Data Approach to Reject Inference on Consumer Credit Scoring”, *Manuscript*, DavidF1967@email.uopx.edu

N. Freed és F. Glover [1981], “A Linear Programming Approach to the Discriminant Problem”, *Decision Science*, 12.évf, 68-74.o.

N. Freed és F. Glover [1981] “Simple but Powerful Goal Programming Approach to the Discriminant Problem” European Journal of Operacional Research, 7.évf., 44-60.o.

J.H. Friedman [1997], “On Bias, Variance, 0/1-loss, and the Curse-of-dimensionality,” *Data Mining and Knowledge Discovery*, 1. évf. 1. szám 55-77.o.

H. Friedman, E. Altman és D.L. Kao [1985], “Introducing Recursive Partitioning for Financial Classification: The Case of Financial Distress”, *Journal of Finance*, 1985. március, Vol. 40. Iss.1, 269-291 old.

A. Gelman, J.B. Carlin, H.S. Stern és D.B. Rubin [1995], “Bayesian Data Analysis”, *Chapman & Hall*, London

W.R. Gilks, S. Richardson, és D. J. Spiegelhalter Eds. [1996]. „ Markov Chain Monte Carlo in Practice”, *Chapman & Hall*, London.

R. Glynn, N. M. Laird és D.B. Rubin [1986], „Selection modeling versus mixture modeling with nonignorable nonresponse”, *In H. Wainer [ed.] Drawing Inferences from Self-Selected Samples,* 119-146. New York: Springer-Verlag.

J.W. Graham és S.I. Donaldson [1993], “Evaluating Interventions with Differential Attrition: the Importance of Nonresponse Mechanisms and Use of Followup Data,” *Journal of Applied Psychology*, 78. évf. 119-128.o.

W.H. Greene [1992], “A Statistical Model for Credit Scoring,” *Working Paper EC-92-29*, Leonard N. Stern School of Business.

W.H. Greene [1993], “Econometric Analysis [second edition],” *Macmillan*, New York.

W.H. Greene [1998], “Sample Selection in Credit-scoring Models,” *Japan and the World Economy*, 10. évf. 299-316.o.

György Erika [2004], “A nemválaszolás elemzése a munkaerő felvételben”, *Statisztikai szemle,* 82. évf.8.szám, 747-772.oldal

Hajdu Ottó [2003]: „Többváltozós statisztikai számítások”, KSH, Budapest

Hajdu Ottó és Virág Miklós [1996], „Pénzügyi mutatószámokon alapuló csődmodell-számítások”, *Bankszemle,* 1996/1-2

Hámori Gábor [2001], „A CHAID alapú döntési fák jellemzői”, *Statisztikai szemle,* 79évf. 8.szám, 703-710.o.

D.J. Hand és W.E. Henley [1993/4], „Can Reject Inference Ever Work?," *IMA Journal of Mathematics Applied in Business & Industry*, 5. évf. 4. szám 45-55.o.

D.J. Hand és W.E Henley [1994], „Inference About Rejected Cases in Discriminant Analysis," *Springer*, 292-299.o., New York.

D.J. Hand [1997], „Construction and Assessment of Classification Rules," *Chichster: Wiley*.

D.J. Hand [1998], "Reject Inference in Credit Operations," *in Credit Risk Modeling: Design and Application* [ed. E. Mays], 181-190.o. AMACOM.

D.J. Hand [2001], "Measuring Diagnostic Accuracy of Statistical Prediction Rules," *Statistica Neerlandica*, 53. évf. 3-16.o.

W.E. Hardy és J.L. Adrian [1985]," A Linear Programming Alternative to Discriminant Analysis in Credit Scoring", *Agribusiness*, 1.évf., 4.szám, 285-292.o.

J.J. Heckman [1979], "Sample Selection Bias as a Specification Error," *Econometrica*, 47. évf. 153-161.o.

D. Hedeker és R. D. Gibbons [1997], „Application of random-effects pattern-mixture models for missing data in longitudinal studies"*, Psychological Methods,* 2[1], 64-78.

C.W. Holsapple et.al. [1988], "Adapting Expert System Technology to Financial Management" *Financial Management*, 1988. ősz, 19.évf. 12-22.o.

D. Holt, T.M.F. Smith és P.D. Winter [1980], "Regression Analysis of Data From Complex Surveys," *Journal of the Royal Statistical Society*, 143. évf. A.

D.C Hsia [1978], "Credit Scoring and the Equal Credit Opportunity Act," *The Hastings Law Journal*, 30. évf. 371-448.o., 1978. november.

Hunyadi László [2001], "A mintavétel alapjai", *Egyetemi Jegyzet SZÁMALK,* Budapest

Hunyadi László és Vita László [2002], "Statisztika közgazdászoknak", *Központi Statisztikai Hivatal,* Budapest

T. Jacobson és K. Roszbach [1998], "Bank Lending Policy, Credit Scoring and Value at Risk," *SSE/EFI Working Paper Series in Economics and Finance 260*, Stockholm School of Economics.

