„MINDEN MODELL ROSSZ, DE NÉMELYIKÜK HASZNOS" HITELEZÉSI SCORING MODELLEK MODELLEZÉSI KOCKÁZATA

Pollák Zoltán – Kocsis Ádám

A pénzügyek világában a döntések támogatására modelleket alkalmazunk, mert a valóság teljes egészében történő megfigyelése lehetetlen. A 2008-as válság élesen felszínre hozta az alkalmazott modellek hibáit és ráirányította a figyelmet a modellkockázat jelentőségére. Tanulmányunk célja speciálisan az adósminősítési modellek modellezési kockázatának számszerűsítése. Ennek szellemében először bemutatjuk a modellhibák okozta lehetséges, portfóliószintű veszteségek meghatározásának módját. Az így nyert veszteségeloszlás széleiben rejlő, minél több információt felhasználva, az extrémérték-elmélet segítségével megadhatóak a modellezési kockázat különböző kockázati mértékei. Az elméleti áttekintést követően az ismertetett eljárásokat egy nyilvánosan elérhető adatbázison, R segítségével mutatjuk be.[1]

JEL-kódok: C01, C13, C19, C25, C52, C58, G21, G28

Kulcsszavak: modellkockázat (model risk), adósminősítés (credit scoring), reject inference, extrémérték-elmélet (extreme value theory), kockáztatott érték (Value-at-Risk – VaR), várható hiány (expected shortfall)

1. A MODELLEZÉSI KOCKÁZAT

Az üzleti életben végbemenő folyamatokat komplexitásukból adódóan minden részletre kiterjedően megfigyelni, leírni lehetetlen. Ezért aztán modelleket alkotunk, amelyekben ismereteinket rendszerezhetjük, tömöríthetjük. Ebből az egyszerűsítésből adódóan mindig szem előtt kell tartanunk, hogy „makettünk" csak a valóság egy kicsinyített mása, és nem azonos azzal. A modell alapján levont következtetéseink pedig csak lokálisan, adott keretek között érvényesek, ezért mindenképpen szükséges, hogy teljesüljenek az előzetesen támasztott feltevéseink.

1 A tanulmány az „Innovatív matematikai modellek kutatása a bázeli banki kockázatok mérésére és tőkekövetelmény számszerűsítésére a piaci, működési, likviditási és másodlagos kockázatok területén; valamint pénzügyi termékek áralakulásának viselkedés-alapú előrejelzése" című Új Széchenyi Terv keretében finanszírozott kutatásfejlesztés során (PIAC_13-1-2013-0073 számú projekt), európai uniós támogatás mellett valósult meg.

A néhány éve kirobbant gazdasági világválság élesen felszínre hozta a korábban használt modellek hibáit, és önmagában véve a modellkockázat kezelésének fontosságát. Dolgozatunk címe egy *George E. P. Box* tollából származó idézet. A 2013-ban elhunyt brit statisztikus nagyon találóan világított rá (még a válság kirobbanása előtt) a modellezési kockázat jelentőségére: *„Lényegében minden modell rossz, de némelyikük hasznos."* (*Box–Draper*, 2007, p. 414.)

A hasznos modellekre tehát szükségünk van, de mivel azok egyszerűsítő jellegükből adódóan hibáznak, érdemes különös hangsúlyt helyezni a modellezési kockázatra. Ezt a pénzügyi szektort szabályozó szervezetek is felismerték, és elvárják a felügyeletük alá tartozó intézményektől ennek figyelembevételét kockázataik mérésénél.

1.1. Rövid szabályozási kitekintő

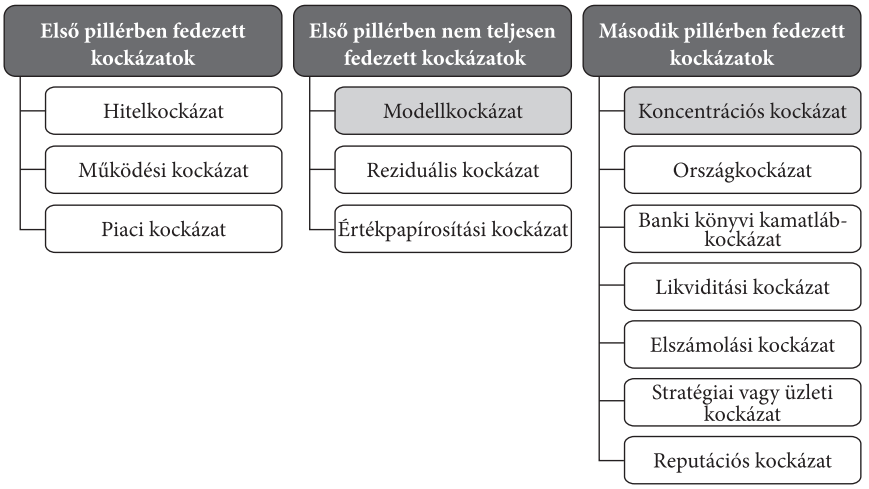
A hitelintézetekre és befektetési vállalkozásokra vonatkozó, bázeli szabályozás második pillére egy belső tőkeszámítást (Internal Capital Adequacy Assessment Process – ICAAP) ír elő, ahol minden (a szabályozás hatálya alá tartozó) intézménynek saját magának kell felmérnie összes releváns kockázatát. Mivel itt a mért kockázatok köre bővebb, mint az első pillérben, az esetek többségében magasabb tőkekövetelmény adódik a szabályozói minimumhoz képest. Amennyiben azonban kevesebb lenne a belső modell által kalkulált tőkekövetelmény az első pillérhez képest – és ezt az MNB felügyeleti szerve a felügyeleti felülvizsgálati folyamat (Supervisory Review and Evaluation Process – SREP) során jóváhagyja –, nem szükséges addicionális tőkét képeznie a pénzügyi intézménynek.

A második pillérben a felügyelet arra szeretné ösztönözni az intézményeket, hogy tudatosabban foglalkozzanak a saját kockázataik mérésével, alkalmazzanak korszerűbb, pontosabb kockázatkezelési technikákat. Az így nyert tudás pedig a folyamatokba ágyazódva támogathatja az adott intézmény prudens működését, amely nemcsak felügyeleti kívánalom, hanem valamennyi érintett (stakeholder) érdeke.

A szabályozó a hitelintézetektől és befektetési vállalkozásoktól minimálisan a következő kockázatok figyelembevételét és mérését várja el:

1. ábra

A modellkockázat elhelyezkedése a banki kockázatok között



Forrás: saját ábra az MNB *„A felügyeleti felülvizsgálati folyamat (SRP)"* című módszertani útmutatója alapján

A továbbiakban tanulmányunk a modellkockázattal foglalkozik részletesebben, amelyet – ahogyan az *1. ábrán* is látszik – az első pillérben nem teljesen vesznek számba, így az intézményeknek mindenképpen át kell gondolniuk helyes kezelését, és ennek megfelelően kell elvégezniük belső tőkeszámításukat.

A modellezési kockázat *„annak a kockázata, hogy a modellek hibáiból kifolyólag gazdasági veszteséget okozó döntéseket (például elbírálás, árazás) hoz az intézmény."* (MNB, 2012b, p. 25.). A szabályozás kiemeli, hogy nem elsősorban az emberi hanyagság okozta modellhibák tartoznak bele a fogalomba, hanem a pénzügyi folyamatokban végbemenő olyan változások miatt bekövetkező veszteségek is, amelyek múltbeli adatokból nem olvashatóak ki. Ez a fajta kockázat önmagában azért merül fel, mert alkalmazott modelljeink sosem lehetnek tökéletesek.

A modellezési kockázatot kvantifikálni nagyon nehéz. A modellhibákat stressz–tesztekkel, valamint érzékenységvizsgálatokkal meg lehet becsülni, ugyanakkor ezeket veszteséggé alakítani annál nehezebben kivitelezhető feladat. Ezt szem előtt tartva, a szabályozó nem elsősorban addicionális tőke tartását várja el, hanem inkább folyamati kezelésmódot javasol a felügyelt intézmények számára.

A pénzügyek világában számos területen alkalmaznak modelleket. A következőkben a modellkockázat egy részterületére, a hitelezési scoring modellekkel kapcsolatban felmerülő kockázatokra fogunk fókuszálni.

1.2. Hitelezési kockázat és a scoring modellek

A hitelintézetek hagyományos tevékenységei közé tartozik a betétgyűjtés, valamint a hitelnyújtás. Amíg azonban a bank az egyik oldalon szinte korlátozás nélkül elfogad betétet, addig a másik oldalon nagyon megválogatja, hogy kinek is adjon kölcsönt, és milyen feltételekkel.

A bankok az aktíváik között található hitelportfólió esetleges veszteségeinek ellentételezésére kötelesek a bázeli szabályozás első pillérének keretében tőkét képezni. A hitelkockázat *„annak kockázata, hogy a kötelezett részben vagy egyáltalán nem fizeti vissza a kötelezettségeit akkor, amikor azok esedékessé válnak"* (*Radnai–Vonnák*, 2010, p. 14.).

A probléma jelentőségét mi sem jelzi jobban, minthogy a hitelezési kockázatra képzik a bankok teljes tőkéjük kétharmadát, nem ritka esetben háromnegyedét is, amivel a hitelkockázat a legjelentősebb banki kockázat (*Krekó*, 2011). Ebből is látszik, hogy a pénzügyi intézményeknél kardinális kérdés a probléma minél átfogóbb kezelése.

A keletkező veszteségek megelőzésének számos módja van a banki gyakorlatban. Ilyen például az, amikor a bank limiteket alkalmaz egyes intézményeknek, vagy akár iparágaknak nyújtandó hitelösszegre. Így elkerülhető a túlzottan koncentrált kihelyezésekből eredő hitelkockázat. További kockázatcsökkentő módszer lehet a fedezetek megkövetelése, amelyeknek az értékesítéséből a bank kielégítést nyerhet az adós nemfizetése esetén.

A hitelkockázat kezelésének legalapvetőbb módja azonban, ha a bank előzetesen próbálja minél hatékonyabban felmérni, hogy a hiteligénylők vissza tudják-e majd fizetni a nyújtott kölcsönt és annak kamatait. A potenciálisan jó, illetve rossz ügyfelek szétválasztásában (vagy más néven a klasszifikációban) nyújtanak segítséget a credit scoring modellek.

