Validációs eljárások a csődelőrejelző modellek teljesítményének megítélésében

Nyitrai Tamás,
a Budapesti Corvinus Egyetem PhD-hallgatója
E-mail: nyitrai.tamas@hotmail.com

A szerző célja, hogy bemutassa és összevesse a szakirodalomban leggyakrabban alkalmazott validációs eljárásokat a csődelőrejelzés területén. Fő kutatási kérdése annak vizsgálata, hogy befolyásolják-e a csődmodellek előrejelző képességét mérni hivatott találati arány nagyságát a különböző validációs módszerek. A kérdésre ezer hazai vállalkozás adataival végzett empirikus vizsgálat eredményei alapján ad választ. Tanulmányában a C4.5 eljárást alkalmazza, ami a hazai szakirodalomban kevésbé elterjedt módszer, ezért röviden kitér elméleti hátterének ismertetésére is. Bemutat egy formulát, amely lehetőséget nyújt a statikus pénzügyi mutatók időbeli változásának figyelembevételére a csődmodellekben.

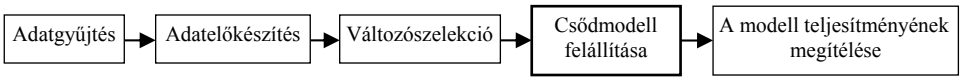
TÁRGYSZÓ:
Csődelőrejelzés.
Döntési fák.
Validációs eljárások.

A vállalatok jövőbeli fizetőképességének előrejelzése évtizedek óta a tudományos érdeklődés tárgyát képezi. Emiatt ez a kutatási terület örökzöldnek tekinthető, kimondottan igaz ez abban a napjainkat jellemző recessziós gazdasági környezetben, amelyben a finanszírozók jóval kockázatkerülőbbek, mint konjunktúra idején. Ezen érdekcsoport részéről tehát minden eddiginél fokozottabb az igény a minél megbízhatóbb teljesítményt nyújtó előrejelző modellek iránt. Fontos azonban hangsúlyozni, hogy nem csak a finanszírozók számára lehet lényeges egy vállalkozás jövőbeli fennmaradási esélyének lehető legpontosabb előrejelzése. A gazdaság egésze szempontjából sem közömbös, hogy hány vállalat működik egy-egy adott iparágban. Emiatt a szabályozó hatóságok, illetve a vállalati tevékenység más érdekcsoportjai részéről éppúgy felmerül az igény, mint a finanszírozókéról a minél jobb és stabilabb teljesítményt nyújtó előrejelző modellek iránt.

A vállalatok jövőbeli fizetőképességének előrebecslését csődelőrejelzés néven ismeri a hazai és a nemzetközi szakirodalom. E terület általános gyakorlata, hogy a vállalatok által közzétett számviteli dokumentumokban (mérleg, eredménykimutatás) szereplő adatokból kalkulálható pénzügyi mutatószámok információtartalmát felhasználva próbálja a gazdálkodóegység esetleges jövőbeli fizetésképtelenségét előre jelezni, valamely klasszifikációra alkalmas többváltozós matematikai-statisztikai módszer, illetve az elmúlt évtizedben dominánsan gépi tanulás (machine learning) eljárásainak segítségével. *Yu et al*. [2014] szerint napjainkban a csődelőrejelzés fő vonalát kizárólag az utóbbi módszerek alkalmazása jellemzi. E főáramba sorolható nemzetközi publikációk jellemzően módszertani összehasonlító kutatások eredményeit mutatják be, melyek célja a csődelőrejelzésben legjobb teljesítményt nyújtó klasszifikációs eljárások elméletének és gyakorlatának bemutatása.

A *Statisztikai Szemle* Olvasói e témáról legutóbb *Kristóf* [2005] tanulmányában olvashattak: a szerző a csődelőrejelzésben leggyakoribb sokváltozós statisztikai módszereket, valamint azok alkalmazását mutatta be az első hazai csődmodell adatbázisán végzett empirikus vizsgálat keretében. Az idézett mű megjelenése óta eltelt közel egy évtized alatt számos jelentős fejlemény történt e szakterületen. Ezek közül az egyik legfontosabb, hogy a korábban bemutatott fő irányzat mellett megjelentek alternatív kutatási területek is, amelyek a csődmodellek találati arányát a már meglevő klasszifikációs módszerek keretein belül próbálják meg növelni. Ezek az alternatív kutatási irányok a csődmodellek építésének egy-egy részterületéhez kötődően végeznek vizsgálatokat, törekedve a modellek előrejelző képességének növelésére. A csődmodellépítés főbb lépéseinek leegyszerűsített vázlatát mutatja az ábra.

A csődmodellépítés folyamata



A tudományos vizsgálódás főáramát az ábrán megvastagított, „Csődmodell felállítása" téglalap jelöli. Az ide sorolható legfontosabb kutatási területek: a klasszifikációs módszer kiválasztása, a paramétereinek optimalizálása, a módszerek esetleges kombinációja stb.

Azonban a többi részterülethez is számos kutatási kérdés kapcsolódik, ezek képezik napjainkban a csődelőrejelzés alternatív kutatási irányait. Jelen tanulmány az adatelőkészítés feladataival, valamint a modellek teljesítményének értékelésével foglalkozik részletesebben. Az adatelőkészítés körében azt vizsgálja, hogy növekszik-e a csődmodellek találati aránya, ha a pénzügyi mutatók nyers értékei mellett azt is szerepeltetjük a lehetséges input változók között, hogy miként viszonyul az adott vállalkozás pénzügyi mutatóinak legutoljára megfigyelt értéke az azt megelőző időszak megfelelő adataihoz.

A cikk másik vizsgálati kérdése az utolsó részterülethez, azaz a csődmodellek teljesítményének értékeléséhez kötődik. E modelleket jellemzően előrejelzési céllal készítik, így fontos szempont azok predikciós erejének értékelése. Ennek felmérését a modellek validációjaként ismeri a szakirodalom. A validáció során az előrejelző képességet olyan megfigyeléseken vizsgálják, amelyek nem szerepeltek a modellépítéshez felhasznált ún. tanulási mintában. A validáció szükségessége tekintetében konszenzus van a csődelőrejelzés szakirodalmában, annak módjában azonban jelentős eltéréseket tapasztalhatunk az egyes publikációk között. Napjainkban a kutatások többségénél az ún. keresztvalidációs eljárást alkalmazzák, de gyakran találkozhatunk azzal a megközelítéssel is, hogy a rendelkezésre álló mintát többször véletlenszerűen osztják fel tanulási és tesztelő mintákra, és a csődmodell teljesítményét az ezen felosztásokon kapott modellek találati arányának átlagában ítélik meg. Bármelyik eljárásra is essen a választás, a szakirodalomban jellemzően nem indokolják meg, hogy miért épp az egyik vagy másik módszert használják. Jelen tanulmány a leggyakrabban alkalmazott validációs eljárások – az egyet kihagyó eljárás (leave one out), a keresztvalidáció, a többszörös tanuló-tesztelő felosztás – összehasonlítására vállalkozik, annak érdekében, hogy megvizsgálja kimutatható-e valamilyen sorrend a különböző módszerekkel kapott validációs eredmények között. Azaz a kutatás másik fő célja annak kiderítése, hogy a validációs technikák között található-e olyan, amelyik jellemzően optimistább vagy pesszimistább becslést ad a modellek előrejelző teljesítményére. Fontos hangsúlyozni, hogy a tanulmány a három validációs eljárást csak a C4.5 alkalmazása mellett vizsgálja, így a levonható következtetések csak e módszer vonatkozásában tekinthetők érvényesnek.

A cikk első fejezete a csődelőrejelzés ábrán bemutatott részterületeihez kötődő legfontosabb hazai és nemzetközik publikációkat, a második a modellépítéshez felhasznált C4.5 klasszifikációs eljárás módszertani hátterét, a harmadik az empirikus vizsgálathoz felhasznált adatbázist, a negyedik a kutatás eredményeit ismerteti. Az ötödik, záró fejezetben az elemzésből levonató következtetéseket foglalom össze.

