A magyarázóváltozók kezelésének egyes kérdései regressziós modellezés során

Hámori Gábor
PhD, a STAT4U Bt.
kutatási igazgatója
E-mail: ghamori@uni-corvinus.hu

A gyakorlati prediktív regressziós modellezés folyamata során a modellek futtatásának egyik megelőző lépése az adatbázis magyarázóváltozóinak előkészítése, esetleges transzformációja a modell illeszkedésének és ezáltal az előrejelző képességének növelése érdekében. Ehhez kapcsolódóan a tanulmány két, tipikusnak mondható problémakört jár körül.

A nagyméretű, komplex adatbázisok elemzésénél gyakori, hogy relatíve alacsony a megfigyelések száma a lehetséges magyarázóváltozók számához képest. Ilyenkor szükségessé válhat a modellezésbe bevonandó változók számának valamilyen logika alapján történő előzetes csökkentése a változók előzetes szelektálása segítségével. A kérdés kezelésére a tanulmány bemutat egy, egyszerű végrehajthatósága okán, kedvelt khinégyzet-alapú előszűrési módszert, és felhívja a figyelmet az ehhez kapcsolódó értelmezési kockázatokra is.

A másik jellemzőnek mondható döntési helyzet a folytonos magyarázóváltozók előzetes transzformációjával kapcsolatos. A cikk írója a közkedvelt logisztikus regressziós modellt alkalmazva megvizsgálja, hogy milyen esetekben és hogyan célszerű az eredeti folytonos magyarázóváltozót kategorizált párjával helyettesíteni a modell magyarázó erejének növelése céljából.

A tanulmány elsősorban a gyakorlatban tevékenykedő modellezők, adatbányászok számára szeretne útmutatást adni, de reményeink szerint hasznos információkkal szolgálhat, az elméleti oktatás területén működő szakembereknek is.

TÁRGYSZÓ:
Előszűrés.
Events per variable (EPV).
Kategorizálás.

DOI: 10.20311/stat2016.01.hu0005

A gyakorlati statisztikai modellek kialakítása során szükségszerű az adatbázis megfelelő előkészítése a modellek futtatását megelőzően. Ez az általában meglehetősen időigényes munkafázis jelentős befolyással van a leendő modellek illeszkedésére és ezáltal gyakorlati alkalmazhatóságukra. A modellezés ezen szakaszának jellemző lépései a hiányzó értékek kezelése (*Oravecz* [2008]), a szélsőséges (outlier) értékek szűrése és az előzetes változótranszformációk. A hiányzó értékek kezelése és az outlierszűrés témakörét itt mellőzve, jelen tanulmány az előzetes változótranszformációk köréből tárgyal két gyakran előforduló problémakört.

A nagyméretű, komplex adatbázisok elemzésénél gyakori, hogy relatíve alacsony a megfigyelések száma a lehetséges magyarázóváltozók számához képest. Ilyenkor szükségessé válhat a modellezésbe bevonandó változók számának valamilyen logika alapján történő előzetes csökkentése, a változók előszűrése. A kérdés kezelésére felidézzük a gyakran alkalmazott khi-négyzet-statisztikára épülő szelekciós technikát, és felhívjuk a figyelmet az ehhez kapcsolódó értelmezési kockázatokra.

A másik jellemzőnek mondható döntési helyzet a folytonos magyarázóváltozók előzetes transzformációjával kapcsolatos. A közkedvelt logisztikus regressziós modellt alkalmazva megvizsgáljuk, hogy milyen esetekben célszerű az eredeti folytonos magyarázóváltozót kategorizált párjával helyettesíteni a modellezés során. Bemutatjuk, hogy a jól ismert CHAID- (chi-squared automatic interaction detector – khi-négyzet-alapú automatikus interakció-detektálás) algoritmus segítségével miként alakítható ki egy olyan kategóriastruktúra, mellyel várhatóan jobb illeszkedés érhető el, mint más egyszerű szabályok által végrehajtott kategorizálások esetén.