Az adósminősítés tulajdonképpen egyidős a hitelezési tevékenységgel. A 20. század első feléig azonban tisztán szakértői alapon, a statisztika eszköztárának használata nélkül bírálták el a hitelkérelmeket. A nagy áttörés 1941-ben következett be, amikor *David Durand* egy diszkriminanciaanalízisen alapuló, pontozásos rendszert alkalmazott autóvásárlási hitelt felvenni készülő magánszemélyekre (*Kiss*, 2003).

Napjainkban az alkalmazott statisztikai eljárások között egyértelműen legelterjedtebb a logisztikus regresszió, melyet először *Delton L. Chesser* javasolt 1974-ben az adósok várható nemteljesítésének előrejelzésére.

Azóta modellek sora látott napvilágot, amelyek anélkül is képesek hatékonyan megoldani a klasszifikáció problémáját, hogy bármilyen előzetes feltételezésünk lenne a sokasággal kapcsolatban. Éppen ez az automatizált jelleg a legfőbb veszé-

lye ezen modelleknek, hiszen sokszor képesek fekete dobozként működni, az azt működtetők pedig néha át sem gondolják a bennük rejlő veszélyeket.

A következőkben megpróbáljuk felderíteni a hitelezési scoring modellek azon gyenge pontjait, ahol fellelhető a modellkockázat. Az első ilyen sarkalatos pont az alapadatok reprezentativitásának kérdése.

2. MODELLKOCKÁZAT MINT A HIÁNYZÓ ADATOK PROBLÉMÁJA

Modellkockázati szempontból a legalapvetőbb probléma nem az adósminősítő modellekben keresendő, hanem magukban az alapadatokban. Használhatunk bármilyen szofisztikált modellt annak eldöntésére, hogy mely ügyfeleknek adjunk kölcsönt, ha már a modellezéshez használt adataink sem megfelelőek. A problémát mintánkkal kapcsolatban a szakirodalomban szelekciós torzításként emlegetett jelenség jelenti (*Little–Rubin*, 2002).

A szelekciós torzítás azért lép fel, mert a modellezéshez használt minta általában nem reprezentatív, mivel csak azon ügyfelek esetében adhatunk minden változónknak értéket, akik már átestek egy kiválasztási folyamaton (kaptak hitelt). Amely ügyfelek nem jutottak kölcsönhöz, azokról nincsen információnk, hogy vajon teljesítették volna-e fizetési kötelezettségeiket.

Egy olyan fiktív intézményben, ahol a hiteligénylők egy pénzérme segítségével, fej vagy írás alapon kapnak hitelt, jogos feltételezés lehetne, hogy a hitelt megkapó ügyfelekre az egyes változók eloszlása azonos az elutasítottakéval, vagyis a minta reprezentálja a teljes sokaságot. A gyakorlatban azonban a bankok különböző modellek segítségével próbálják előre jelezni, hogy az adott ügyfél jó vagy rossz (csődös) lesz. A befogadás így nem véletlenszerű, vagyis a meghitelezettek (akiknek az adataiból modellt építünk) nem fogják megfelelően reprezentálni az összes kérelmezőt.

A szelekciós torzítás ráadásul általában egy rossz irányú kockázat (wrong-way risk). Ennek belátására folytassuk az imént megkezdett gondolatmenetet. A bankok többnyire kifinomult credit scoring modelljei feltehetően jobb kiválasztást biztosítanak a fej vagy írás alapú hitelminősítésnél. Ebben az esetben reális feltételezés lehet, hogy a befogadottak körében nagyobb arányt képviselnek a jó ügyfelek, mint az elutasítottak körében. Így tehát egy olyan mintán fogjuk elvégezni a modellépítést, amelyben a jó ügyfelek felülreprezentáltak, az arányaiban kevesebb rossz ügyfél bekerülése pedig azt eredményezi, hogy kisebb megbízhatósággal lesz képes a modellünk felismerni a rossz ügyfelek karakterisztikáját, mintha a sokaságot teljesen reprezentáló, arányaiban több rossz ügyfelet magában foglaló mintánk lenne.

A probléma figyelmen kívül hagyása esetén romolhat az adósminősítési modellek klasszifikációs képessége, a modellhibák (vagyis a téves besorolás) miatt pedig vesztesége keletkezhet a hitelintézetnek.

A szelekciós torzítás problémájának kezelésére különféle – összefoglalóan *reject inference*[2]-nek nevezett – technikákat javasol a szakirodalom, amelyeknek az a lényege, hogy be kell építeni az elutasítottakat is a modellbe, például úgy, hogy egy becslést adunk azok viselkedésére, ha kaptak volna hitelt.

2.1. Az adathiány fajtái

A hiányzó adatok kezelése egy viszonylag új tudományterülete a statisztikának. Az 1970-es évek elején, a számítástechnika fejlődésével jelentek meg az első törekvések a probléma átfogóbb orvoslására. A következőkben szeretnénk röviden áttekinteni az adathiány legalapvetőbb típusait, hogy a későbbiekben a most bevezetett elnevezés- és fogalomrendszert tudjuk használni a probléma modellkockázati vonatkozásainak elemzésekor.

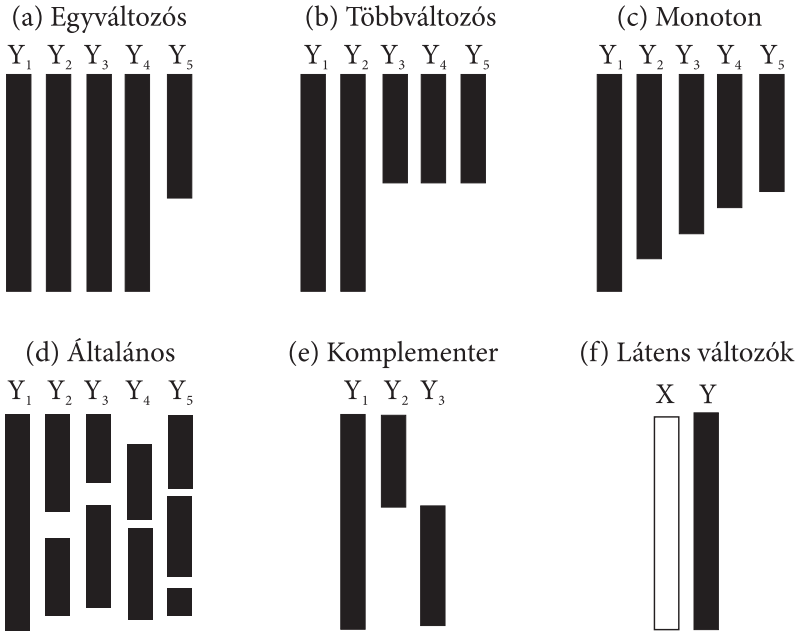
Az adathiány leírásának egyik megközelítése, amikor annak mintázatát próbáljuk meghatározni (*Little–Rubin*, 2002). Hagyományosan egy credit scoring rendszer építésénél az adatmátrixunk úgy épül fel, hogy soraiban találhatjuk az egyes megfigyeléseket, az oszlopokban pedig a vizsgált változókat. Ezek a változók lehetnek hiánytalanul feltöltve (nyilvánvalóan ez az ideális állapot), vagy tartalmazhatnak adathiányokat egyes megfigyelési egységekre nézve. Attól függően, hogy ezek a hiányzó értékek hogyan helyezkednek el alapadatmátrixunkban, hatféle különböző adathiány-mintázatról (missing-data pattern) beszélhetünk.

A 2. *ábrán* az (a) eset mutatja az egyváltozós adathiányt, amikor a hiányzó értékek egyetlen változó tekintetében jelentenek gondot, a többi változó teljes. A dolgozatunkban vizsgált kérdés tipikusan ezzel a mintázattal írható le, hiszen esetünkben egyetlen változóban, a hitelkockázatot megtestesítő magyarázó változóban hiányoznak értékek (azok esetén, akiket elutasítottak), a hiteligénylők karakterisztikáját leíró, többi változó minden megfigyeléshez rendel értéket. Itt természetesen azzal a feltételezéssel élünk, hogy a hiteligényléskor minden kérelmező hiánytalanul kitöltött valamennyi bank által kért adatot. Fókuszáljunk tehát a továbbiakban az egyváltozós adathiány esetére!

2 A szelekciós torzítás csökkentésére irányuló módszerek összefoglaló elnevezése. A hazai szakirodalomban nem jelent meg magyar nyelvű fordítása, az eredeti angol kifejezés az elterjedt.

2. ábra

Példák adathiány-mintázatra*



Megjegyzés: *Minden változónál (oszlopok) a beszínezett rész jelöli a megfigyelt értékeket.
Forrás: Little–Rubin (2002)

Egy másik megközelítés szerint akkor lehet megfelelően felmérni, kezelni az adathiányt, ha rendelkezünk valamilyen ismerettel a hiányzás és az egyes változók kapcsolatrendszeréről, vagyis akkor, ha tudjuk, hogy milyen folyamat vezetett az adathiány kialakulásához. Három fő csoportba oszthatjuk az eseteket (ezek az adathiány-mechanizmusok – missing-data mechanism), attól függően, hogy mennyire véletlenszerű az adathiány (*Oravecz*, 2008):

– Teljesen véletlenszerű adathiány (Missing Completely at Random – MCAR)
– Véletlenszerű adathiány (Missing at Random – MAR)
– Nem véletlenszerű adathiány (Missing not at Random – MNAR)

Jelölje $Y = (y_{ij})$ az adatmátrixunkat, amelyben n darab megfigyelés K darab változó szerinti értéke található. Vezessünk be egy $M = (m_{ij})$ indikátormátrixot is, amely m_{ij} elemeinek értéke 1, ha az adat hiányzik és 0, ha megfigyelt. Formálisan az adathiány mibenléte leírható M adott Y melletti feltételes eloszlásával ($f(M|Y, \theta)$, ahol θ ismeretlen paramétereket jelöl (*Little–Rubin*, 2002). Teljesen véletlenszerű adathiányról (MCAR) akkor beszélünk, ha a teljes körűen megfigyelt

egyedek és a hiányosan megfigyeltek eloszlása azonos, vagyis a korábban definiált feltételes eloszlásban *M* mátrix nem függ *Y*-tól:

$$f(M|Y,\theta) = f(M|\theta)\ teljesül\ \forall Y,\theta\ esetén. \quad (1)$$

Ilyen adathiány-mechanizmussal találkozunk, ha például a bank a korábban említett fej vagy írás alapon dönti el, hogy egy igénylő kapjon-e hitelt.