1. Szakirodalmi áttekintés

Beaver [1966] publikációjának megjelenése óta folyamatosnak mondható a tudományos érdeklődés a vállalatok jövőbeli fizetőképességének előrejelzése iránt. Tekintettel arra, hogy a csődelőrejelzés statisztikai szempontból klasszifikációs probléma, a tudományterület alapvetően módszertan-orientált. Emiatt fejlődését leginkább a klasszifikációs problémák megoldására alkalmas eljárások, valamint az azt támogató informatika fejlődése befolyásolja. Vélhetően pontosan emiatt az 1990-es évektől kezdődően indult rohamos növekedésnek a csődelőrejelzés témakörében megjelent tudományos publikációk száma.

A tudományterület fejlődésének első húsz évében az ábrán bemutatott részterületek között nem volt jellemző olyan mértékű arányeltolódás, mint az elmúlt két évtizedben, amikor a kutatás főáramát – a dinamikusan fejlődő módszertannak köszönhetően – az egyes klasszifikációs eljárások teljesítményének összehasonlító elemzése képezte. A részterület súlyát indokolja, hogy *Du Jardin* [2010] összesítése szerint az elmúlt ötven évben több mint ötszázféle klasszifikációs eljárást alkalmaztak a csődelőrejelzésben, melyek közül a diszkriminanciaanalízis, a logisztikus regresszió és a neurális hálók tekinthetők domináns eljárásnak. *Nikolic et al.* [2013] szerint a logisztikus regresszió napjainkig tartó népszerűsége könnyű gyakorlati alkalmazhatóságának köszönhető. Ennek ellenére a tudományos kutatás főáramát továbbra is a módszertani összehasonlító tanulmányok uralják. Szinte kivétel nélkül mindegyik arra a következtetésre jut, hogy a gépi tanulásra épülő adatbányászati módszerek – mint például a neurális hálók vagy újabban az SVM (support vector machine) – sokkal hatékonyabban képesek feltárni a függő és a független változók közötti komplex nemlineáris kapcsolatrendszert, mint a vizsgált adathalmazzal szemben súlyos előfeltevéseket támasztó matematikai-statisztikai eljárások (diszkriminanciaanalízis, logisztikus regresszió). Ennek pozitív hatása a csődmodellek jobb előrejelző teljesítményében érhető tetten, ami önmagában kedvező tendencia.

A fő kutatási irány létezését ma már nemzetközi folyóiratcikkek explicit módon is kimondják: *Yu et al*. [2014] megállapítása szerint ugyanis napjainkban a főáram gépi tanulásra épülő eljárásokat alkalmaz a csődelőrejelzésben. Ebből adódóan az ezen

kívül eső témakörök a csődelőrejelzés alternatív kutatási irányait képezik, ahol számos nyitott kérdés van, melyek megválaszolása akár a módszertani fejlesztésekhez hasonló mértékben is javíthatja a modellek előrejelző képességét. Ezt illusztrálandó bemutatok néhány nemzetközi kutatási eredményt a csődmodell-építés folyamatának egyes részterületeihez kötődően.

Az olvasó számára meglepőnek tűnhet, de már maga az adatgyűjtés is önálló kutatási területnek tekinthető. A nemzetközi kutatási eredmények azt mutatják, hogy a gazdaság egészének állapota jelentős hatással van a csődelőrejelző modellek pontosságára. A *Du Jardin–Séverin* [2012] szerzőpáros francia vállalatok adatait elemezve arra a következtetésre jutott, hogy a konjunkturális gazdasági környezet időszakában összegyűjtött adatokra épített csődmodellek szignifikánsan gyengébb teljesítményt mutatnak recesszió idején. Portugál kutatók pedig arra hívták fel a figyelmet, hogy egyes iparágak olyan speciális sajátosságokkal rendelkeznek, amelyek iparágspecifikus csődmodellek felépítését indokolják (*Horta–Camanho* [2013]).

García et al. [2012] szerint a pontos előrejelzés szempontjából fontos szerepet játszik a modellépítéshez felhasznált adatok minősége, ami a megfigyelések számával, a független változók relevanciájával jellemezhető leginkább. Mivel a legtöbb módszer érzékeny az outlier megfigyelésekre, az adatok megfelelő előkészítése nagyon fontos tevékenység a hitelkockázati modellek felállítása során (*García et. al* [2012]). Az idézett szerzők azt is kiemelik, hogy napjainkban relatíve kevés tanulmány foglalkozik az outlier megfigyelések kezelésével annak ellenére, hogy azok jelenléte a csődelőrejelzés alapvető sajátosságának tekinthető (*McLeay–Omar* [2000]). Előbbire példaként *Federova et al.* [2013] tanulmánya említhető. A szerzők orosz vállalkozások csődjét vizsgálják oly formában, hogy a modell felállítását megelőzően a kiugró értékkel bíró megfigyeléseket kitörlik a tanuló adathalmazból azok torzító hatására hivatkozva. Feltételezésem szerint az outlier megfigyelések fontos információtartalommal rendelkeznek, így modellben tartásuk indokolt lehet.

Hasonló következtetésre jutott *Yu et al.* [2014] is. Ők azt találták, hogy az egyes változók outlier értéke önmagában hiba nélkül képes azonosítani a fizetésképtelen vállalkozásokat az általuk vizsgált mintában. A kiugró értékkel rendelkező megfigyelések modellben tartásának módjára vonatkozóan azonban nincs egyértelmű iránymutatás a szakirodalomban. E problémára gyakran alkalmazott megoldás, hogy az outlier megfigyeléseket valamely szélső percentilis értékéhez igazítják. Ez a megközelítés akkor okozhat gondot, ha a maga szélső percentilis is outlier. Ekkor az adathalmazban továbbra is maradnak kiugró értékek. Problémát jelent az is, hogy jelenleg nincs egyértelmű definíció arra, hogy mikor nevezünk egy megfigyelést outliernek. Ilyen körülmények között praktikus lehet statisztikai hüvelykujjszabályokat alkalmazni. A cikk két megközelítést hasonlít össze: az egyik azon értékeket tekinti outliernek, amelyek standardizált értéke a 3 szórás terjedelmen kívülre esik; a másik pedig azokat, amelyek a 2 szórás terjedelmen kívül találhatók.

A főáramba tartozó publikációk többsége a csődmodellek input változóiként a legutolsó megfigyelt év „nyers" pénzügyi mutatóit használja. Ezáltal azonban a csődmodellek csak a mérleg fordulónapján készült „pillanatfelvételből" nyerhető információkat tudják felhasználni a fizetőképes és a fizetésképtelen vállalkozások megkülönböztetése céljából. E problémára praktikus megoldást javasolt *Berg* [2007], aki a legutolsó három év változókörét használta fel a csődmodellek input változói között. Az így létrejött csődmodell találati aránya szignifikánsan meghaladta azon modell pontosságát, amelyet az imént idézett szerző csak a legutolsó megfigyelt év adatai alapján állított fel.

Jelen tanulmány abból indul ki, hogy a vállalati gazdálkodás egy folyamat, melynek végső stádiuma a vállalkozás fizetésképtelensége, ebből adódóan azt feltételezem, hogy a mutatószámok nyers értéke mellett fontos információt hordozhatnak azok időbeli változásai is. E hipotézist a hazai vállalkozások köréből vett ezer elemű mintán vizsgáltam. Meghatároztam a csődelőrejelzésben leggyakrabban alkalmazott pénzügyi mutatószámokat, illetve azokat a változókat, amelyek időbeli változásukat hivatottak számszerűsíteni. További kérdés, hogy a pusztán nyers pénzügyi mutatókat tartalmazó modellekhez képest növelhető-e az elérhető találati arány, ha az input változók körében szerepeltetjük a mutatók időbeli változását kifejező változókat is. Ez a kérdés már túlmutat az adatok modellezésre történő előkészítésén, ugyanis a változószelekció folyamatában kaphatunk majd választ arra, hogy hordoz-e releváns információt a pénzügyi mutatók dinamikája.