T. Jacobson és K.F. Roszbach [1999], "Evaluating Bank Lending Policy and Consumer Credit Risk," *in Computational Finance 1999* [edited by Y.S. Abu-Mostafa et al.] the MIT Press, 2000.

D.N Joanes [1993], "Reject Inference Applied to Logistic Regression for Credit Scoring," *IMA Journal of Mathematics Applied in Business and Industry*, 5. évf. 4.szám 35-43.o.

N.M. Kiefer és C.E. Larson [2003], "Specification and Informational Issues in Credit Scoring", kézirat

J.O. Kim és J. Curry [1977], „ The treatment of missing data in multivariate analysis", *Sociol. Meth. Res.*, 6.évf, 215-240.o.

Kiss Ferenc [2003], "A credit scoring fejlődése és alkalmazása", *Ph.D. értekezés,* BME

P.W. Lavori, R. Dawson és D. Shera [1995], „ A multiple imputation strategy for clinical trials with truncation of patient data"*, Statistics in Medicine*, 14 évf., 1913-1925.o.

S. Y. Lee, Y.M. Chiu [1990], "Analysis of Multivariate Polychoric Correlation Models with Incomplete Data", *British Journal of Mathematical and Statistical Psychology*, vol. 43., 145-154. o

K.J. Leonard [1992], "Credit-scoring Models for the Evaluation of Small-business Loan Applications," *IMA Journal of Mathematics Applied in Business & Industry*, 4. évf. 89-95.o.

L.A. Lipsitz, I. Nakajima, M. Gagnon, T. Hirayama, C.M. Connelly, H. Izumo és T. Hirayama [1994], "Muscle Strength and Fall Rates Among Residents of Japanese and American Nursing Homes: an International Cross-Cultural Study," *Journal of the American Geriatrics Society*, 42. évf. 9. szám 953-959.o.

R.J.A. Little [1979], "Maximum Likelihood Inference for Multiple Regression with Missing Values: A Simulation Study", *Journal of the Royal Statistical Society*, vol.41., 76-87.o.

R.J.A. Little és D.B. Rubin [1987], "Statistical Analysis with Missing Data", *John Wiley & Sons*, New York.

R.J.A. Little és D.B. Rubin [2002], "Statistical Analysis with Missing Data," 2. Edition, *John Wiley & Sons*, New York.

R.J.A. Little [1993], "Pattern-mixture Models for Multivariate Incomplete Data," *Journal of the American Statistical Association*, 88. évf. 125-134.o.

R.J.A. Little [1995], "Modeling the Dropout Mechanism in Repeated-measures Studies," *Journal of the American Statistical Association*, 90. évf. 1112-1121.o.

R.J.A. Little és N. Schenker [1994], "Missing Data" in *Handbook for Statistical Modeling in the Social and Behavioral Sciences* [G. Arminger, C. C. Clogg és M. E. Sobel szerk.] New York: Plenum 39-75.o.

G.S. Maddala [1983], "Limited Dependent and Qualitative Variables in Econometrics," *Cambridge University Press*, Cambridge, UK.

Máder Miklós Péter [2005],"Imputálási eljárások hatékonysága", *Statisztikai Szemle*, 83.évf. 7.szám, 628-644.o.

N.K. Malhotra [1987], "Analyzing Marketing Research Data with Incomplete Information on the Dependent Variable", *Journal of Marketing Research*, vol.24, 74-84.o

O.L. Mangasarian [1965], "Linear and nonlinear separation of patterns by linear programming", *Operation Research*, 13 évf., 444-452.o.

M.L. Marais, J.M. Patell és M.A. Walfson [1984], "The Experimental Design of Classification Models: An Application of Recursive Partitioning and Bootsrapping to Commercial Bank Loan Classifications", Journal of Accounting Research, 1984 Vol.22, 87-115o.

E. Mays [2004], "Credit Scoring for Risk Managers" *South Western Thomson Learning*

M. McDermit, R. Funk, M. Dennis [1999], "Data Cleaning And Replacement of Missing Values", manuscript

G.J. McLachlan és K.E. Basford [1988], "Mixture Models, Inference and Applications to Clustering," *Marker Dekker*, New York.

G.J. McLachlan [1992], "Discriminant Analysis and Statistical Pattern Recognition," *Wiley*, New York.

R.W. McLeod et.al. [1993], "Predicting Credit Risk: A Neural Network Approach", *Journal of Retail Banking,* 15.évf., 3.szám, 37-40.o.

X.L. Meng [1995], „Multiple-imputation inferences with uncongenial sources of input [with discussion]", *Statistical Science*, 10.évf., 538-573. o.

C.L. Meng és P.Schmidt [1985], "On the Cost of Partial Observation in the Bivariate Probit Model," *International Economic Review*, 26. évf. 1. szám 71-85.o., 1985. február.

R. Nath, W.M. Jackson és T.W. Jones [1992], "A comparison of the classical and the linear programming approaches to the classification problem in discriminant analysis" *Journal of Statistical and Computational Simulations*, 41.évf., 73-93. o.

M.C. Paik, R. Sacco és I.F. Lin [2000], "Bivariate Binary Data Analysis with Nonignorably Missing Outcomes," *Biometrics*, 56. évf. 1145-1156.o.