1. Előzetes változószelekció khi-négyzet-statisztika alapján

A többváltozós statisztikai modellek kialakítása esetében problémát okozhat, ha túl kevés a megfigyelések száma a lehetséges magyarázóváltozók számához képest. Logisztikus regresszió esetében az alacsony EPV- (events per variable – egy változóra jutó esetszám) értékek mellett a paraméterbecslések torzítottá válnak, és megnő a szélsőséges maximum likelihood becslés esélye is (*Peduzzi et al.* [1996]). A hivatkozott szerzők különböző EPV-értékek mellett elvégzett Monte-Carlo-szimuláció alapján, az EPV = 10 értéket javasolják használni minimumkritériumként a modellezési adatbázis kialakítása során.

Amennyiben az adatbázisra a választott modell tekintetében túl alacsony EPV jellemző,[1] és nincsen lehetőség, vagy nem ésszerű az esetszám pótlólagos bővítése, szükségessé válik a modellezésbe bevont lehetséges magyarázóváltozók számának csökkentése. Folytonos magyarázóváltozók esetén ennek egy lehetséges módja a hagyományos főkomponens-analízis alkalmazása, melynek során a kiinduló változószet információtartalmának jelentős részét néhány, a módszer által meghatározott főkomponensbe tömörítjük. Az eljárás hátránya, hogy csak folytonos változók esetében kínál megoldást. Ezért a továbbiakban bemutatunk egy hagyományos khi-négyzet-statisztikára épülő változószelekciós technikát, mely egyszerűsége, teljeskörűsége és gyors alkalmazhatósága okán méltán népszerű a gyakorlati modellezők körében.

A módszer abból a logikából indul ki, hogy a modellezés során azok a változók a legkevésbé értékesek, melyeknél a hordozott információtartalom nem, vagy csak csekély hozzájárulással bír az előre jelezni kívánt változó vonatkozásában. A lehetséges magyarázóváltozók és a célváltozó közötti egyváltozós kapcsolat erősségét alkalmasan megválasztott kapcsolatszorossági mérőszámmal tudjuk jellemezni. Közkedvelt és minden statisztikai programcsomagban megtalálható a khi-négyzet-statisztika, melynek segítségével két kategóriás mérési szintű változó közötti kapcsolat erősségét jellemezhetjük. Kategóriás mérési szintű változók esetén a célváltozó vonatkozásában minden kategóriás magyarázóváltozó esetében számolható a khi-négyzet-statisztikákhoz tartozó empirikus szignifikancia- (*p*-) érték. Folytonos mérési szintű változók esetén a khi-négyzet-statisztika használatának feltétele az, hogy először a folytonos változót célszerűen megválasztott számú osztóponttal kategóriaváltozóvá alakítjuk.[2] Az így kategorizált változókon a khi-négyzet-statisztikákat és a hozzájuk tartozó empirikus szignifikanciákat az előzőkhöz hasonlóan számoljuk. Az eljárás végére minden változóra rendelkezni fogunk empirikus szignifikancia-értékekkel. A változókat ezen értékek alapján sorba rendezve, az alacsonyabb *p*-értékek jelzik a szorosabb, míg a magasabb értékek a kevésbé szoros kapcsolatokat. A szűrést ezek után úgy hajtjuk végre, hogy a legmagasabb *p*-értékkel rendelkező változók közül elhagyunk annyit, amennyi a minimális EPV-arány eléréséhez szükséges.

[1] Enter vagy backward változószelekciós módszer használata esetén fordulhat elő leginkább ez a helyzet, mert ilyenkor az összes változó egyszerre kerül a modellbe.

[2] A módszer szoftvertámogatása egyes statisztikai/adatbányászati programcsomagokban megtalálható. A kategorizálás segítségével képet kaphatunk némely folytonos magyarázóváltozó és célváltozó közötti kapcsolatról, annak monoton vagy nemmonoton jellegéről. Az eljárás közben az eredeti folytonos változót célszerű megtartani annak érdekében, hogy amennyiben szükséges, a változó eredeti mérési szintjével tudjunk továbbdolgozni.