Véletlenszerű adathiánnyal (MAR) akkor van dolgunk, ha az adathiányra a hiányos változóból nem tudunk következtetni, de előre jelezhető a többi (teljes) változó segítségével.

$$f(M|Y,\theta) = f\left(M\middle|Y_{megfigyelt},\theta\right)\ teljesül\ \forall Y_{hiányzó},\theta\ esetén\ , \quad (2)$$

ahol $Y_{megfigyelt}$ az *Y* mátrix hiánytalan megfigyeléseket tartalmazó komponense, míg $Y_{hiányzó}$ azon rész, ahol felbukkan az adathiány. A (2) egyenletnek megfelelő, véletlenszerű adathiányt mutatja be a következő eset.

Tegyük fel, hogy rendelkezünk egy mintával, amelyben ügyfeleink hiteligényléshez bekért adatai hiánytalanul rendelkezésre állnak. Ezután egy credit scoring modellt építünk, amely alapján eldöntjük, hogy mely ügyfelek kapjanak hitelt. A hitelnyújtást követően megfigyeljük, hogy mintánkból mely egyedek fizették, illetve kik nem teljesítették kötelezettségeiket (default). Ez utóbbi, a hitelkockázatot megtestesítő paraméter esetén természetesen az elutasítottak körében hiányzó értékeket találunk, de mivel egy egyértelmű, jól dokumentált metódus alapján választottuk ki, hogy kik kapjanak hitelt, az adathiányra tudunk következtetni a többi (teljes) változó segítségével, hiszen hitelezési scoring modellünket is ezen teljes változók segítségével építettük.

Az imént leírt példában a hangsúly azon van, hogy a hitelelbírálás világosan lefektetett szabályok szerint történt. Amennyiben ad hoc kiválasztási elemeket is beleviszünk a szelekciós algoritmusba (például kivételágon kapnak egyesek kölcsönt), adathiányunkra nem lehet következtetni a többi változó segítségével, így átcsúszunk a következő kategóriába, amely már kedvezőtlenebb tulajdonságokkal rendelkezik.

Ez a típus a nem véletlenszerű adathiány (NMAR), amely tehát azt jelenti, hogy a nem teljes változó adathiányára nem tudunk következtetni a többi változóból. Ez a verzió az adathiány legnehezebben kezelhető esete.

Az adathiány-mechanizmusok felismerése nagyon fontos feltétele a probléma megfelelő kezelésének, illetve az abból adódó kockázat (vagy bizonytalanság) számszerűsítésének.

2.2. Egy logisztikus regressziós imputációs modell Y_i dichotóm változóra

Az eddigiekben igyekeztünk rávilágítani arra, hogy a szelekciós torzítás jelensége egy gyakorlatban létező és igen jelentős probléma az adósminősítő modelleknél. Általános érvényű módszer a torzítás csökkentésére nincs, különböző szempontok szerint mérlegelni kell tehát, hogy milyen eljárást is válasszunk.

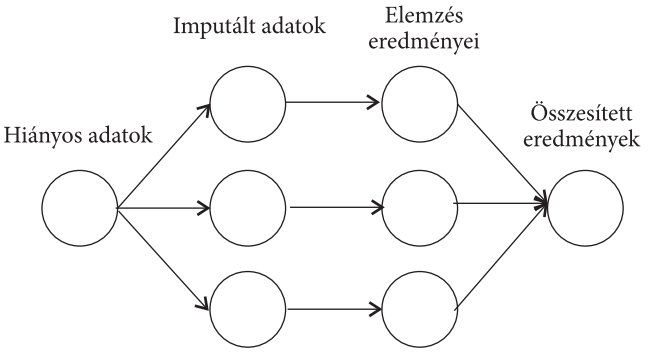
Azzal, hogy egy hiányos adatbázisban megpróbálunk nem létező adatokat becsülni, akaratunk ellenére mi magunk viszünk bizonytalanságot a becslésünkbe. A választásunk egy logisztikus regresszió alapú, többszörös imputációs eljárásra esett, mert segítségével becsülhető a becslőfüggvények varianciája, valamint beépíthető az adathiány okozta bizonytalanság a rendszerbe (*Oravecz*, 2008).

A többszörös imputációs (multiple imputation) modellek esetén a „többszörös" jelző arra utal, hogy minden hiányzó érték helyére több, *m* darab becslést készítünk, és a végén az *m*-féle teljes adatbázison elvégzett elemzések eredményeit összesítjük (pooling) a becsült paraméterek és a standard hibák segítségével (*Little–Rubin*, 2002). A pótlások bizonytalansága a metódus segítségével beépül a modellbe, így az imputált adatbázis közelíteni képes a teljes adatbázisban lévő változékonyságot.

A többszörös imputáció logikáját és az előbb bemutatott főbb lépéseket mutatja a következő ábra ($m = 3$):

3. ábra

A többszörös imputáció főbb lépései



Forrás: Buuren–Groothuis–Oudshoorn (2011)

A *3. ábrán* szereplő 3 iteráció a gyakorlatban azért még kevés, de már 10-20 a legtöbb esetben elegendő (*Buuren–Groothuis–Oudshoorn*, 2011).

Az általunk használt modell egy többszörös imputációs eljárás, amely speciálisan egyváltozós adathiány esetén alkalmazható, ahol a nem teljes változó dichotóm (értékei kizárólag 0 vagy 1 lehetnek).[3]

Jelölje θ egy ismeretlen paramétereket tartalmazó vektort, X_i pedig a hiánytalanul megfigyelt változók halmazát (ezek lesznek a magyarázó változók). Legyen Y_i a hiányos dichotóm változó, amelynek hiányzó értékeit becsülni szeretnénk. Ekkor az Y_i dummy változó feltételes eloszlása a következő:

$$f(Y_i|X_i,\theta) = logit^{-1}(X_i\theta)^{Y_i}[1 - logit^{-1}(X_i\theta)]^{1-Y_i}\,, \tag{3}$$

ahol az inverz logit függvény a következőt jelenti:

$$logit^{-1}(a) = \frac{e^a}{1+e^a}\,. \tag{4}$$

A θ paraméterek becslése maximum likelihood módszerrel történik a teljes körűen megfigyeltek (vagyis a befogadott kérelmezők) adatai alapján. Az egyes hiányzó értékek imputációját az eljárás három lépésben valósítja meg.

1) Először a megfigyeltek adatai alapján számított, becsült várható értékű és varianciájú normális eloszlásból húzunk annyi darab véletlen számot, ahány X_i magyarázó változónk van (a kapott vektort jelöljük θ_*-gal).

2) Ezen becsült paramétervektor segítségével minden hiányzó megfigyelésre kiszámoljuk a $logit^{-1}(X_i\theta_*)$ értékét, amely egy 0 és 1 közötti valós számot ad eredményül.

3) Végül pedig minden hiányzó érték esetén generálunk a (0,1) intervallumon egy egyenletes eloszlású véletlen számot (ezt jelöljük v_i-vel). Ezek után pedig nincs más dolgunk, mint hogy $Y_i = 0$ értéket írunk a hiányzó adat helyére, amennyiben $v_i > logit^{-1}(X_i\theta_*)$ és $Y_i = 1$ értéket egyébként.

Többszörös imputációról lévén szó, ezen három lépést ismételve, létrehozunk *m* darab független, teljes adatbázist (mindig új véletlen számokat generálva), amelynek eredményeit aztán a *3. ábrán* illusztrált módon összesítjük.

Az imént ismertetett eljárás egy viszonylag egyszerű, többszörös imputációs módszer, amely alkalmazható egyváltozós, MAR-típusú (azaz véletlenszerű) adathiány kezelésére, ahol a hiányosan megfigyelt változó dichotóm. A *4. fejezetben* egy valós adatbázison fogjuk bemutatni ezt az eljárást, és segítségével megpróbáljuk csökkenteni a szelekciós torzítást.

3 Az eljárást lásd részletesebben: Rubin (1987), p. 169.

3. ADÓSMINŐSÍTÉSI MODELLEK MODELLKOCKÁZATÁNAK MÉRÉSE

Az előző fejezetben bemutattuk a szelekciós torzítás jelenségét, valamint egy annak a csökkentésére irányuló technikát. Erre egyrészről azért volt szükség, mert ezen módszer segítségével javítható a hitelezési scoring modellek klasszifikáló képessége, másrészt a modellezési kockázat mérésénél azzal a feltételezéssel fogunk élni, hogy a logisztikus regressziós scoring modell eredményei az egyes ügyfelek csődvalószínűségeiként értelmezhetőek. Ez utóbbi feltevés csak akkor teljesül, ha a modellépítési minta reprezentálja a hiteligénylők teljes sokaságát, ami az elutasítottak adatainak beépítése nélkül nem valósulna meg.

3.1. Klasszifikáció és a scoring modellek lehetséges veszteségei

Ebben a fejezetrészben rátérünk az adósminősítési modellek modellkockázatának becslésére. A kockázati mértékek meghatározásához először szükségünk van a scoring modell által elkövetett hibák következtében fellépő lehetséges veszteségekre.

A klasszifikáció során kétféle hibát követhetünk el, amelynek különböző költsége van (*Thomas* et al., 2002). Egyrészt elutasíthatunk egy valójában jó hiteligénylőt (ez a másodfajú hiba), ami miatt az ügyfélen elérhető, potenciális profitot a bank elveszíti. A másik lehetséges hiba, ha a rendszer jónak titulál egy valójában rossz egyedet, és ezért a bank meghitelezi (elsőfajú hiba). Ez esetben tényleges vesztesége keletkezik a hitelintézetnek, ha az ügyfél bedől, és nem fizeti kötelezettségét.

A modell által előre jelzett, illetve a ténylegesen bekövetkező állapot (az adott jelentkező csődbe ment, vagy rendesen törlesztett) összevetésére elkészíthető egy 2 × 2-es mátrix, az úgynevezett klasszifikációs tábla (confusion matrix).