Hasonlóképp nem húzható éles határ a főáramnak tekinthető módszertani összehasonlító kutatások és a változószelekció között. Ez utóbbi szintén kritikus kérdése a csődelőrejelzésnek, mivel máig nem született olyan egységesen elfogadott elmélet, amely meghatározná a csődmodellekben szerepeltetendő magyarázóváltozók körét (*Nikolic et al.* [2013]). Ez azért jelent problémát, mert *Du Jardin* [2010] szerint csaknem határtalan azon pénzügyi mutatók száma, amelyek a csődmodellek input változóiként felhasználhatók. A számosságot az idézett szerző azzal érzékelteti, hogy az elmúlt ötven évben megjelent publikációkban több mint 500 különböző pénzügyi mutatót használtak fel a szerzők a modellépítés során. Kiemelkedően fontos a legoptimálisabb változócsoport azonosítása a rendkívül nagyszámú lehetséges magyarázóváltozók közül, mert a feleslegesek jelentős mértékben befolyásolják az adatbányászati módszerek teljesítményét. E negatív hatás egyrészt a módszerek futási idejének növekedését, másrészt klasszifikációs teljesítményének romlását eredményezheti (*Wang et al.* [2014]). Ebből kiindulva vált önálló kutatási területté a csődelőrejelzés szempontjából legjobb teljesítményt nyújtó változószelekciós eljárások összehasonlítása, ami napjainkban is a tudományos érdeklődés tárgyát képezi (lásd például *Lin et al.* [2014] tanulmányát).

A csődmodellek teljesítményének megítélésére számos mutató áll az elemző rendelkezésére, melyek közül a legelső és leggyakrabban alkalmazott a csődmodell talá-

lati aránya, amelyet a helyesen klasszifikált megfigyelések számának az összes megfigyelés számához viszonyított arányként határozhatunk meg. Mivel a modellek előrejelzési céllal készülnek, fontos tényező e képességük a tanulási mintában nem szereplő megfigyelések vonatkozásában. Ennek megítélésére leggyakrabban az ún. keresztvalidációs eljárást alkalmazzák, melynek lényege, hogy a rendelkezésre álló adatbázist véletlenszerűen n egyenlő részre osztják, melyek közül $n - 1$ részt a modell felállítására, a kimaradt egy részt pedig annak tesztelésére használják fel. A modell így összesen n alkalommal állítható fel annak érdekében, hogy minden részlete szerepeljen egyszer tesztelő mintaként. Az n modellfuttatás elvégzését követően az egyes modellek klasszifikációs teljesítményét átlagolni kell. Ezen átlagos találati arány szolgál a csődmodell előrejelző képességének megítélésére.

A keresztvalidációs eljárás alkalmazásával csökken annak esélye, hogy egy véletlenszerű mintafelosztáson elért eredmény alapján téves következtetéseket vonjunk le a modell teljesítményéről. Ekkor is fennáll azonban annak kockázata, hogy az n egyenlő részre történő felosztás során az osztópontok kijelölése torzítja a modell előrejelző képességének objektív megítélését. Ennek elkerülése érdekében az n-szeres keresztvalidációs eljárást gyakran megismételik újabb véletlenszerű osztópontok kijelölésével. Ezt az eljárást k-szor n-szeres keresztvalidációnak nevezhetjük. Ebben az esetben a validációt k-szor kell ismételni, minden lépésben újabb véletlenszerű osztópontok alkalmazásával.

A keresztvalidáció speciális esete az ún. egyet kihagyó keresztvalidáció (leave one out). Az eljárás során az n elemű mintából minden lépésben egyetlen elemet hagyunk ki tesztelési céllal, míg a maradék $n - 1$ megfigyelés a modell tanuló adathalmazaként szolgál. Ekkor a minta elemszámával azonos számú modell állítható fel annak érdekében, hogy minden megfigyelés szerepeljen egyszer tesztelő elemként. E módszer alkalmazása esetén a modell teljesítményét a tesztelő elemek besorolásai alapján számított találati arány segítségével számszerűsíthetjük.

Egy másik gyakran alkalmazott validációs eljárás lényege, hogy a rendelkezésre álló mintát egyetlen véletlen osztópont segítségével – előre rögzített arányban – tanuló és tesztelő mintákra osztjuk fel. Annak érdekében, hogy minél objektívebb legyen a modell előrejelző képességének megítélése, a véletlenszerűen kiválasztott osztópontok segítségével a felosztást többször is megismételjük. Fontos megjegyezni, hogy a mintafelosztás arányára szintén nincs egyértelmű iránymutatás a szakirodalomban, ami azért fontos kérdés, mert *Hu* [2009] eredményei szerint a felosztás arányának változtatása akár 2-3 százalékpontos változást is eredményezhet a modellek besorolási pontosságában.

A csődmodellek validációjának szükségessége szempontjából konszenzus mutatkozik a szakirodalomban. A validációs eljárás módja tekintetében azonban már lényeges különbségek tapasztalhatók. Alig találni olyan publikációkat, amelyek azonos eljárást alkalmaznak a modellek validációja során. Arra sem láttam példát, hogy a

szerző megindokolta volna, miért épp az általa alkalmazott validációs eljárást választotta egy másik helyett. Ebből kiindulva a tanulmány az előbbiekben bemutatott három validációs technikát kívánja összehasonlítani annak érdekében, hogy feltárja, mutatkozik-e érdemi különbség a csődmodellek becsült előrejelző képességében attól függően, mely validációs módszer segítségével becsülték azt.

A nemzetközi tendenciák nem hagyták érintetlenül a magyar csődelőrejelzés fejlődését sem. *Kristóf* [2005] munkáját követően számos módszertani összehasonlító tanulmány született Magyarországon is, melyek közül fontos kiemelni *Virág–Kristóf* [2009] cikkét, melyben a többdimenziós skálázás és a logisztikus regresszió módszereinek együttes alkalmazásával kiemelkedő besorolási pontosságot mutató modellt állítottak fel. A későbbiekben ugyanez a szerzőpáros az adatok előkészítése területén folytatott kutatásokat, amely során a főkomponens-elemzés információsűrítő eszközét, valamint a CHAID-alapú (Chi-squared Automatic Interaction Detector) döntési fák segítségével kategorizált változók előrejelző képességét vizsgálták a csődelőrejelzésben leggyakrabban alkalmazott klasszifikációs eljárások keretein belül (*Kristóf–Virág* [2012]).

Szintén az adatok megfelelő előkészítésének fontosságára hívja fel a figyelmet *Virág–Nyitrai* [2013]. Ők az első hazai csődmodell adatbázisán végeztek kísérleti modellfuttatásokat az SVM-módszer alkalmazásával, amelyet napjainkban a legkorszerűbb módszerként tartanak számon a csődelőrejelzésben (*Horta–Camanho* [2013]). Az előbb idézett magyar szerzőpáros eredményei szerint az SVM-módszerrel felállított modellek besorolási pontossága meghaladja a korábban *Virág–Kristóf* [2005] által ugyanezen az adatkörön neurális hálókkal elért találati arányt.

Jelen tanulmányban a hazai szakirodalomban kevésbé elterjedt klasszifikációs eljárást, a döntési fák csoportjába tartozó C4.5 módszert alkalmazom. A modell elméleti alapjait ismerteti röviden a következő fejezet.

2. A C4.5 eljárás módszertani háttere

A döntési fát felállító eljárás kidolgozása *Quinlan* [1993] nevéhez kötődik. A döntési fák hallatán sokaknak a CHAID-módszer jut eszébe, melynek módszertani hátteréről korábban *Hámori* [2001] munkájában olvashattak a *Statisztikai Szemle* Olvasói. A CHAID-módszer a döntési fa ágaztatását a függő és a független változók egyes kategóriái közötti χ^2-alapú függetlenségvizsgálat eredményének figyelembevételével végzi. Ezzel szemben a C4.5 eljárás a fa felállításakor az egyes ágaztatásokból származó információs hasznot (information gain) veszi alapul. A módszer elmé-

leti hátterét *Quinlan* [1993] munkája szolgáltatja. Az eljárás működési elvét az imént idézett szerző jelöléseinek felhasználásával ismertetem a továbbiakban.