2. Előszűrés problémái – változók értékelésének egyváltozós és többváltozós megközelítése

Az előszűrés korábbiakban leírt khi-négyzet-alapú eljárása egyváltozós statisztikai technikára épül. A módszer célja, hogy a változók egyedi előrejelzési ereje szerint a változókat rangsorolni tudjuk. Az eljárás eredményeképpen eltávolítjuk az adatbázisból a célváltozóval látszólag gyenge vagy kapcsolatban nem levő magyarázóváltozókat. A problémát az okozza, hogy az egy változóban független prediktor többváltozós környezetben akár jelentős parciális hatással lehet a függőváltozó értékére, így elhagyása a modellből a predikciós erő csökkenését okozhatja. A probléma szemléltetésére tegyük fel például, hogy a budapesti lakásárak alakulására akarunk regressziós modellt készíteni. A mintát egy budai frekventált, valamint egy pesti alacsony presztízsű kerület megfigyelései alkotják. Az elemzés során azt találjuk, hogy az egyváltozós elemzés alapján a lakásárak és a lakások alapterülete között nincs statisztikailag szignifikáns kapcsolat. A többváltozós regressziós modell illesztése során azonban a lakás nagysága szignifikáns magyarázóváltozónak bizonyul. A jelenség oka alapvetően az, hogy a két kerületben a lakások négyzetméterára nagyban különbözik, a frekventált kerületben a kis lakásnak is lehet olyan magas ára, mint az alacsony presztízsű kerületben egy nagy lakásnak. Így látszólag a lakás nagysága nem hat az árra, azonban, ha a területi hovatartozást is bekapcsoljuk a modellbe a nagyság parciális hatása már jelentős lesz. A lakásokat a két kerületben külön-külön vizsgálva az ár és a nagyság között már van kapcsolat. A lakásnagyság változó előzetes kiszűrése az adatmintából tehát a modell előrejelző képességének romlását okozta volna. Hogy milyen jelentős következményei lehetnek az ilyen természetű egyváltozós tévkövetkeztetéseknek, azt a következő valós példa jól illusztrálja. Pár évvel ezelőtt egy hazai egészségügyi intézetben a műtét utáni szövődmények kialakulásának statisztikai előrejelzése volt a feladatunk. A műtéti eseteket két csoportba soroltuk annak megfelelően, hogy a műtétet követően kialakult-e az illető betegnél szövődmény, vagy sem. Az orvosi kutatásoknál megszokott és elvárt módon elvégeztük a szövődmény bekövetkezése és a lehetséges magyarázó faktorok egyváltozós kapcsolatának elemzését. Meglepő módon a dohányzási szokásokat leíró változónál az volt tapasztalható, hogy minél több cigarettát szív el valaki naponta, annál kisebb a szövődmény bekövetkezésének esélye. Az egyváltozós elemzés alapján úgy tűnt tehát, hogy a dohányzás védő hatású a műtét utáni szövődmény bekövetkezése tekintetében! A józan észnek ellentmondó eredmény azonnal érthetővé vált akkor, amikor a dohányzási szokásokat leíró változókat a többi változóval együtt vizsgáltuk. Megállapíthattuk, hogy az adatmintában a fiatalok körében jóval magasabb volt a dohányosok aránya, mint az idősebbek esetében. Így hipotézisünk szerint a magasabb cigarettaszám valójában az alacsonyabb életkor információját hordozta magában, és a dohányzás parciális hatása ténylegesen negatív a szövődményráta alakulására nézve. A

többváltozós elemzés igazolta ezt a vélekedést azáltal, hogy a dohányzás és az életkor növekedésének parciális hatása is negatívnak bizonyult.

Ebből a gondolatmenetből következik, hogy a minimális EPV-kritérium megtartása mellett az összes magyarázóváltozót célszerű bevonni a modellezésbe. Viszont amennyiben a magyarázóváltozók nagy száma miatt előzetes változószelekció indokolt, az elkészült modell interpretációjánál, a szignifikáns magyarázóváltozók paramétereinek értelmezésénél kellő óvatossággal célszerű eljárni. Ez különösen igaz akkor, ha a modellezés célja nem a maximális prediktív erejű modell elkészítése, hanem a modellezendő jelenség mechanizmusának megértése, a magyarázóváltozók és a célváltozók közötti kapcsolatrendszer feltárása a kutatás célkitűzése.