1. táblázat

Klasszifikációs tábla, benne az első- és másodfajú hibával

Klasszifikációs tábla		**Ténylegesen**		**Σ**
		Jó (G)	**Rossz (B)**	
Előrejelzés	**Jó (G)**	Helyesen jónak ítélt (g_G)	Tévesen jónak ítélt (g_B) **elsőfajú hiba**	g
	Rossz (B)	Tévesen rossznak ítélt (b_G) **másodfajú hiba**	Helyesen rossznak ítélt (b_B)	b
Σ		n_G	n_B	n

Forrás: Thomas et al. (2002)

A klasszifikációs tábla tehát úgy épül fel, hogy a soraiban találhatjuk adott *C* cutoff (a befogadás/elutasítás küszöbértéke) mellett a modellünk által jónak (*g*), illetve rossznak vélt kérelmezők számát (*b*). A hitelkockázatot jelző változó segítségével (ténylegesen milyen az ügyfél) pedig meg tudjuk nézni, hogy scoring modellünk hány egyedet sorolt be helyesen, illetve mennyiszer tévedett. Egy adott *n* elemű minta mellett a klasszifikációs tábla utolsó, összegző sorában szereplő értékek (n_G, n_B és *n*) fixek, míg a modell által befogadott (*g*), valamint elutasított kérelmezők száma (*b*) a cutoff megválasztásától függ.

A klasszifikációs tábla szerkezete az elkövetett hibák száma adott csődvalószínűségek mellett a cutoff megválasztásától függ. A következő két táblázat mutatja a szélsőséges eseteket:

2–3. táblázat

Klasszifikációs táblák 0 és 1 cutoff értékek mellett

Cutoff = 0		**Ténylegesen**		**Σ**
		Jó	**Rossz**	
Előrejelzés	**Jó**	0	**0**	**0**
	Rossz	$\mathbf{n_G}$	n_B	**n**
Σ		$\mathbf{n_G}$	$\mathbf{n_B}$	**n**

Cutoff = 1		**Ténylegesen**		**Σ**
		Jó	**Rossz**	
Előrejelzés	**Jó**	n_G	$\mathbf{n_B}$	**n**
	Rossz	**0**	0	**0**
Σ		$\mathbf{n_G}$	$\mathbf{n_B}$	**n**

Forrás: saját táblázatok

Az első esetben (bal oldali táblázat) a cutoff értékét a minimális 0-nak választottuk. Ekkor minden kérelmező elutasításra kerül, így az ilyen kiválasztás során csak másodfajú hiba léphet fel, mértéke pedig az adatbázisban szereplő, jó ügyfelek számával egyezik meg (hiszen ezen kérelmeket is elutasították).

A jobb oldali táblázat azt a szituációt mutatja, amikor az elutasítás küszöbértékét a maximális *C* = 1 értékben határozzuk meg. Ekkor minden kérelmezőt meghitelez a bank. Ez az eset tulajdonképpen a szakirodalomban gyakran emlegetett „nyitott kapuk" módszere, amely a korábban ismertetett, szelekciós torzítás elkerülésének legjobb módja, hiszen segítségével valóban a teljes sokaságot reprezentáló modellépítési mintánk lesz. Ha ennyire hatásos, akkor miért nem alkalmazzák a bankok a hitelezési scoring modelljeik építésénél?

A válasz igen egyszerű: mert „*a tapasztalat drága iskola*".[4] A bank hatalmas veszteséget szenvedne el egy ilyen hitelportfólión a rossz ügyfelek nemfizetése nyo-

4 *Benjamin Franklin* (idézi Jorion, 1999, p. 40.)

mán, ezért a pénzügyi intézmények inkább vállalják a szelekciós torzítás okozta, magasabb számú modellhibát, vagy keresnek kevésbé megbízható megoldást az alapadatok reprezentativitásának kérdésére.

A cutoff tehát egy olyan változó, amelynek a változtatása esetén adott csődvalószínűségek mellett a klasszifikációs tábla szerkezete változik. Tanulmányunkban a modellezési kockázatot szeretnénk számszerűsíteni, vagyis azt a veszélyt, hogy a modellhibák miatt veszteséget szenved el a bank. A modellhiba miatti veszteség pedig az elsőfajú és másodfajú hibákból ered, így a továbbiakban a 2 × 2-es klasszifikációs tábla (*1. táblázat*) jobb felső, illetve bal alsó negyedére kell fókuszálnunk.

Először valamilyen eljárás (például logisztikus regresszió) segítségével megbecsüljük minden ügyfélnél a $p(x)$ feltételes bedőlési valószínűségeket, amelyek valószínűségkénti értelmezése azon alapszik, hogy a modellépítési minta reprezentálja-e a hiteligénylők teljes, „ajtón bejövő" sokaságát. Különböző reject inference technikák segítségével csökkenthető a szelekciós torzítás, így az elutasított kérelmezők adatainak beépítésével reális feltételezés lehet, hogy a minta reprezentatív. Adottak tehát a csődvalószínűségek, amelyek a cutoffhoz hasonlóan a (0,1) intervallumon vehetik fel értékeiket.

Ha sorba rendezzük az egyedeket becsült csődvalószínűségeik alapján, akkor – amennyiben nincs két teljesen azonos megfigyelés – bármely két szomszédos $p(x)$ bedőlési valószínűség értéke közötti cutoff különböző klasszifikációs táblához vezet. Így a klasszifikációs tábla $(n + 1)$ különböző szerkezete állítható elő, amelyek adott bedőlési valószínűségek mellett a default változó megfigyelt értékeitől függnek. Ezen különböző klasszifikációstábla-összetételek segítségével pedig modellezhető a hitelezési scoring rendszer modellkockázata.

Bontsuk fel a (0,1) intervallumot az imént említett módon megfelelően sok részintervallumra, az egyes osztópontok mint cutoff értékek mentén pedig nézzük meg a teljes modellhibák okozta veszteséget!

Tegyük fel, hogy az elsőfajú hiba költsége D (debt), amely minden hiteligénylőre azonos mértékű. Például $D = 0{,}45$ azt jelenti, hogy egy befogadott rossz kérelmező csődje esetén várhatóan a bank nem kapja vissza az adott ügyféllel szembeni kitettségének 45%-át. Jelölje L (lost profit) a másodfajú hiba bekövetkezése esetén felmerült alternatívaköltséget, amelyet az elmaradt kamatbevétel miatt szenved el a bank. A modellkockázat becslése során végig fixnek feltételezzük a D és L modellhibából eredő veszteségek arányát.

Adott C elutasítási küszöbérték mellett a modellhibák okozta veszteséget a teljes hitelportfólióra a következőképpen számíthatjuk ki:

$$Teljes\ veszteség_C = D \sum_{i=1}^{g_B} E_i \cdot \widehat{p(x)}_i + L \sum_{i=1}^{b_G} E_i \cdot \widehat{p(x)}_i \,, \tag{5}$$

ahol E_i az i-edik egyed esetén fennálló kitettség (exposure) nagysága (fedezet hiányában ez a folyósított hitelösszeg); $\widehat{p(x)}i$ pedig az adott ügyfél becsült feltételes csődvalószínűsége.

Az (5) egyenletben az első szumma azon ügyfelekre vonatkozik, ahol elsőfajú hibát vétett a modell, míg a második szummában összegződnek az olyan egyedek által okozott veszteségek, akik esetén másodfajú hiba következett be.

Ebben a részben bemutattuk, hogy hogyan kaphatjuk meg a lehetséges klasszifikációs modell okozta, portfóliószintű veszteségértékeket. A következő fejezetrészben pedig ismertetjük, hogy a veszteségek eloszlásának felhasználásával hogyan határozhatóak meg a modellezési kockázat különböző mértékei.

3.2. A modellkockázat mérése az extrémérték-elmélet segítségével

A hitelezési scoring rendszerek modellkockázatának kockázati mértékére az empirikus kvantilisnél jobb mérőszámot adhat, ha a kockáztatott érték meghatározásához az extrémérték-elmélet eszköztárát felhasználva, egy megfelelő eloszlást illesztünk a veszteségek szélére. Erre azért van szükség, mert a veszteségeloszlás szélein jellemzően kevés megfigyelés található, amelyek között nagy távolságok is lehetnek, így pedig pontbecslésünk félrevezető lehet.

Az extrémérték-elmélet (extreme value theory – EVT) az extrém (kiugró) események statisztikai elemzésével foglalkozik. Az elmélet részterületei közül a pénzügyi alkalmazásokat tekintve talán a legelterjedtebb a küszöbtúllépések (treshold exceedances) modellje, amely az összes olyan veszteséget figyelembe veszi az eloszlás szélének becslésére, amelyek meghaladnak egy bizonyos u veszteségküszöböt (*Tulassay*, 2013). Segítségével jobb becslést tudunk adni a VaR-ra az eloszlás szélének teljes körű figyelembevételével.

Legyen X egy valószínűségi változó, amely a modellezendő veszteségeket képviseli, és $(Fx)=P(X\leq x)$a veszteségek eloszlásfüggvénye. Tekintsük extrém veszteségnek az egy adott u küszöböt meghaladó értékeket. A túllépések eloszlása (feltéve, hogy meghaladtuk az u határt) ekkor

$$F_u(y) = P(X - u \leq y | X > u)\,, \tag{6}$$

ahol $(X - u)$ nem más, mint a túllépés mértéke.

A Bayes-tétel segítségével összefüggés kereshető az $F(x)$ veszteségek eloszlásfüggvénye és az $F_u(y)$ küszöbtúllépések feltételes eloszlásának eloszlásfüggvénye között.

$$F_u(y) = P(X - u \leq y | X > u) = \frac{P(u < X \leq y + u)}{P(X > u)} = \frac{F(y + u) - F(u)}{1 - F(u)}\,. \tag{7}$$

A *Pickands*, *Balkema* és *de Haan* által bizonyított tétel azt mondja, hogy az eloszlások széles osztályára létezik olyan ξ és $\beta(u)$, hogy ha az u küszöb tart az eloszlás felső végpontjához, akkor a küszöbtúllépések feltételes eloszlásának eloszlásfüggvényére igaz, hogy

$$F_u(y) = P(X - u \leq y | X > u) \approx G_{\xi,\beta(u)}(y), \tag{8}$$

ahol $G_{\xi,\beta}(y)$ az általánosított Pareto-eloszlás (Generalized Pareto Distribution – GPD). A Pickands–Balkema–de Haan-tétel tehát kimondja, hogy elég magas küszöb esetén a túllépések közelítőleg GPD-eloszlást követnek, vagyis az általánosított Pareto-eloszlás a küszöbtúllépések természetes modellje (*McNeil* et al., 2005).