Tegyük fel, hogy mintánkban szerepel S darab megfigyelés, amelyek $C_j \ \left(j = 1,2,...,k\right)$ darab osztályba sorolhatók. Az információelmélet az üzenet információtartalmát azzal méri, hogy milyen valószínűséget fejez ki. Az információ mennyiségét egy p valószínűségű esemény esetén a $-\log_2 p$ mértékkel adhatjuk meg. Abban az esetben, ha ismertté válik, hogy egy megfigyelés egy C_j osztályba tartozik, akkor a következő valószínűség értéke áll információként rendelkezésre:

$$\frac{f\left(C_j, S\right)}{\left|S\right|},$$

ahol f a C_j osztály gyakoriságát, míg az $\left|S\right|$ a minta teljes elemszámát jelöli.

Ezeket felhasználva az S elemű adathalmaz entrópiáját a következő kifejezéssel adhatjuk meg:

$$info\left(S\right) = -\sum_{j=1}^{k} \frac{f\left(C_j, S\right)}{\left|S\right|} \cdot \log_2 \frac{f\left(C_j, S\right)}{\left|S\right|}. \qquad /1/$$

Ez lényegében azt a bizonytalanságot méri, amely azzal kapcsolatban merül fel, hogy a megfigyelések melyik osztályba tartoznak. Másként fogalmazva az /1/ kifejezés azt az információmennyiséget fejezi ki, amely ahhoz szükséges, hogy az S darab megfigyelés kategorizálásával kapcsolatos bizonytalanságot megszüntesse.

Ha egy X ismérv n változatának megfelelően az adathalmazt felosztjuk, akkor annak entrópiáját az /1/ kifejezés alapján kapjuk:

$$info_X\left(T\right) = \sum_{i=1}^{n} \frac{\left|T_i\right|}{\left|S\right|} \cdot info\left(T_i\right), \qquad /2/$$

ahol T_i az X ismérv i-edik változata szerint képzett csoportot jelöli.

Ha az adathalmaz X ismérv változatai szerinti felosztása nem triviális, akkor annak hatására az adathalmaz kezdeti entrópiája a következő mértéknek megfelelően csökken:

$$gain\left(X\right) = info\left(S\right) - info_X\left(T\right). \qquad /3/$$

Ebben a megközelítésben azon ismérv szerint érdemes a fát ágaztatni, amely esetén ez az információs nyereség a legnagyobb. A megközelítés problémája, hogy a minél több változattal rendelkező megfigyeléseket preferálja, emiatt az egyes felosztásokból nyerhető információs többletet célszerű valamilyen viszonyítási alap tekintetében értékelni. E célnak *Quinlan* [1993] szerint a leginkább az X ismérv n változatának teljes entrópiája felel meg, ami az /1/ kifejezés alapján:

$$split\ info(X) = -\sum_{i=1}^{n} \frac{|T_i|}{|S|} \cdot \log_2 \left(\frac{|T_i|}{|S|} \right).$$

Az adathalmaz X ismérv változatai szerint történő felosztásának hatására a teljes adathalmaz entrópiája csökken. Ennek mértékét a /3/ kifejezés adja meg, amelyet az előbb ismertetett probléma elkerülése érdekében célszerű az X ismérv változatainak teljes entrópiájához viszonyítani, ugyanis annak mértéke független a klasszifikációs feladattól. Az így kapott hányados:

$$gain\ ratio(X) = \frac{gain(X)}{split\ info(X)}. \qquad /4/$$

A *gain ratio* azt fejezi ki, hogy egy X ismérv változatainak információtartalmán belül mekkora arányt képvisel az az információs többlet, amely a klasszifikációs feladat szempontjából hasznosnak tekinthető.

A C4.5 eljárás a döntési fa ágaztatása során a /4/ kifejezés szerint definiált *gain ratio*-t veszi alapul. Ez azt jelenti, hogy a döntési fa első ágaztatása azon változó szerint történik, amely tekintetében az elágaztatás a legnagyobb relatív információs haszonnal jár. A további lépések során az egyes ágakon belül tovább folytatódik az ágaztatást a *gain ratio* maximalizálásának elvét követve mindaddig, amíg az eljárás valamely leállítási kritériuma be nem következik.

A *gain ratio* maximalizálása során figyelembe kell venni azt is, hogy értéke akkor is lehet magas, ha egy adott ismérv változatai szerint történő felosztásnak nincs nagy információtartalma. Emiatt a vizsgálat során a felhasználó megadhatja, hogy mi az a minimális információtartalom, amelyet elvár egy adott ismérv szerinti felosztástól. Ha egy adott ágon nem található olyan ismérv, amelynek változatai szerint történő felosztás információtartalma elérné ezt az előre meghatározott nagyságot, akkor az adott ág nem kerül továbbágaztatásra – még akkor sem, ha a *gain ratio* nagysága egyébként magas lenne. Ha a fa egyik ágán sem található ilyen ismérv, akkor a fa képzése megáll.

A C4.5 algoritmusának van más leállási feltétele is. Az e tanulmány elkészítéséhez felhasznált szoftver (Tanagra 1.4.49) esetén a felhasználó meghatározhatja a fa

ágaztatása során az újabb ágak létesítéséhez szükséges megfigyelések minimális számát is. A további ágaztatáshoz szükséges minimális információs többlet nagyságát azonban az alkalmazott szoftver esetén a felhasználó nem specifikálhatja.

A C4.5 segítségével egy klasszikus fa struktúrát kapunk eredményül. Más döntési fát generáló módszerekkel szemben (mint például a CHAID) annyi a különbség, hogy a fa kialakítása során az eljárás nem támaszkodik sem paraméteres technikákra, sem statisztikai hipotézisvizsgálatok eredményeire. Ebben a tekintetben tehát a C4.5 nem paraméteres adatbányászati módszernek tekinthető.

A döntési fák általános sajátossága a túltanulásra való hajlam. Ez azt jelenti, hogy az eljárás eredményeképp előálló fa túlságosan leképezi a minta egyedi sajátosságait, s emiatt a mintán kívüli megfigyeléseken jóval gyengébb teljesítményt mutat, mint a tanulási minta megfigyelésein. A túltanulás elkerülésének leghatékonyabb módja a fa nyesése, ami azt jelenti, hogy a döntési fa egyes ágait egy levéllel helyettesítik annak érdekében, hogy a fa ne „tanulja meg" a tanulási minta azon speciális sajátosságait, amelyek az elhagyandó ágon szerepelnek.

A C4.5 alkalmazása során a fa nyesése automatikus folyamat, amit szintén *Quinlan* [1993] alapján mutatok be. Tegyük fel, hogy egy ágon *N* megfigyelés szerepel, melyek közül *E*-t tévesen soroltuk be. E két adatból pedig becslés adható a téves besorolás sokasági arányának konfidenciaintervallumára. Quinlan szerint a 75 százalékos megbízhatóságú intervallum felső határa egyfajta pesszimista becslésnek tekinthető az egyes ágakon a téves besorolás valószínűségére. Ezek segítségével pedig mérhető, hogy mekkora a tévesen besorolt megfigyelések várható száma az eredeti, illetve a nyesésen átesett fa alkalmazásakor. Ha nyeséssel a tévesen besorolt megfigyelések várható száma alacsonyabb, mint az eredeti fán, akkor a nyesést érdemes elvégezni; ennek szigorúságát a felhasználó a megbízhatósági szint módosításával szabályozhatja.