3. Folytonos változók kategorizálása

Amennyiben az adatbázisra jellemző magas EPV-érték miatt nincs szükség előzetes változószelekcióra, felmerül a kérdés, hogy az előszűrt folytonos változókat eredeti alakjukban vagy kategorizált párjukkal szerepeltessük a modellben. A kérdés megválaszolásához figyelembe kell venni, hogy a folytonos változók regressziós/prediktív modellben betöltendő szerepére hatást gyakorol, hogy van-e, és ha van, milyen jellegű a kapcsolata a célváltozóval.

Mint ismeretes, mind a lineáris, mind a logisztikus regressziós modell esetén, az alkalmazott függvénytípus következtében, csak a célváltozó tekintetében monoton kapcsolatot mutató folytonos változóknál várhatunk megfelelő illeszkedést (*Schechtman–Yitzhaki* [2012]). Ezért a nemmonoton kapcsolatot mutató változók[3] esetén a folytonos változót kategorizált alakjában javasolt szerepeltetni a modellben a prediktív erő növelése céljából. Kevésbé magától értetődő, hogy nemegyszer monoton kapcsolat esetén is előállhat olyan eset, amikor a folytonos változó nem szignifikáns, míg a kategorizált párja szignifikáns magyarázóváltozónak bizonyul modellünkben. Erre a helyzetre mutat példát a következő eset: egy egészségügyi adatbázisban az életkor (KOR) és a testtömegindex (BMI) folytonos változókkal szeretnénk előre jelezni a műtét utáni szövődmények (TARGSSI) előfordulását.[4] A következőkben láthatjuk a két változót egyszerre bevonó (enter) logisztikus regressziós modell paraméterbecslésével kapcsolatos statisztikákat feltüntető táblázatot.

[3] Tegyük fel, hogy a fizetéseket akarjuk regresszálni az életkor segítségével. A fizetések jellemzően az életkor előrehaladásával eleinte növekszenek, majd egy idő után csökkenni kezdenek. A kapcsolat fordított U alakú, folytonos formájában jó eséllyel nem szignifikáns változóként fog szerepelni a modellben.

[4] Az adatbázis 350 beteg adatait tartalmazta, akik közül 98 esetben fordult elő műtét utáni szövődmény. Az outputok SPSS 20 programcsomaggal készültek.

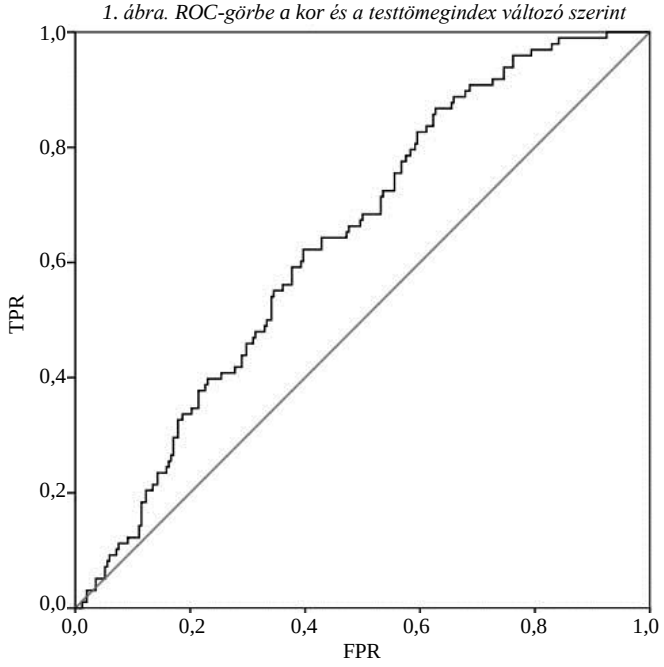
1. táblázat

Paraméterbecslés a kor és a testtömegindex változó bevonásakor

Változó	β	Standard hiba	Wald-statisztika értéke	Szabadságfok	Szignifikancia-szint	Exp(β)
BMI	0,010	0,009	1,187	1	0,276	1,010
KOR	0,032	0,009	13,088	1	0,000	1,032
Konstans	–3,024	0,585	26,709	1	0,000	0,049