Az általánosított Pareto-eloszlás eloszlásfüggvénye a következő általános alakban írható fel:

$$G_{\xi,\beta}(y) = \begin{cases} 1 - \left(1 + \frac{\xi y}{\beta}\right)^{-\frac{1}{\xi}}, & \xi \neq 0 \\ 1 - e^{-\frac{y}{\beta}}, & \xi = 0 \end{cases}, \tag{9}$$

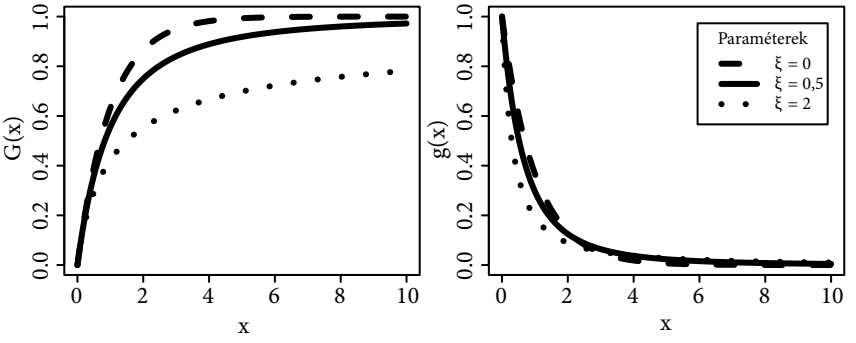
ahol ξ az úgynevezett alakparaméter, $\beta > 0$ pedig a skálaparaméter. A GPD-eloszlás várható értéke pedig

$$E(X) = \frac{\beta}{1 - \xi}. \tag{10}$$

A (9) felírásból jól látszik, hogy $\xi = 0$ esetben a GPD-eloszlás exponenciális eloszlást követ $\lambda = \frac{1}{\beta}$ paraméter mellett, vagyis az exponenciális eloszlás az általánosított Pareto-eloszlás egy speciális eseteként is felfogható. A GPD-eloszlás eloszlásfüggvényét (baloldalon) és sűrűségfüggvényét (jobb oldalon) mutatja a *4. ábra* három különböző ξ alakparaméter esetén ($\beta = 1$ mindegyik esetben):

4. ábra

A GPD-eloszlás eloszlás- és sűrűségfüggvénye különböző ξ értékek esetén



Forrás: saját ábra, R-ben elkészítve

Az ábrákon szaggatott vonal jelzi az exponenciális eloszlást ($\xi = 0$), folytonos vonal a 0,5-ös, míg pontozott görbe a 2-es alakparaméterrel rendelkező GPD-eloszlást. Amint az a *4. ábra* sűrűségfüggvényein ($g(x)$) látszik, az alakparaméter megválasztásával rugalmasan az eloszlás szélére szabható, a skálaparaméter segítségével pedig illeszthető pénzegységben értendő abszolút veszteségekre, de akár százalékban értelmezett hozamokra is az általánosított Pareto-eloszlás.

A veszteségeloszlás szélének modellje egy bizonyos u küszöb felett a (7) és (8) egyenletek felhasználásával és kihasználva, hogy $x = y + u$, a következő ($x > u$):

$$F(x) = [1 - F(u)]G_{\xi,\beta}(x - u) + F(u)\,, \tag{11}$$

ahol $F(u)$-t általában a historikus adatokból becsüljük: $\hat{F}(u) = \frac{n - N_u}{n}$, ezt a mód szert a szakirodalomban historikus szimulációnak hívják (a képletben szereplő N_u az u küszöböt meghaladó veszteségek számát jelöli, n pedig az összes vizsgált veszteség darabszámát) (*McNeil*, 1999).

Ekkor az $x > u$ veszteségeket a következőképpen modellezhetjük:

$$\hat{F}(x) = \left[1 - \hat{F}(u)\right]G_{\hat{\xi},\hat{\beta}}(x - u) + \hat{F}(u) = 1 - \frac{N_u}{n}\left[1 + \hat{\xi}\left(\frac{x - u}{\hat{\beta}}\right)\right]^{-\frac{1}{\hat{\xi}}}, \tag{12}$$

amely már egy pontosabb modell, mintha csak az empirikus eloszlást használnánk.

A küszöbátlépések modelljének leírásában gyakran szerepelt az a bizonyos, megfelelően választott u küszöb, amely feletti túllépéseket az elmélet modellezi. Ennek az u értéknek a meghatározása a gyakorlatban nem egyszerű feladat. Egyik lehetséges út a határ megválasztására, ha megvizsgáljuk az átlagos küszöbtúllépések függvényét (mean excess function). Egy X valószínűségi változó átlagos küszöbtúllépés-függvénye a következő (amennyiben X várható értéke véges):

$$e(u) = E(X - u | X > u)\,. \tag{13}$$

Mint azt említettük, a (7) egyenletben szereplő $F_u(y)$ az u küszöböt meghaladó küszöbtúllépések eloszlása, feltéve, hogy a veszteség átlépte az adott küszöböt. A (13) egyenletben szereplő, átlagos küszöbtúllépés-függvény pedig megadja $F_u(y)$ várható értékét az u függvényében. Az, hogy a küszöbtúllépések feltételes eloszlása GPD-eloszlást követ, vagyis $F_u(x) = G_{\xi,\beta(u)}(x)$, a $\beta(u) = \beta + \xi u$, esetén teljesül. Felhasználva a (10) egyenletben szereplő várhatóérték-képletet, az átlagos küszöbtúllépés-függvény átalakítható a következő formába:

$$e(u) = \frac{\beta(u)}{1 - \xi} = \frac{\beta + \xi u}{1 - \xi} = \frac{\beta}{1 - \xi} + \frac{\xi}{1 - \xi}u\,. \tag{14}$$

A (14) egyenletből látszik, hogy a GPD-eloszlás esetén az átlagos küszöbtúllépés-függvény u-ban lineáris, vagyis a megfelelő határ kiválasztásánál az a feladatunk, hogy ábrázolva $e(u)$-t, az u függvényében találjunk egy olyan küszöbértéket, amely felett a függvény közel lineáris, hiszen ekkor vélhetően jól fog illeszkedni az általánosított Pareto-eloszlás.

Az imént bemutatott küszöbtúllépések modellje a veszteségeloszlás egyszerű kvantilisénél jobban kihasználja az eloszlás széleiben rejlő információt, így felhasználható az eddigieknél pontosabb VaR-becslésre a következő módon:

$$\hat{F}(x) = 1 - \frac{N_u}{n}\left[1 + \hat{\xi}\left(\frac{x-u}{\hat{\beta}}\right)\right]^{-\frac{1}{\hat{\xi}}} = q\,, \tag{15}$$

ezt invertálva, megkapjuk a kockáztatott értéket (q>$F(u)$):

$$\widehat{VaR}_q = F^{-1}(q) = u + \frac{\hat{\beta}}{\hat{\xi}}\left[\left(\frac{n}{N_u}(1-q)\right)^{-\hat{\xi}} - 1\right]. \tag{16}$$

A VaR egyértelmű előnye az egyszerű értelmezhetősége, de számos hátrányos vonása van, például nem tekinthető koherens kockázati mértéknek. A legnagyobb probléma mégsem ez a VaR-ral kapcsolatban, hanem hogy nem mond semmit az azt meghaladó veszteségekről, vagyis az eloszlás legmagasabb veszteségeket tartalmazó széléről.

A modell segítségével a kockáztatott értéknél talán jobb, koherens kockázati mértékek is egyszerűen számíthatóak, mint amilyen például az expected shortfall. Az ES már felhasználja az eloszlás szélében rejlő információt, és azt mutatja meg, hogy mekkora a VaR-t meghaladó veszteségek (feltételes) várható értéke.

$$\widehat{ES}_q = \widehat{VaR}_q + E\left(X - \widehat{VaR}_q \middle| X > \widehat{VaR}_q\right) = \frac{\widehat{VaR}_q}{1-\hat{\xi}} + \frac{\hat{\beta} + \hat{\xi}u}{1-\hat{\xi}}\,. \tag{17}$$

A küszöbtúllépések modelljének segítségével jobban kihasználható az eloszlás széleiben rejlő információ, pontosabb és stabilabb eredményeket kapunk, mintha empirikus kvantilist számolnánk. Az elmélet szépsége mellett ugyanakkor hátrányként megemlítenénk, hogy a nagyon magas kvantilisek még így is csak nagy hibával becsülhetőek. A GPD-eloszlás paramétereinek becslését és a tanulmányunk témája szempontjából egyéb lényeges kérdéseket a gyakorlati példán fogjuk bemutatni.

4. A MODELLKOCKÁZAT BEMUTATÁSA EGY VALÓS ADATBÁZISON

Ebben a részben az a célunk, hogy egy valós, nyilvánosan elérhető adatbázison is bemutassuk az eddig áttekintett módszereket. Az adattábla rövid ismertetését követően először alkalmazunk a reject inference technikák közül egy többszörös imputáció alapú eljárást, amellyel célunk a szelekciós torzítás csökkentése. Ezek után az imputált adatbázison egy logisztikus regresszión alapuló scoring modellt építünk, amelynek modellezési kockázatát egy koherens (expected shortfall) és egy nem koherens (VaR) kockázati mérték segítségével mérni fogjuk.

4.1. A German Credit Data adatbázis bemutatása[5]

A hitelezési scoring adatbázisok, illetve az azokból készített hitelminősítő modellek a bankok leginkább védett adatai közé tartoznak. Ez az oka annak, hogy nagyon nehéz hozzájutni egy olyan teljes adattáblához, amelyen bemutathatóak lennének az általunk korábban ismertetett eljárások. Néhány kisebb adatbázis oktatási célra azért elérhető, egy ilyenen fogjuk elvégezni a szükséges elemzéseket az *R* statisztikai programcsomag segítségével.

A *German Credit Data* című adatbázist a Hamburgi Egyetem Statisztika és Ökonometria Tanszéke publikálta. Az adattábla 1000 lakossági hiteligénylő jellemzőit tartalmazza. A sorokban találhatók a megfigyelési egységek (ügyfelek), az oszlopokban az egyes változók, amelyek mentén elbírálhatjuk a hitelkérelmeket. Összesen 20 magyarázó változó, illetve egy hitelkockázatot jelző eredményváltozó (default) áll rendelkezésünkre annak eldöntésére, hogy egy adott kérelmező jó ügyfél lesz-e vagy rossz.

Miután nagy vonalakban megismerkedtünk az adatbázis lényeges jellemzőivel, következő fontos állomásként a korábban bemutatott módszer segítségével megpróbáljuk csökkenteni az alapadatokban jelen lévő szelekciós torzítást.