3. A vizsgált adathalmaz

A kutatási kérdések megválaszolására saját adatgyűjtésből összeállítottam azt az ezer elemű mintát, amely 50-50 százalékos arányban tartalmaz fizetőképes és fizetésképtelen vállalatokat. Ebből adódóan a minta nyilvánvalóan nem tekinthető reprezentatívnak, azonban ez általánosnak mondható a csődelőrejelzésben. A csődös cégek felülreprezentálása azzal magyarázható, hogy a gépi tanulásra épülő eljárások egyenlőtlen megoszlás esetén hajlamosak a domináns csoport sajátosságaira specializálódni (*Horta–Camanho* [2013]). Ez a csődelőrejelzés kapcsán azért kimondottan hátrányos, mert a sokaságban kisebbséget alkotó csődös vállalatok téves besorolása

jóval költségesebb hiba, mint a működő vállalatok téves klasszifikációja. Ezeket rendre első és másodfajú hibaként ismeri a szakirodalom. Az aszimmetrikus költség miatt a modellezésben leginkább az elsőfajú hiba minimalizálására érdemes törekedni (*Du Jardin* [2010]).

A mintavétel során érvényesített szempontok:

> *1*. Alapvető elvárás volt, hogy legalább három évre visszamenőleg rendelkezésre álljanak a vizsgált vállalkozások pénzügyi mutatói (beszámoló, mérleg, eredménykimutatás). Ennek oka a kutatás azon célkitűzése, hogy vizsgálja azt is, bírnak-e diszkrimináló erővel azon változók, amelyek egy adott pénzügyi mutató értékét saját korábbi értékeinek tükrében mutatják.
>
> *2*. Kihagytam azokat a vállalkozásokat, melyeknek volt olyan pénzügyi mutatószáma, amely nem mutatott időbeli szóródást. Ez ugyanis lehetetlenné tenné olyan változók konstruálását, amelyek az egyes mutatók nagyságát saját korábbi értékeinek függvényében ítélik meg.
>
> *3*. Szintén kimaradtak a mintából az olyan megfigyelések, amelyek legalább két egymást követő évben nem realizáltak árbevételt. Ennek oka, hogy az ilyen vállalkozások vélhetően nem folytatnak érdemi gazdálkodást, így mintába kerülésüknek torzító hatása lenne a modellek eredményeire.

Du Jardin [2010] szerint a csődelőrejelzésben gyakran alkalmazott megközelítés azon pénzügyi mutatók használata magyarázóváltozóként, amelyek más vizsgálatokban eredményesnek bizonyultak. E tanulmány is ezt a megközelítést követi. Az input változók kiválasztása során az első hazai csődmodell változóit (*Kristóf* [2005]) és saját megfontolásaimat vettem figyelembe. A 18 mutató nevét és számításmódját az 1. táblázat tartalmazza. A mutatók számítása során az egyes mérlegtételeket, illetve az eredménykimutatás érintett sorait azok fordulónapi záró értékükön vettem figyelembe.

A csődmodellekben gyakran használt mutatószámok közé tartozik a sajáttőkearányos nyereség, amely gyakran veti a fel a *Kristóf* [2008] munkájában is felmerülő kettős negatív osztás problémáját. A probléma kezelésére nincs egyértelműen preferált megoldás, ezért e mutatót nem vettem figyelembe a számítások során.

A hányados típusú mutatószámok másik jellemző problémája akkor merül fel, amikor a tört nevezőjében nulla érték adódik. Ezt a gyakorlatban gyakran kezelik úgy, hogy az ilyen adatokat hiányzó értéknek tekintik, és a többi megfigyelés valamilyen középértékével, vagy azok valamelyik szélső percentilisével helyettesítik. Véleményem szerint e megközelítés nem feltétlenül visz konzisztens értéket a csődelőrejelző modellekbe. Az ebben a tanulmányban javasolt megoldás a következő példával illusztrálható. Tekintsünk egy olyan vállalkozást, amely rövid lejáratú kötelezett-

ségeit mindig azonnal, vagy jellemzően minden évben a mérleg fordulónapját közvetlenül megelőzően teljesíti, így ebben az időpontban nem rendelkezik rövid lejáratú kötelezettséggel, ami lehetetlenné teszi a likviditási ráta kalkulációját. Tételezzük fel, hogy a példában szereplő vállalkozás jelentős forgóeszköz-állománnyal is rendelkezik, ami lehetővé teszi számára, hogy egy később felmerülő esetleges „likviditási sokkot" képes legyen finanszírozni. Ha egy ilyen vállalkozás likviditási rátáját a mintában szereplő többi megfigyelés átlagával helyettesítenénk, akkor ezt a modell átlagos likviditású vállalatként tekintené, ami az adatai alapján nem helytálló. A másik lehetőség: valamely szélső percentilissel történő helyettesítés, amely már konzisztensebb információt visz a modellbe, de ekkor mintaspecifikus, hogy egy konkrét mutatószámot mivel helyettesítünk.

Ezen okfejtésből kiindulva az adatok előkészítése során azt a megoldást alkalmaztam, hogy azokban az esetekben, ahol a nevező értéke nulla lenne, ezt az értéket 1-gyel helyettesítettem. Így a példában szereplő vállalat likviditási rátája meglehetősen nagy értéket vesz fel, jelezve, hogy a vállalat likviditása rendkívül magas.

1. táblázat

Az empirikus vizsgálatban felhasznált mutatók neve és számításának módja

Mutató	Számítás módja
Likviditási ráta	Forgóeszközök/Rövid lejáratú kötelezettségek
Likviditási gyorsráta	(Forgóeszközök – Készletek)/Rövid lejáratú kötelezettségek
Pénzeszközök aránya	Pénzeszközök/Forgóeszközök
Cash flow/Kötelezettségek	(Adózás utáni eredmény + Értékcsökkenési leírás)/Kötelezettségek
Cash flow/Rövid lejáratú kötelezettségek	(Adózás utáni eredmény + Értékcsökkenési leírás)/Rövid lejáratú kötelezettségek
Tőkeellátottság	(Befektetett eszközök + Készletek)/Saját tőke
Eszközök forgási sebessége	Értékesítés nettó árbevétele/Mérlegfőösszeg
Készletek forgási sebessége	Értékesítés nettó árbevétele/Készletek
Követelések forgási ideje	Követelések/Értékesítés nettó árbevétele
Eladósodottság	Kötelezettségek/Mérlegfőösszeg
Saját tőke aránya	Saját tőke/Mérlegfőösszeg
Bonitás	Kötelezettségek/Saját tőke
Árbevétel-arányos nyereség	Adózás utáni eredmény/Értékesítés nettó árbevétele
Eszközarányos nyereség	Adózás utáni eredmény/Mérlegfőösszeg
Követelések/Rövid lejáratú kötelezettségek	Követelések/Rövid lejáratú kötelezettségek
Nettó forgótőke aránya	(Forgóeszközök – Rövid lejáratú kötelezettségek)/Mérlegfőösszeg
Vállalat mérete	Az eszközállomány természetes alapú logaritmusa
Évek	A megfigyelt évek száma

Az előzőkben bemutatott mintavételi szempontok alapján csak olyan hazai vállalkozások kerülhettek be a mintába, amelyek beszámolói legalább három évre visszamenőleg hozzáférhetők voltak. Az adatgyűjtés eredményeképp minden megfigyelés esetén rendelkezésre állt egy legalább 3 és legfeljebb 12 elemű idősor valamennyi pénzügyi mutatószám esetén. A teljes adatbázis a vizsgált 500 működő vállalkozás tekintetében összesen 4 194 üzleti évre, az 500 fizetésképtelen megfigyelés esetén pedig 3 398 üzleti évre vonatkozóan tartalmazott adatokat.

Az egyes megfigyelések pénzügyi mutatószám-idősorának elemeiből többféle formában is konstruálhatók olyan változók, amelyek azt fejezik ki, hogyan viszonyul egy vállalat legkésőbb megfigyelt pénzügyi mutatója azokhoz a korábbi értékekhez, amelyet ugyanezen vállalat azonos mutatója a korábbi üzleti évek során felvett. Az empirikus vizsgálatban a következő formulát használtam:

$$\frac{X_{i,t-1} - X_{i,\min_{(t-2;t-n)}}}{X_{i,\max_{(t-2;t-n)}} - X_{i,\min_{(t-2;t-n)}}} \qquad /5/$$

Az /5/-ös formula egy adott vállalkozás *i*-edik pénzügyi mutatószáma esetén azt számszerűsíti, hogy a legutolsó megfigyelt érték hol helyezkedik el az azt megelőző időszak szóródásának terjedelmén belül.