Látható, hogy a két változó közül csak a KOR szignifikáns a szokásos szinteken. Ennek megfelelően a modell illeszkedése is nagyon gyenge, amit a Nagelkerke $R^2 = 0{,}089$ érték is mutat. Ezzel összhangban a Gini-együttható értéke is alacsony (28%) az 1. ábra ROC- (receiver operating characteristic – vevő működési karakterisztika) görbéjének megfelelően.[5]

1. ábra. ROC-görbe a kor és a testtömegindex változó szerint



Megjegyzés. TPR (true positive rate – igaz pozitív arány), FPR (false positive rate – hamis pozitív arány).

[5] A ROC-görbe és a Gini-együttható használatában kevésbé jártas olvasók a Függelékben találják ezen módszerek rövid leírását.

Nézzük meg, mi történik, ha a testtömegindex folytonos változó helyett a kategorizált párját használjuk a modellezéshez. A kategóriahatárokat, az új változó, a korrigált testtömegindex (BMIKAT2) kódolását és az egyes kategóriákhoz tartozó szövődményrátákat (kategórián belüli szövődményes esetek aránya a kategória összes esetéhez) láthatjuk a 2. táblázatban.

2. táblázat

A testtömegindex kategorizálása

BMI	BMIKAT2	TARGSSI (százalék)
0–28	1	12,70
28–34	2	31,40
34–	3	44,60

Látható, hogy kategóriák szerint monoton növekszik a szövődményráta értéke, tehát a BMI folytonos változónak a célváltozóval való kapcsolata monoton, azaz a magasabb BMI-értékek nagyobb szövődményrátákat vonzanak. A változó ennek ellenére nem bizonyult szignifikánsnak eredeti modellünkben. A regressziót újra futtatva az életkor folytonos és a testtömegindex kategorizált változóval a 3. táblázat szerinti eredményeket kapjuk.

3. táblázat

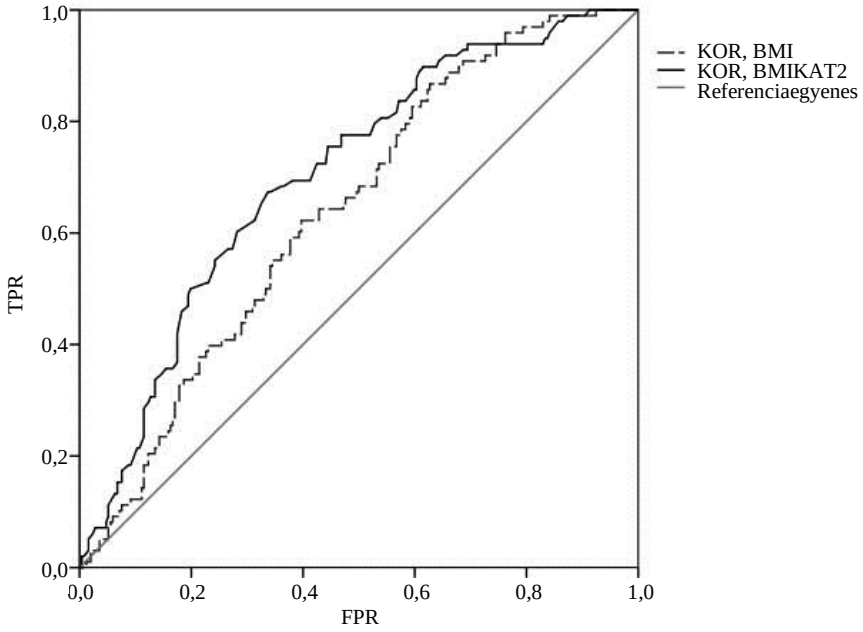
Paraméterbecslés a kor
és a kategorizált testtömegindex változó bevonásakor