4.2. A szelekciós torzítás csökkentése

Ahogyan azt az elméleti bevezetőben is említettük, a szelekciós torzítás kezelésének szakirodalma számtalan módszert kínál a probléma orvoslására. Nincsen olyan általános érvényű eljárás, amely minden típusú adathiányra a legjobb megoldást adja. Az egyes szerzők a témában sok esetben ellentétes eredményre jutottak, mert a módszerek sikere nagyban függ az adott adatbázis karakterisztikájától.

5 A *German Credit Data* adatbázis nyilvános, elérhető az alábbi helyen: https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data), letöltve: 2014. 10. 09.

A valóságban a bankok ismerik az elutasított kérelmezők minden bekért adatát (természetesen a csődösséget jelző változó kivételével), így rendelkezésükre áll egy olyan minta, amely valóban az „ajtón bejövő", a sokaságot reprezentálni képes kérelmezők adatait tartalmazza. Ehhez hasonló, teljes adatbázis nyilvánosan nem érhető el, ezért feltételeztük, hogy az előző fejezetrészben bemutatott *German Credit Data* adattábla ilyen, vagyis valamennyi típusú igénylőt a sokasági aránynak megfelelően tartalmaz.

Ekkor az elutasított kérelmezők default változójának értékét nem ismerjük, így ezeket törölni kell az adatbázisból. Annak eldöntésére, hogy mely ügyfelek legyenek azok, akik nem kaptak hitelt, lefuttattunk egy logisztikus regressziót valamennyi megfigyelési egységre. A rendelkezésre álló 20 magyarázó változóból egy backward[6] típusú modellszelekciós eljárás segítségével csak az 5%-on szignifikáns változókat vontuk be a modellbe. Így végül a 11 magyarázó változós, szűkített modell alapján meghatároztunk minden egyedre egy pontszámot.

Tegyük fel, hogy a bank ezen becsült értékek alapján a legrosszabb 50 hiteligénylőt elutasította, a maradék 950 magánszemély pedig megkapta a kért összeget. Azért kellett ilyen magas, 95%-os befogadási arányt feltételeznünk, mert nagyobb fokú elutasítás mellett túl kevés rossz ügyfél maradt volna a mintában, amelyre a később bemutatásra kerülő, imputációs eljárás nem tudna megfelelően illeszkedni. Egy valós modellépítési adatbázis az általunk használt, 1000 elemű mintánál általában jóval több megfigyelési egységet tartalmaz, így ott a magasabb elutasítási arány sem okoz gondot.

Az elutasított 50 hiteligénylő default változóját ezután kitöröltük[7], mert ezt csak a befogadott ügyfelek esetében ismerhetjük. Amennyiben csak a meghitelezett 950 kérelmező adataival dolgoznánk a továbbiakban, akkor vélhetően torzított eredményeket kapnánk a korábban bemutatott, szelekciós torzítás jelenléte miatt. Be kell tehát építeni az elutasítottakat is az elemzésbe. Ezt például úgy tehetjük meg, ha a meghitelezettek teljes körűen megfigyelt adatai alapján megbecsüljük az elutasítottak hiányzó értékeit.

Az elutasítottak default változójának becslését egy logisztikus regresszión alapuló, többszörös imputációs eljárással végeztük el a „*mice*" R package segítségével. Amennyiben a hiányos adatbázis véletlenszerű (MAR) adathiánnyal rendelkezik, gyakran használt eljárás annak kezelésére a többszörös imputáció. Az általunk ismertetett eljárással megbecsültük tehát a default változó ismeretlen értékeit. Összehasonlítva az imputált értékeket a korábban törölt valósakkal, az 50 esetből

6 A backward típusú modellszelekciós eljárás lépésről lépésre szűkíti a modellt, egészen addig, amíg nem lesz valamennyi bevont változó szignifikáns (Kovács, 2011)

7 A törlés előtt az értékeket egy másik objektumba elmentettük, hogy a későbbiekben az alkalmazott imputációs eljárás hatékonyságát a segítségével vizsgálni tudjuk.

13-szor tévedett ez az eljárás. Ez azt jelenti, hogy az esetek 74%-ában helyesen becsülte a hiányzó megfigyeléseket.

Ezután az imputációt elvégeztük az eljárás bootstrapet felhasználó változatával is. Egyes szerzők szerint ugyanis a korábbiakban említett, normalitási feltevés a többszörös imputációs eljárások esetében általában sérül, ami azt eredményezheti, hogy torzított becsléseket kaphatunk a θ paraméterekre. Ez azonban *White* et al. (2010) kutatásai alapján elkerülhető bootstrapet alkalmazó eljárásokkal. Ennek során a megfigyelt adatokból mintákat veszünk, amelyeken újra és újra elvégezzük az imputációs eljárást, elmentve az egyes esetekben adódott paramétereket. Végül pedig ezen paraméterek eloszlásából húzva végezzük el a hiányzó adatok pótlását.

A bootstrapet alkalmazó, többszörös imputációs eljárás csupán 6 esetben tévedett az eredeti értékekhez képest, ami 88%-os hatékonyságot jelent. Mind a hat félreklasszifikált esetben a módszer másodfajú hibát vétett, vagyis valójában jó ügyfeleket sorolt be a rosszak közé. Mivel a gyakorlati tapasztalatok azt mutatják, hogy jóval nagyobb az elsőfajú hiba okozta veszteség, mint a másodfajú miatt bekövetkező, ezért az általunk alkalmazott eljárás amellett, hogy kevés esetben tévedett, mindezt a kisebb költséggel járó irányba tette. Ezek figyelembe vételével végül a bootstrap alapú második eljárás által kapott becsült eredményekkel egészítettük ki a hiányos adatmátrixot.

4.3. A lehetséges modellkockázati veszteségek meghatározása

Az immár hiányzó értékektől mentes adattáblán az előző fejezetrészben leírtakhoz hasonlóan lefuttattunk egy logisztikus regressziót először minden magyarázó változó bevonásával, majd létrehoztunk egy szűkebb modellt, csak az 5%-on szignifikánsak felhasználásával.

A logit becsült paramétereinek segítségével minden ügyfélre meghatároztuk annak bedőlési valószínűségét (probability of default – PD). Az eredmények valószínűségként való értelmezése egy erős feltevés, de amennyiben az adatbázisban szereplő 1000 kérelmező valóban reprezentálja a teljes sokaságot, és a szelekciós torzítás kezelésére irányuló törekvés sikeres volt, akkor ez a feltételezés megfelelő lehet.

A modellkockázat számszerűsítéséhez szükségünk van egy becsült nemteljesítéskori veszteségrátára (loss given default – LGD), amely azt mutatja meg, hogy egy befogadott ügyfél csődje esetén várhatóan a kitettség mekkora része nem térül meg. Ennek becsléséhez támpontot nyújthat a bázeli szabályozás, amely a hitelkockázati alap IRB (internal rating based – belső minősítésen alapuló) módszer keretében 45%-os LGD-értéket ír elő nem alárendelt, elismert biztosíték nélküli hitelekre (575/2013/EU-rendelet, 161. cikk (1)).

Ez az érték azonban a nem lakossági kitettségekre vonatkozik, a magánszemélyeknek nyújtott kölcsönök esetében a nemteljesítéskori veszteségrátát a hitelintézeteknek és befektetési vállalkozásoknak saját maguknak kell becsülniük. Az elemzéshez kiindulópontnak ezt a szabályozói értéket fogjuk használni, de ezen input paraméternek bármelyik intézmény beírhatja az általa becsült LGD-t.

Folytatva a lehetséges modellkockázati veszteségek meghatározásának gondolatmenetét, legyen az elsőfajú hiba költsége minden egyedre 45% (D = 0,45). A másodfajú hiba költsége egy jó ügyfél elutasítása esetén felmerülő alternatívaköltséget jelenti, vagyis az elmulasztott kamatjövedelmeket. A *German Credit Data* adatbázis leírásában megtalálható, hogy *Hans Hofmannak*, a Hamburgi Egyetem professzorának becslése szerint hozzávetőleg ötször akkora az elsőfajú hiba költsége, mint a másodfajú hibáé (*Hofmann*, 1994). Ezt felhasználva, legyen a másodfajú hiba során elszenvedett veszteség a kitettség 9%-a (L = 0,45/5 = 0,09). Az elemzés végén el fogjuk végezni ezen becsült veszteségek érzékenységvizsgálatát, vagyis megnézzük azok végső kockázati mértékre kifejtett hatását.

A modellhibák okozta lehetséges veszteségeket a *3.1 alfejezetben* leírt módon határoztuk meg úgy, hogy először létrehoztuk azon cutoff értékek vektorát, amelyek mentén a különböző klasszifikációstábla-szerkezetek esetén előforduló veszteségeket vizsgálni fogjuk. Ezután a különböző elutasítási küszöbértékek esetén megnéztük, hogy mely egyedeket sorolt be a modell tévesen jónak, illetve helytelen módon rossznak.

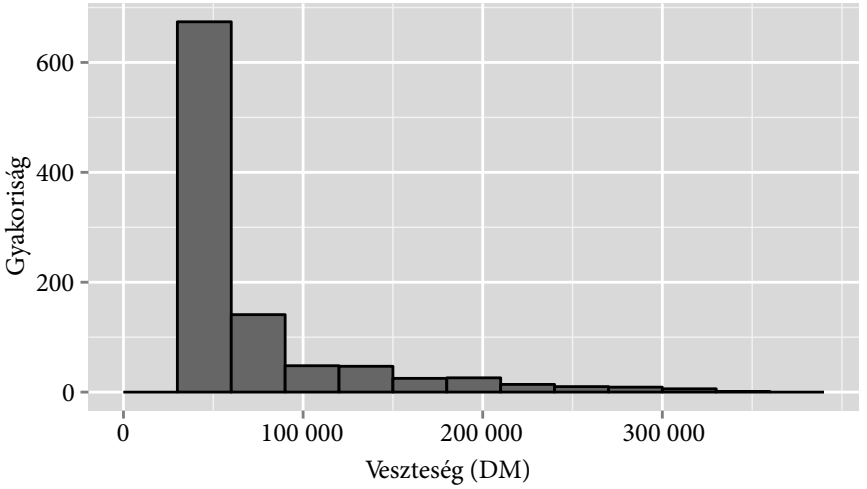
Minden kérelmezőre, ahol a modell hibát ejtett, kiszámoltuk a várható veszteséget. Ezt úgy végeztük el, hogy vettük a kitettség (vagyis a hitelösszeg) értékének 45%-át, illetve 9%-át attól függően, hogy első- vagy másodfajú hibát követett el a modell, majd ezt megszoroztuk az adott egyedre becsült csődvalószínűséggel. Itt azzal a feltételezéssel élünk, hogy ezek a kockázati tényezők (PD, LGD, E) egymástól függetlenek, valamint fedezettel nem rendelkező hitelkihelyezésekről van szó, vagyis a kitettség mértéke megegyezik a hitelösszeggel.