A javasolt változó számításához tehát felhasználtam a legutolsó megfigyelt évet megelőző legalább 2 és legfeljebb 11 elemű pénzügyi mutatószám-idősor tagjait a rendelkezésre álló adatsor hosszúságának függvényében. A gyakorlati elemzés azt mutatta, hogy adott vállalat konkrét pénzügyi mutatószámának idősorán belül mutatkoznak olyan évek, amikor annak értéke a többi évhez képest kiugróan magas, illetve alacsony értéket vett fel. Ez azért jelent problémát, mert az /5/-ös formula nevezőjében a mutatók szóródásának terjedelme szerepel, amit a legnagyobb és legkisebb megfigyelt érték különbségeként határozhatunk meg. Annak érdekében, hogy ezen intervallum hosszára ne legyenek hatással a szélsőségesen magas, illetve alacsony értékkel jellemezhető évek mutatói, az outliereket pótoltam azon év minimális vagy maximális értékével, amely még nem minősül outliernek. E korrekcióhoz azonban egyértelműen definiálni kell, hogy melyik tekinthető szélsőségesen alacsony vagy magas értéknek. A gyakorlati elemzés számára erre vonatkozóan nincs egységes definíció. Emiatt statisztikai hüvelykujj-szabályokat alkalmaztam az outlierek azonosítására. Minden megfigyelés esetén az egyes mutatószám-idősorokat standardizáltam az idősor átlagával és szórásával. Ezen standardizált értékek alapján az minősült outliernek, amely *a*) a 3, vagy *b*) a 2 szórás terjedelmen kívüli értéket vett fel. Ezt követően az outliereket helyettesítettem az adott megfigyelés olyan értékével, amely a korrigálandó értékhez a legközelebb esik, de még nem minősül kiugrónak.

4. Az empirikus vizsgálat eredményei

Az előző fejezetben bemutatott C4.5 módszert három adatbázison alkalmaztam, azok standard beállításai mellett; a számításokat az oktatási és kutatási célokra szabadon hozzáférhető Tanagra 1.4.49 szoftverrel végeztem. A szoftver alapbeállításai szerint a túltanulás elkerülése érdekében követelmény volt, hogy egy új ág keletkeztetéséhez legalább 5 megfigyelés szükséges, valamint az eljárás kidolgozójának javaslatára a fa nyesésénél 75 százalékos megbízhatóságú konfidenciaintervallum felső határát alkalmaztam (lásd részletesebben a harmadik fejezetet, illetve *Quinlan* [1993] munkáját). A modellek előrejelző képességét háromféle validációs technikával becsültem. Az egyet kihagyó eljárásnál a minta 999 eleme szolgált tanuló mintaként, míg az egyetlen kihagyott megfigyelés a tesztelő mintaként. E validációs módszernél minden megfigyelés szerepel egyszer tesztelő elemként, az ezekre kapott előrejelzések adják a modell összesített validációs eredményét. A második validációs eljárás a 100-szor 10-szeres keresztvalidáció volt. Ennek lényegét a második fejezetben röviden bemutattam. A harmadik validációs technika az ezer elemű minta véletlenszerű felosztása volt tanuló és tesztelő mintákra. Annak érdekében, hogy az eredményeket ne befolyásolja a felosztás mértéke, négy felosztási arányt is alkalmaztam: 90:10, 80:20, 70:30, 60:40 (tanuló:tesztelő). E véletlen felosztásnál is fontos szerepe lehet a felosztási pont kiválasztásának, ezért mind a négy esetben 100 véletlenszerűen kiválasztott osztópontot használtam, majd a 100 tesztelő mintán elért találati arány átlaga került be az eredményeket összesítő táblázatba.

Összegezve a számításokat: az egyet kihagyó eljárásnál 1 000 modellfuttatás átlagos eredményeit, a keresztvalidációs eljárás 100-szoros alkalmazásával 1 000 modell átlagos eredményeit, míg a tanuló-tesztelő felosztásnál a négy arány esetén 100-100, azaz összesen 400 modell eredményeit átlagoltam; ez adathalmazonként 2 400, összességében pedig 7 200 modellfuttatás átlagos eredményét jelenti. Ilyen elemszámnál az egyes kiugró értékek már kevésbé képesek az átlagok érdemi elmozdítására, így vélhetően az itt kapott sorrend kellően jó becslést nyújt az egyes validációs módszerek közötti különbségek megítélésére. A modellezés eredményeit a 2. táblázat foglalja össze.

A korrigálatlan adathalmazon nem került sor az outlier értékek miatt változtatásra. A „2 és 3 szórás" névvel illetett adathalmazok esetén az egyes vállalatok pénzügyi mutatószám-idősorában azon értékek, amelyek az idősoron belül a 2, illetve 3 szórás tartományon kívülre estek, helyettesítettem a hozzájuk legközelebb eső, de már nem outlier értékkel. Fontos hangsúlyozni, hogy az idősorok legutolsó éveit ez a korrekció nem érintette. Így a kizárólag nyers pénzügyi mutatókat tartalmazó változókon futtatott modellek esetén csak egyetlen eredményt tartalmaz a táblázat.

2. táblázat

A csődmodellek átlagos találati aránya
(százalék)

Validációs módszer	Adathalmaz	Változókör		Átlag
		Nyers	Nyers és dinamikus	
Egyet kihagyó	Korrigálatlan	74,0	74,9	75,2
	3 szórás		76,1	
	2 szórás		75,7	
Keresztvalidáció (100 × 10)	Korrigálatlan	74,9	75,5	75,5
	3 szórás		75,7	
	2 szórás		75,9	
Tanuló-tesztelő (90:10)	Korrigálatlan	74,7	75,1	75,4
	3 szórás		75,7	
	2 szórás		76,3	
Tanuló-tesztelő (80:20)	Korrigálatlan	74,9	75,0	75,3
	3 szórás		75,3	
	2 szórás		76,0	
Tanuló-tesztelő (70:30)	Korrigálatlan	74,5	75,1	75,1
	3 szórás		75,2	
	2 szórás		75,6	
Tanuló-tesztelő (60:40)	Korrigálatlan	74,6	74,7	75,0
	3 szórás		75,2	
	2 szórás		75,5	
Átlag		74,6	75,5	75,3

A csak nyers pénzügyi mutatókra épített modellek stabilan 74,5 százalék körüli előrejelző teljesítményt mutattak. Az olvasó számára ez relatíve alacsonynak tűnhet. Fontos azonban hangsúlyozni, hogy a tanulmánynak nem volt célja egy konkrét előrejelző modell felállítása. Illetve ki kell emelni, hogy a minta rendkívül heterogén a vállalatok mérete, kora és tevékenységi köre szempontjából. Továbbá figyelembe kell venni az eredmények megítélésénél, hogy a minta 50 százalékban tartalmaz csődös vállalatokat, így a modellek előrejelző képessége a véletlen találgatást érdemben meghaladja. A különböző validációs eljárások között számottevő eltérés nem adódott, csupán az egyet kihagyó módszer mutat valamivel alacsonyabb találati arányt a többi technikához képest.