L.F. Pau ed. [1986], "Artificial Intelligence in Economics and Management" *North-Holland Publishing Co*., Amsterdam

D.J. Poirier [1980], "Partial Observability in Bivariate Probit Model," *Journal of Econometrics*, 12. évf. 209-217.o.

P.L. Roth és F.S. Switzer III [1995], "A Monte Carlo Analysis of Missing Data Techniques in a HRM Setting," *Journal of Management*, 21. évf. 5. szám 1003-1023.o.

D.B. Rubin [1976], "Inference and Missing Data," *Biometrika*, 63. évf. 581-592.o.

D.B. Rubin [1987], "Multiple Imputation for Nonresponse in Surveys," *John Wiley & Sons*.

D. B. Rubin [1996], "Multiple imputation after 18+ years [with discussion]", *Journal of the American Statistical Association*, 91, 473-489.o.

D. B. Rubin [2003], "Nested Multiple Imputation of NMES via Partially Incompatible MCMC", *Statistica Neerlandica,* 57, 3-18.o.

Rudas T. [1998],"Hogyan olvassunk közvélemény-kutatásokat?", Új Mandátum Könyvkiadó, Budapest

J.L. Schlafer [1997], "Analysis of Incomplete Multivariate Data," *Chapman & Hall*, London.

J.L. Schafer és J.W. Graham [2002], "Missing Data: our View of the State of the Art," *Psychological Methods*, 7. évf. 2. szám 147-177.o.

J.L. Schafer és M.K. Olsen [1998],"Multiple imputation for multivariate missing-data problems: a data analyst's perspective", *Multivariate Behavioral Research*, 33. évf., 545-571. o.

P. Sebastiani és M. Ramoni [2000], "Bayesian Inference with Missing Data Using Bound and Collapse," *Journal of Computational and Graphical Statistics*, 9. évf. 4. szám 779-800.o.

V. Srinivasan és Y. H. Kim [1987], "Credit Granting: A Comparative Analysis of Classification Procedures" *Journal of Finance*, 1987. július, 42.évf. 3.szám, 665-683.old.

Statistical Solutions, Inc. [1998], „SOLAS for Missing Data Analysis”, *Version 1. Cork, Ireland: Statistical Solutions.*
K.Y. Tam (1991), “Neural Network Models and the Prediction of Bank Bankruptcy” *Omega. The International Journal of Management Science*, 19.évf, 5.szám 429-445.o

K.Y. Tam és M.Y. Kiang [1992], “Managerial Applications of Neural Networks: The Case of Bank Failure Predictions”. *Management Science*, 1992. július, 38.évf. 7.szám, 926-947o.

L.C. Thomas [2000], “A Survey of Credit and Behavioural Scoring: Forecasting Financial Risk of Lending to Consumers,” *International Journal of Forecasting*, 16. évf. 149-172.o.

L.C. Thomas, D.B. Edelman és J.N. Crook [2002], “Credit Scoring and Its Applications”, *Society for Industrial and Applied Mathematics, Philadelphia*

Virág M. és Kristóf T. [2006], “Iparági rátákon alapuló csődelőrejelzés sokváltozós statisztikai módszerekkel”, *Vezetéstudomány*, 37. évf., 1.szám. 25-35.o.

D. West [2000], ”Neural Network Credit Scoring Models”, *Computers and Operations Research,* 27-évf. 1131-1152.o

J.C. Wiginton [1980], ”A note on the comparison of logit and discriminant models of consumer credit behaviour”, *Journal of Financial Quantitative Anal.,15,* 757-770.o.

W. Wothke [1998], ”Longitudinal and multi-group modeling with missing data”, *Mahwah, NJ: Lawrence Erlbaum Associates.*

H.A. Ziari, D.J. Leatham és P.N. Ellinger [1997], ”Development of statistical discriminant mathematical programming model via resampling estimation techniques”, *American Journal of Agricultural Economics, 79, 1352-1362.o.*

A témakörrel kapcsolatos saját publikációk jegyzéke

Folyóiratcikkek:

Oravecz Beatrix (2007): Credit scoring modellek és teljesítményük mérése. Hitelintézeti Szemle, 6.évf. 6.szám, 607-627.o.

Oravecz Beatrix : Hiányzó adatok és kezelésük a statisztikai elemzésekben. Statisztikai Szemle, várható megjelenés 2008. január

Oravecz Beatrix: Szelekciós torzítás és csökkentése a credit scoring modelleknél. Hitelintézeti Szemle, várható megjelenés 2008. február

Cikk szerkesztett könyvben:

Oravecz Beatrix (2004): Imputációs eljárások. Egy reneszánsz statisztikus. Tanulmánykötet Hunyadi László tiszteletére, KSH

Egyéb:

Oravecz Beatrix (2000): Hitelmonitoring, Szakdolgozat, BKAE Pénzügy Tanszék

Oravecz Beatrix (2002): Dijkstra, W.: Új módszer az interjúk közbeni kölcsönkapcsolatok tanulmányozására. cikkismertetés, Statisztikai Szemle, 79. évf. 7. szám, 636-637.o.