Adott cutoff mellett a portfóliószintű modellkockázati veszteséget ezen egyedi várható költségek összege adja. A korábban leírt módon meghatározott, különböző befogadási küszöbértékek mellett kiszámítva mindezt, az eltérő összetételű klasszifikációs táblák melletti, modellhibák okozta veszteségeket kaptuk.

4.4. A modellkockázat számszerűsítése

Az előzőekben kiszámított, téves besorolásokból eredő, lehetséges portfóliószintű veszteségeket mutatja a következő hisztogram:

5. ábra
A portfóliószintű veszteségek hisztogramja



Forrás: saját ábra

Az *5. ábrán* látható veszteségeloszlás erős aszimmetriával rendelkezik, a jobb széle hosszan elnyúlik. Ez azt mutatja, hogy az alacsony veszteségek gyakoriak, míg az igazán nagy modellhibák okozta veszteségből kevés van.

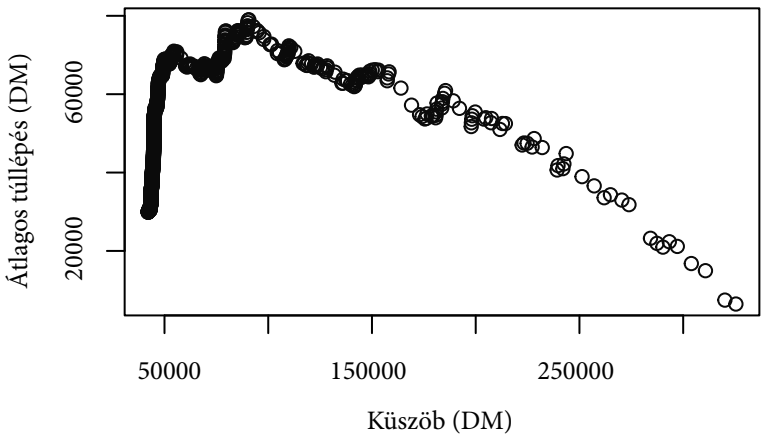
A következőkben az *5. ábrán* látható veszteségeloszlás szélére fogjuk alkalmazni a küszöbtúllépések modelljének eszköztárát. Az R-ben való kalkulációkat, illetve ábrákat az „*evir*" package segítségével készítettük el. Mint azt korábban említettük, az extrémérték-elméletnek számos alkalmazási területe van, a szóban forgó R csomag kifejezetten a pénzügyi alkalmazásra fókuszálva számít különböző kockázati mértékeket a veszteségeloszlásra. Az illesztett GPD-eloszlás paramétereinek becslését maximum likelihood módszerrel végzi a program (*Gilleland* et al., 2013).

Szükségünk van először egy megfelelően megválasztott *u* küszöbre amelyet meghaladó veszteségeknél a küszöbtúllépések feltételes eloszlása közelítőleg GPD-eloszlást követ. A határ megtalálásának egyik leggyakrabban használt eszköze az átlagos küszöbtúllépések függvénye, amely megmutatja, hogy különböző *u* értékek esetén (vízszintes tengely) mekkora az azt meghaladó veszteségek átlaga.

A GPD-eloszlás esetén ezen várható érték az *u* küszöb lineáris függvénye, vagyis feladatunk az, hogy meghatározzunk egy olyan határt, amely felett az átlagos küszöbtúllépések függvénye közel lineáris.

6. ábra

Átlagos küszöbtúllépések függvénye (mean excess function)



Forrás: saját ábra

A *6. ábrán* láthatjuk, hogy a küszöbválasztás problémája nem feltétlenül egyértelmű, több megoldás is lehetséges. Hozzávetőlegesen a 60 000 német márkás határ alatt a függvény határozottan pozitív meredekségű, míg ezen érték felett egy negatív trend látszik. Ezek alapján u = 60 0000 küszöb mellett illesztettünk GPD-eloszlást a veszteségekre. Mivel több lehetséges határ is megfelelő lenne, az elemzés végén be fogjuk mutatni a kockázati mértékek érzékenységét a küszöbválasztásra. Az illesztett GPD-eloszlás becsült ξ alak- és β skálaparamétere, valamint a standard hibák a következők lettek:

4. táblázat

Az illesztett GPD-eloszlás becsült paraméterei és a standard hibák

Megnevezés	ξ	β
Becsült paraméter	–0,0822	72 744
Standard hiba	0,0532	3 964

Forrás: saját táblázat

A megfigyelésekre illesztett GPD-modell lehetőséget nyújt a veszteségeloszlások felső kvantiliseinek becslésére, és ezáltal különböző kockázati mértékek kiszámítására.

5. táblázat

VaR-értékek, historikus percentilisek, valamint expected shortfall

Szignifikanciaszint	VaR	Historikus percentilis	ES
95%	5,70%	6,05%	7,46%
99%	8,57%	8,87%	10,12%

Forrás: saját táblázat

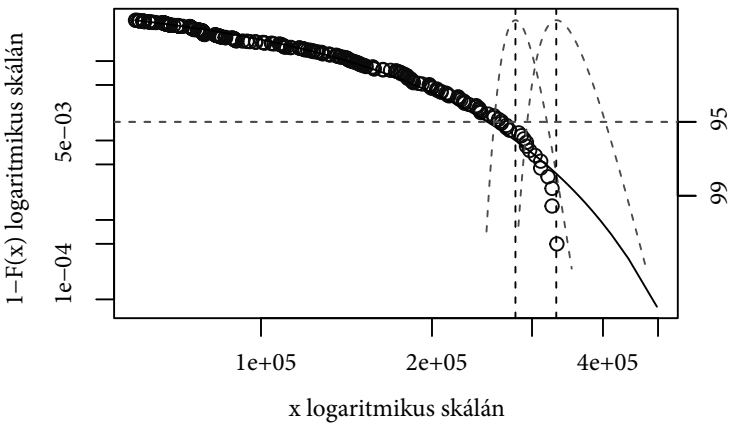
Az 5. *táblázatban* a kockázati mértékek a teljes hitelportfólió értékének százalékában vannak kifejezve. Látható, hogy a veszteségeloszlás historikus percentilise minden esetben viszonylag közel esik az illesztett GPD-eloszláson alapuló VaR-becsléshez. A két érték közül ez utóbbit tekinthetjük a jobb becslésnek, mert a korábban leírtaknak megfelelően a küszöbtúllépések modellje jobban kihasználja a veszteségeloszlás szélében rejlő információt.

A táblázat alsó sorában található VaR-érték például úgy értelmezhető, hogy 99%-os megbízhatósággal az adósminősítési modell hibáiból adódóan elszenvedhető maximális veszteség a portfólió értékének 8,57%-a a következő periódusban. Az expected shortfall a kockáztatott értéket meghaladó veszteségek (feltételes) várható értékét mutatja, ezért az minden esetben magasabb lesz a VaR-nál. Ennek megfelelően a táblázatban szereplő 10,12%-os érték azt jelenti, hogy 99%-os megbízhatósági szinten a lehetséges modellkockázati veszteségek legrosszabb 1%-os tartományában (a 280 506 német márka feletti veszteségek esetén) várhatóan az intézményt 10,12%-os veszteség érheti.

Az imént értelmezett kockázati mértékeket, azok konfidenciaintervallumait, valamint az illesztett általánosított Pareto-eloszlást szemlélteti az alábbi ábra:

7. ábra

A 99%-os VaR- és ES-értékek, valamint azok konfidenciaintervallumai



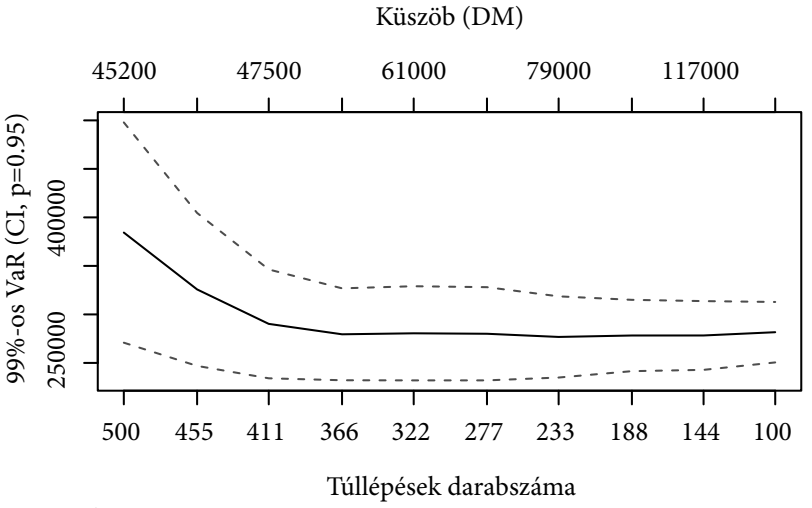
Forrás: saját ábra az *„evir"* package segítségével elkészítve

A *7. ábrán* karikák jelzik az egyes megfigyelt veszteségeket (*x*), a folytonos vonal pedig az illesztett GDP-eloszlást. Látszik, hogy néhány outliertől eltekintve, elég jól illeszkednek az adatpontok az általánosított Pareto-eloszlásra. A bal oldali, függőleges szaggatott vonal a 99%-os VaR-értéket, a tőle jobbra eső, függőleges egyenes pedig az expected shortfallt jelöli. A két konkáv görbe a kockázati mértékek konfidenciaintervallumait mutatja. Az ábrán látható, vízszintes vonal és ezen két görbe metszéspontjai megadják a becsült kockázati mérőszámok 95%-os (jobb oldali tengely) konfidenciaintervallumának végpontjait. Ezen egyenest párhuzamosan lefele tolva a 99%-os érték felé, egyre nagyobb megbízhatósági szintű intervallumokat kapunk. Természetesen – ahogyan az az ábrán is látszik – magasabb megbízhatósági szintű becslés egyre szélesebb konfidenciasávot eredményez.

Korábban említettük, hogy az átlagos küszöbtúllépések függvénye (*6. ábra*) alapján nem mindig egyértelmű az *u* küszöb megválasztása.