A táblázat legalsó sora az egyes validációs eredmények egyszerű számtani átlagát tartalmazza. Az eredmény arra utal, hogy javul a csődmodellek várható besorolási pontossága, ha a nyers mutatószámok mellett figyelembe vesszük a mutatószámok időbeli változását is. Az 1 százalékpontos differencia ugyan alacsonynak tűnik, de az

a 7 200 modellfuttatást tartalmazó kísérleti kutatás eredményeiből adódott, ahol a futtatások számából kiindulva nagyon alacsony annak esélye, hogy véletlen tényezők hatására tér el a két eredmény. Élesebb különbség mutatkozik, ha az összevetést a különböző adathalmazokon végezzük el. Például a tanuló-tesztelő (90:10) mintafelosztás esetén a 2 szórás szabály szerint korrigált adatokon a különbség 1,6 százalékpont; az egyet kihagyó eljárásnál a 3 szórás szabály szerinti korrekció mellett pedig 2,1 százalékpont. Ezeken túl fontos hangsúlyozni, hogy valamennyi vizsgált esetben nagyobb találati arány adódott a nyers pénzügyi mutatók önálló alkalmazásához képest abban az esetben, amikor azok mellett a dinamikus változók is szerepeltek a független változók között. Az eredmények tehát empirikus bizonyítékkal szolgálnak a változók időbeli trendjének csődmodellekben történő szerepeltetésének létjogosultságára.

A 2. táblázat arra a kérdésre is választ nyújt, hogy érdemes-e a pénzügyi mutatószám-idősorok esetleges kiugró értékeit korrigálni. Bármely validációs eljárást tekintjük, a válasz: „igen", mivel mindkét korrigált adathalmaz esetén a találati arány meghaladta a korrigálatlan adathalmazon elért pontosságot. A tanulmány arra a kérdésre is választ keresett, hogy e korrekció során szigorúbb vagy megengedőbb szabályt érdemes-e alkalmazni az outlier értékek azonosítására. Az eredmények alapján a pénzügyi mutatószám-idősorok esetén azokat az értékeket, amelyek az idősoron belül a 2 szórás terjedelmen kívül esnek, célszerű helyettesíteni a hozzájuk legközelebb eső, de még nem outlier értékkel. Ekkor ugyanis jó eséllyel növekszik az elérhető találati arány. Az eredmény alól csak egyetlen kivétel adódott, az egyet kihagyó eljárás, ahol a 3 szórás terjedelem alkalmazásával kapott eredmények hoztak magasabb besorolási pontosságot. Ettől függetlenül kijelenthető, hogy a mutatószám-idősorok adatai esetén szükséges az outlier értékek korrekciója a jobb előrejelző teljesítmény érdekében.

A tanulmány fő célkitűzése azonban a különböző validációs eljárások eredményei közötti különbségek elemzése volt. Az egyes adatkörökön kapott eredmények egyszerű számtani átlagait mutatja a 2. táblázat utolsó oszlopa. A különbségek ezúttal is alacsonynak tűnhetnek, de a modellfuttatások számossága alapot nyújthat az átlagok közötti különbségek értékelésére. Megállapítható, hogy az egyet kihagyó keresztvalidáció valamelyest pesszimistább technika. Vélhetően ez és a rendkívül nagy számításigény állhat annak hátterében, hogy sokkal ritkábban alkalmazzák a tudományos publikációkban. A legoptimistább validációs eredményeket a nemzetközi szakirodalomban leggyakrabban választott keresztvalidációs módszer, valamint a 90-10 arányú véletlen felosztás mutatta. A tanuló és tesztelő minták véletlen felosztásainak eredményeinél megfigyelhető az is, hogy a tesztelő minta arányának növekedésével egyre csökken a modellek találati aránya. Ennek oka vélhetően az, hogy a tanuló minta méretének csökkenésével folyamatosan fogy a modellek számára rendelkezésre álló információ a tesztelő mintában szereplő vállalatok helyes besorolásához, így csökken a tesztelő mintákon elérhető találati arány is.

Mivel az átlagos modellteljesítmények között tapasztalt különbség meglehetősen alacsony (0,5 százalékpont körüli) az egyes validációs módszerekkel kapott eredmények tekintetében, azokat további statisztikai vizsgálatnak vetettem alá. Az egyet kihagyó módszer esetén nincs, de a keresztvalidációs módszerrel, illetve az adathalmaz véletlenszerű tanuló-tesztelő mintára történő felosztása tekintetében van lehetőség az egyes modellfuttatások eredményeképp kapott találati arányok eloszlásának statisztikai vizsgálatára. A modellfuttatások alapstatisztikai mutatóit foglalja össze a 3. táblázat.

3. táblázat

A modellfuttatások alapstatisztikai mutatói

Mutató	Keresztvalidáció (100 × 10)	Tanuló-tesztelő			
		(90:10)	(80:20)	(70:30)	(60:40)
Átlag (százalék)	75,51	75,41	75,27	75,05	75,01
Medián (százalék)	75,60	75,25	75,00	75,08	75,00
Módusz (százalék)	75,70	77,23	74,50	74,42	76,25
Szórás (százalék)	1,18	4,45	3,10	2,35	2,27
Csúcsosság	0,187	0,012	–0,269	–0,009	0,049
Ferdeség	0,022	0,007	0,034	0,133	0,195
Maximum (százalék)	78,80	89,11	83,50	82,06	80,75
Minimum (százalék)	70,90	60,40	66,00	67,77	67,25

Megjegyzés. A számítások az Excel 2007 programcsomag használatával készültek.

A 3. táblázat eredményei az egyes validációs módszerek közötti különbségek további elemzését teszik lehetővé, és a szakirodalomban leggyakrabban alkalmazott keresztvalidációs eljárás alkalmazását indokolják nemcsak amiatt, mert ez mutatja a legoptimistább képet a vizsgált modellek becsült előrejelző képessége tekintetében, hanem azért is, mert ezzel a módszerrel a legalacsonyabb a találati arányok szóródása. Jóval nagyobb szóródás jellemzi a minta véletlenszerű tanuló és tesztelő mintára történő felosztását. Minél alacsonyabb a tesztelő minta aránya, annál jelentősebb a szóródás. A 90:10 arány esetén például a szóródás terjedelme közel 30 százalékpont, ami azt jelenti, hogy „szerencsés” osztópont választása esetén közel 90 százalékos előrejelző teljesítmény is adódhat, míg „szerencsétlenebb” esetben ez az érték mindössze 60 százalék körüli. Ez az eredmény felhívja a figyelmet a nagyobb számításigényű validációs módszerek alkalmazásának fontosságára annak érdekében, hogy objektív és reális képet kaphassunk a klasszifikációs modellek teljesítményének megítélése szempontjából.

5. Összegzés

A tanulmányban bemutattam a csődmodell-építés folyamatát és legfontosabb lépéseit, valamint egy rövid áttekintést adtam az egyes részterületek vonatkozó nemzetközi szakirodalmából.

Elsődleges célom annak vizsgálata volt, hogy a tudományos publikációkban, illetve a modellezéshez felhasználható programcsomagokban általánosan elterjedt modellvalidációs eljárások között tapasztalható-e különbség az egyes modellek becsült előrejelző képessége között. Három validációs technikát vetettem össze: az egyet kihagyó eljárást, a keresztvalidációt és a minta véletlenszerű felosztását tanuló és tesztelő mintákra. A fő kutatási kérdés megválaszolásához 1 000 vállalkozás pénzügyi adatait tartalmazó adatbázist gyűjtöttem össze, amely több mint 7 500 üzleti évre vonatkozóan tette hozzáférhetővé a mintába került vállalkozások pénzügyi mutatóit.

További célként tűztem ki annak vizsgálatát, hogy érdemes-e a nyers pénzügyi mutatók mellett azok időbeli tendenciáját is figyelembe venni a csődmodellek input változóinak körében, illetve szükséges-e az e mutatók kalkulációja előtt a mutatószám-idősorokban esetlegesen előforduló kiugró értékeket korrigálni.

Tanulmányomban egy döntési fát felállító eljárást, a C4.5 algoritmust alkalmaztam. Tekintettel arra, hogy a módszer kevésbé elterjedt a gazdasági témakörökkel foglalkozó szakirodalomban, a cikkben röviden bemutattam a módszer elméleti hátterét is. A munkának nem volt célja konkrét előrejelzésre optimalizált csődmodell felállítása, emiatt a C4.5 módszert a standard beállítások mellett használtam.