8. ábra

A 99%-os VaR értéke az *u* küszöb függvényében



Forrás: saját ábra

A *8. ábra* azt mutatja, hogy az általunk választott 60 000 német márkás határ (felső tengely) fölött szinte bárhol megválaszthattuk volna az *u* értékét, a 99%-os VaR becslése nem változna szignifikánsan. Ez azt jelenti, hogy a küszöbtúllépések modellje a kockázati mértékek robusztus becslését adja (*McNeil* et al., 2005).

A különféle feltételezések mellett a bemutatott eljárás függ bizonyos bemenő paraméterek értékétől. Ilyen például az első-, valamint a másodfajú hiba költsége. A korábbiakban az elsőfajú hibát $D = 45\%$-ként határoztuk meg a bázeli szabályozás

előírt LGD-je alapján, valamint egy szakértői becslés segítségével a hibamértékek arányából számítottuk ki a másodfajú hiba költségét (L = 9%). A következő táblázat mutatja a 99%-os megbízhatósági szintű VaR ezen paraméterekre vonatkozó érzékenységét:

6. táblázat

Az első- (D) és másodfajú (L) hiba költségének hatása a 99%-os VaR-ra

Érzékenység-vizsgálat		D		
		–1%	**0%**	**+1%**
L	**–1%**	–0,91%	**–0,64%**	0,38%
	0%	**–0,73%**	0%	**0,93%**
	+1%	–0,70%	**0,03%**	1,02%

Forrás: saját táblázat

A *6. táblázat* oszlopaiban láthatjuk, hogy hány százalékkal változik a 99%-os VaR értéke, amennyiben a D = 45%-ról kiindulva 1%-kal növeljük, illetve csökkentjük az elsőfajú hiba költségét. A sorok pedig azt mutatják meg, hogy hány százalékkal változik ugyanezen kockázati mérték, ha a másodfajú hiba költségét változtatjuk az L = 9%-os szinthez képest.

Jól látszik, hogy a 99%-os kockáztatott érték jóval érzékenyebb a D bemenő paraméterre. Amennyiben ceteris paribus a D-t növeljük 1%-kal, akkor a VaR értéke 0,93%-kal nő, míg ugyanez a változás L növelésének hatására csupán 0,03%. Az is jól látható, hogy a változások mértéke nem szimmetrikus, például együttesen emelve a kétféle hiba költségét, a VaR 1,02%-kal nő, míg szimultán csökkentve azokat, csupán 0,91%-kal csökken a kockáztatott érték. Ezen megállapításokat mindenféleképpen szem előtt kell tartania a hitelintézetnek, amennyiben a D és L bemenő paramétereket megbecsli a modellkockázat méréséhez, mert a VaR értéke érzékeny ezen input adatokra.

Elemzésünk zárásaként még mindenképpen érdemes megvizsgálni, hogy az általunk számszerűsített kockázati mértékek képesek-e pontosan mérni a modellezési kockázatot. A válasz természetesen nem, hiszen a kockázat egy látens, vagyis közvetlenül nem mérhető fogalom. És mint ilyen, annak *„bármely kvantifikálható mértéke csupán közelítő érték, kiemelve egy tényezőjét ama valós kockázatnak, amely elméletileg komplex, többjelentésű és soktényezős"* (*Bélyácz*, 2011, p. 309.)

5. ÖSSZEGZÉS

Tanulmányunk céljaként az adósminősítési modellek modellezési kockázatának mérését tűztük ki, hogy az így kapott eredmények segítségével a hitelintézetek jobban megismerhessék az általuk használt modellek hibáiból adódó, esetleges veszteségeket, valamint az így nyert információ támogathassa a vezetőket a döntések meghozatalában.

A modellezési kockázat definiálását követően egy rövid kitekintéssel bemutattuk annak helyét a bázeli szabályozásban. Ezt követően a problémakör egy vékony, de annál jelentősebb szeletét állítottuk a vizsgálódás középpontjába, a hitelezési scoring modellek modellezési kockázatát.

Ezután az alapadatok reprezentativitásának kérdésével foglalkoztunk. A szelekciós torzításként ismert probléma azért lép fel, mert a bank csak azon kérelmezők estén rendelkezik hiánytalanul megfigyelt adatokkal, akik már kaptak hitelt. Azon ügyfelek esetén, akiket a hitelintézet elutasított, nem ismert a hitelkockázatot jelző változó értéke, vagyis hogy mi történt volna abban az esetben, ha megkapják a kért kölcsönt. Itt tehát a modellkockázat kérdése felfogható hiányzó adat problémájaként.

Ezt szem előtt tartva, áttekintettük az adathiány főbb típusait, majd kiemeltünk az eljárások közül egy többszörös imputáció alapú metódust. Azért esett a választásunk erre, mert segítségével beépíthető az adathiány okozta bizonytalanság a becslésbe, amely kockázatkezelési szempontból kiemelt jelentőségű.

Ezt követően bemutattuk, hogy adott becsült bedőlési valószínűségek, valamint csődösséget jelző változó (default) mellett hogyan használható fel a cutoff különböző szerkezetű klasszifikációs táblák modellezésére, és így a lehetséges modellkockázati veszteségek előállítására. Ha pedig a veszteségeloszlás szélére alkalmazzuk a küszöbtúllépések modelljét, már könnyedén meghatározhatóak a modellezési kockázat különböző mértékei.

A szükséges elméleti alapok bemutatása után áttértünk a leírtak gyakorlati példán történő alkalmazására. Az adattábla rövid bemutatását követően egyváltozós és véletlenszerű (MAR) adathiány feltételezése mellett alkalmaztunk egy többszörös imputációs eljárást az elutasított kérelmezők hiányzó default változójának becslésére. Itt azt tapasztaltuk, hogy ezen eljárásnak a bootstrapet felhasználó változata pontosabb becsléseket eredményez.

Ezután megbecsültük az egyes ügyfelek csődvalószínűségeit, amelyek alapján előállítottuk a lehetséges modellkockázati veszteségeket. Ezen veszteségeloszlás szélére általánosított Pareto-eloszlás illesztése mellett meghatároztunk két kockázati mértéket, a VaR-t és az expected shortfallt.

Ezt követően megvizsgáltuk a 99%-os kockáztatott érték érzékenységét először az extrémérték-elmélet alkalmazása során használt küszöbre, majd pedig az első- és másodfajú hiba költségére. Itt arra a következtetésre jutottunk, hogy a VaR robusztus a küszöb megválasztására, viszont az egyes modelltévedések esetén elszenvedett veszteségek értékei komolyan befolyásolják a kockázati mértéket, ezek minél pontosabb becslése tehát létfontosságú a végeredmény szempontjából.

Ahogyan azt tanulmányunk címe is kiemeli, tökéletes modell definíciószerűen nem létezhet, vagyis mindenképpen szükség van a modellhibákból adódó, lehetséges veszélyek minél sokoldalúbb feltérképezésére. Az általunk használt eljárás is csupán a végtelenül komplex modellezési kockázat egy leegyszerűsített metszetét mutatja, így – visszautalva *George E. P. Box* szavaira – bár ez sem adhat hibátlan eredményt, de azért reményeink szerint hasznos lehet a modellkockázat megismeréséhez, az ehhez kapcsolódó banki döntések meghozatalához.

IRODALOMJEGYZÉK

575/2013/EU-rendelet a hitelintézetekre és befektetési vállalkozásokra vonatkozó prudenciális követelményekről és a 648/2012/EU rendelet módosításáról (2013. június 26.).

Bélyácz Iván (2011): Kockázat, bizonytalanság, valószínűség. *Hitelintézeti Szemle* 10 (4), pp. 289–313.

Box, G. E. P. – Draper, N. R. (2007): *Response Surfaces, Mixtures, and Ridge Analyses*. 2nd ed., New Jersey: John Wiley & Sons.

Buuren, S. – Groothuis-Oudshoorn, K. (2011): Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software* 45 (3), pp. 1–67.

Gilleland, E. – Ribatet, M. – Stephenson, A. G. (2013): A software review for extreme value analysis. *Extremes* 16, pp. 103–119.

Hofmann, H. (1994): German Credit Dataset. https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data) (letöltve: 2014. október 9.)

Jorion, P. (1999): *A kockáztatott érték*. Budapest: Panem Kiadó.

Kiss Ferenc (2003): A credit scoring fejlődése és alkalmazása. PhD-értekezés, Budapesti Műszaki Egyetem.

Kovács Erzsébet (2011): *Pénzügyi adatok statisztikai elemzése.* Budapest: Tanszék Kiadó.

Krekó Béla (2011): Kockázat, bizonytalanság és modellkockázat kockázatkezelési szemmel. *Hitelintézeti Szemle* 10 (4), pp. 370–378.

Little, R. J. A. – Rubin D. B. (2002): *Statistical Analysis with Missing Data*. 2nd ed., New Jersey: John Wiley & Sons.

MNB (2012a): A felügyeleti felülvizsgálati folyamat (SRP). Módszertani útmutató, Magyar Nemzeti Bank, május.

MNB (2012b): A tőkemegfelelés belső értékelési folyamata (ICAAP). Magyar Nemzeti Bank, május.

McNeil, A. J. (1999): *Extreme Value Theory for Risk Managers. Internal Modelling and CAD II*. London: Risk Books, pp. 93–113.

McNeil, A. J. – Frey, R. – Embrechts, P. (2005): *Quantitative Risk Management. Concepts, Techniques and Tools*. New Jersey: Princeton University Press.

Oravecz Beatrix (2008): Szelekciós torzítás és csökkentése az adósminősítési modelleknél. PhD-értekezés, Budapesti Corvinus Egyetem, Gazdálkodástudományi Doktori Iskola.

Radnai Márton – Vonnák Dzsamila (2010): *Banki tőkemegfelelési kézikönyv*. Budapest: Alinea Kiadó.

Rubin D. B. (1987): *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons,.

Thomas, L. C. – Edelman, D. B. – Crook, J. N. (2002): *Credit Scoring and Its Applications.* Philadelphia: Society for Industrial and Applied Mathematics.

Tulassay Zsolt (2013): Extreme Value Theory. In Daróczi, G. – Puhle, M. – Berlinger, E. – Csóka, P. – Havran, D. – Michaletzky, M. – Tulassay, Zs. – Váradi, K. – Vidovics-Dancs, Á.: Introduction to R for Quantitative Finance. Birmingham: Packt Publishing.

White, I. R. – Daniel, R. – Royston, P. (2010): Avoiding bias due to perfect prediction in multiple imputation of incomplete categorical variables. *Computational Statistics and Data Analysis* 54, pp. 2267–2275.