A vizsgált három validációs módszer összehasonlítására összesen 7 200 modellfuttatás átlagos eredményei alapján került sor. A legpesszimistább modellteljesítmény az egyet kihagyó validációs módszer alkalmazásával adódott, míg a legoptimistább a keresztvalidációs eljárás, illetve a 90:10 arányban történő tanuló-tesztelő mintafelosztás alkalmazása esetén. Vélhetően ezzel magyarázható az előbbi alul, míg utóbbiak felülreprezentáltsága a szakirodalomban. A kutatási eredmények alapján arra jutottam, hogy a tanuló-tesztelő mintafelosztás vonatkozásában a tesztelő minta arányának növekedésével fokozatosan csökkent a modellek előrejelző képessége. Ez arra utal, hogy robusztus csődmodellek felállításához minél nagyobb mintaelemszámra van szükség.

A bemutatott eredmények arra is rámutattak, hogy a modellek előrejelző képessége növekszik, ha a mutatók nyers értékei mellett szerepeltetjük az azok időbeli változását kifejező változót a cikkben javasolt /5/-ös formula szerint. Az empirikus vizsgálat eredményei szerint e javulás mértéke tovább fokozható, ha /5/ alkalmazása előtt az idősorokban előforduló outliereket helyettesítjük a hozzájuk legközelebb eső, de már nem kiugró értékkel az idősoron belül.

Az elemzés azonban közel sem tekinthető teljesnek. A cikkben bemutatott vizsgálatok eredményei alapján levonható következtetések számos további kutatási kérdést vetnek fel, amelyek megválaszolása jövőbeli kutatások tárgya lehet.

A számítások eredményeképp meglehetősen alacsony, mindössze fél százalékponton belüli különbség adódott a vizsgált validációs módszerek átlagos teljesítményében. Kérdéses lehet, hogy ez az alacsony differencia mennyiben a választott klasszifikációs eljárás (C4.5), illetve az empirikus vizsgálat céljából felhasznált minta sajátossága. Általánosabb következtetések levonására azon jövőbeli kutatások adhatnak lehetőséget, amelyek a tanulmányban bemutatott elemzést más módszerek alkalmazásával, valamint homogénebb (például egy konkrét iparágra szűkített) minta adatain végzik el.

Érdekes jövőbeli kutatási terület lehet a különböző döntési fát generáló eljárások (mint a cikkben bemutatott C4.5, a CHAID, vagy a klasszifikációs és regressziós fák [classification and regression tree – CART]) előrejelző teljesítményének összevetése akár a csődelőrejelzés, akár más klasszifikációs feladat megoldása szempontjából, illetve annak vizsgálata, hogy javul-e a csődmodellek találati aránya, ha a nyers pénzügyi mutatók értékeit is korrigáljuk az outlier értékeknek megfelelően. Ebben a tekintetben pedig ismét kérdéses lehet, hogy javít-e a modellek találati arányán a dinamikus pénzügyi mutatószámok szerepeltetése az input változók körében.

Irodalom

BEAVER, W. [1966]: Financial Ratios as Predictors of Failure. Empirical Research in Accounting: Selected Studies. *Journal of Accounting Research*. Vol. 4. pp. 71–111.

BERG, D. [2007]: Bankruptcy Prediction with Generalized Additive Models. *Applied Stochastic Models in Business and Industry*. Vol. 23. No. 1. pp. 129–143.

DU JARDIN, P. – SÉVERIN, E. [2012]: Forecasting Financial Failure Using a Kohonen Map: A Comparative Study to Improve Model Stability Over Time. *European Journal of Operational Research*. Vol. 221. Issue 2. pp. 378–396.

DU JARDIN, P. [2010]: Predicting Bankruptcy Using Neural Networks and Other Classification Methods: The Influence of Variable Selection Techniques on Model Accuracy. *Neurocomputing.* Vol. 73. No. 10–12. pp. 2047–2060.

FEDEROVA, E. – GILENKO, E. – DOVZHENKO, S. [2013]: Bankruptcy Prediction for Russian Companies: Application of Combined Classifiers. *Expert Systems with Applications*. Vol. 40. Issue 18. pp. 7285–7293.

GARCÍA, V. – MARQUÉS, A. I. – SÁNCHEZ, J. S. [2012]: On the Use of Data Filtering Techniques for Credit Risk Prediction with Instance-Based Models. *Expert Systems with Applications.* Vol. 39. Issue 18. pp. 13267–13276.

HÁMORI G. [2001]: A CHAID alapú döntési fák jellemzői. *Statisztikai Szemle*. 79. évf. 8. sz. 703–710. old.

HORTA, I. M. – CAMANHO, A. S. [2013]: Company Failure Prediction in the Construction Industry. *Expert Systems with Applications*. Vol. 40. Issue 16. pp. 6253–6257.

HU, Y. C. [2009]: Bankruptcy Prediction Using ELECTRE-Based Single-Layer Perceptron. *Neurocomputing*. Vol. 72. Issue 13–15. pp. 3150–3157.

KRISTÓF, T. – VIRÁG, M. [2012]: Data Reduction and Univariate Splitting – Do They Together Provide Better Corporate Bankruptcy Prediction? *Acta Oeconomica*. Vol. 62. Issue 2. pp. 205–227.

KRISTÓF T. [2005]: A csődelőrejelzés sokváltozós statisztikai módszerei és empirikus vizsgálata. *Statisztikai Szemle.* 83. évf. 9. sz. 841–863. old.

KRISTÓF T. [2008]: A csődelőrejelzés és a nem fizetési valószínűség módszertani kérdéseiről. *Közgazdasági Szemle*. LV. évf. 5. sz. 441–461. old.

LIN, F. – LIANG, D. – YEH, C. C. – HUANG, J. C. [2014]: Novel Feature Selection Methods to Financial Distress Prediction. *Expert Systems with Applications*. Vol. 41. Issue 5. pp. 2472–2483.

MCLEAY, S. – OMAR, A. [2000]: The Sensitivity of Prediction Models to the Non-Normality of Bounded and Unbounded Financial Ratios. *British Accounting Review*. Vol. 32. Issue 2. pp. 213–230.

NIKOLIC, N. – ZARKIC-JOKSIMOVIC, N. – STOJANOVSKI, D. – JOKSIMOVIC, I. [2013]: The Application of Brute Force Logistic Regression to Corporate Credit Scoring Models: Evidence from Serbian Financial Statements. *Expert Systems with Applications*. Vol. 40. Issue 15. pp. 5932–5944.

QUINLAN, J. R. [1993]: C4.5: *Programs for Machine Learning.* Morgan Kaufmann. San Mateo.

VIRÁG M. – KRISTÓF T. [2005]: Az első hazai csődmodell újraszámítása neurális hálók segítségével. *Közgazdasági Szemle*. LII. évf. 2. sz. 144–162. old.

VIRÁG M. – KRISTÓF T. [2009]: Többdimenziós skálázás a csődmodellezésben. *Vezetéstudomány*. 40. évf. 1. sz. 50–58. old.

VIRÁG, M. – NYITRAI, T. [2013]: Application of Support Vector Machines on the Basis of the First Hungarian Bankruptcy Model. *Society and Economy*. Vol. 35. No. 2. pp. 227–248.

WANG, G. – MA, J. – YANG, S. [2014]: An Improved Boosting Based on Feature Selection for Corporate Bankruptcy Prediction. *Expert Systems with Applications*. Vol. 41. Issue 5. pp. 2353–2361.

YU, Q. – MICHE, Y. – SÉVERIN, E. – LENDASSE, A. [2014]: Bankruptcy Prediction Using Extreme Learning Machine and Financial Expertise. *Neurocomputing.* Vol. 128. March. pp. 296–302.

Summary

The author presents and compares various validation techniques applied in the literature of bankruptcy prediction. The main purpose of the study is to investigate the impact of different validation techniques on hit rates, which are widely used performance indicators in bankruptcy prediction. The answer is given by the results of an empirical research based on a database consisting of 1 000 Hungarian firms. The C4.5 method applied in this study is rare in the Hungarian literature, thus, the author also discusses its theoretical background. The article presents a formula which enables the readers to take into account the time dynamics of static financial ratios in bankruptcy prediction models.