Változó	β	Standard hiba	Wald-statisztika értéke	Szabadságfok	Szignifikancia-szint	Exp(β)
KOR	0,034	0,009	13,767	1	0,000	1,035
BMIKAT2	0,879	0,189	21,561	1	0,000	2,409
Konstans	–4,615	0,711	42,113	1	0,000	0,010

A testömegindex itt már szignifikáns, és ennek megfelelően alakul a modell illeszkedése is, amely most már sokkal jobb (Nagelkerke $R^2 = 0{,}121$, Gini-együttható = 40 százalék), mint az eredeti modellünk esetében. A 2. ábrán a két modellhez tartozó ROC-görbéket egyszerre tanulmányozhatjuk.

A példa rávilágít arra, hogy a folytonos változó kategorizálása még abban az esetben is javíthatja modellünk illeszkedését, ha a célváltozó és a folytonos magyarázóváltozó kapcsolata monoton.

2. ábra. ROC-görbe a kor és a kategorizált testtömegindex változó szerint



Az új kategóriaváltozók kódolása az adatbázisban dummy változók segítségével történik. Egy *K* kategóriával rendelkező változó esetében ez *K* – 1 darab dummy változó bevonását jelenti az adatbázisba.[6] A folytonos változók kategorizálásával az adatbázisban szereplő változók száma ezáltal jelentősen megnőhet. A kategorizálás során tehát mindig figyelemmel kell lenni az adatbázist jellemző EPV-érték alakulására.

4. Kategorizálás CHAID-algoritmussal

Miután megállapítottuk, hogy mely folytonos változó esetében javasolható a kategorizált párral történő helyettesítés, felmerül a kérdés, hogy hány kategóriával rendelkezzen az új változó, illetve hol legyenek a kategóriahatárok.[7] Mivel a kategóriák száma és elhelyezkedése hatást gyakorol a végső változó predikciós erejére a modellben, nem mindegy, hogy a folytonos skála felosztását milyen módon hajtjuk végre. Az illeszkedés szempontjából kedvező felosztás megtalálására több algoritmus is használható. Közös vonásuk, hogy a célváltozó tekintetében a felosztást úgy hajtják végre, hogy a létrejövő kategóriákon belüli homogenitás és a kategóriák közötti heterogenitás a

[6] Azzal a megszokott feltételezéssel élve, hogy modellünk tengelymetszettel rendelkezik.

[7] Az adatbányászati szakzsargonban a kategorizálási eljárást angolul „data-binning” illetik. Az elnevezés után a hazai adatbányász szóhasználatban közkedvelt a „változó binnelése” fogalom használata.

legnagyobb legyen. Bináris kategorizálás esetében az ún. osztópont-analízis (cutpoint-analysis) segítségével határozható meg a bináris kategóriaváltozó optimális osztópontja (*Vargha–Bergman* [2012]). A több kategóriával rendelkező kategóriaváltozók kialakítására a következőkben, elérhetősége és egyszerűsége okán, kategóriaegyesítési céllal,[8] a döntési fák családjába tartozó rekurzív klasszifikáló eljárás, az ún. CHAID-algoritmus (*Hámori* [2001]) részalgoritmusát alkalmazzuk. Célunk, hogy a K különböző kategóriával rendelkező változó[9] esetében összevonjuk azokat a kategóriákat, melyek legkevésbé különböznek egymástól az m különféle kategóriával rendelkező célváltozó tekintetében.[10] Ehhez az algoritmus X_i kategorizált folytonos változó kategóriái közül az összes lehetséges módon kiválaszt kettőt. Amennyiben a vizsgált magyarázóváltozó K különböző kategóriával rendelkezik, a kiválasztás $K\cdot(K-1)/2$ féleképpen történhet. Ezt követően $K\cdot(K-1)/2$ különböző, $(2 \times m)$ méretű kontingenciatáblázatra Pearson-féle khi-négyzet-teszt segítségével kiszámolja, hogy milyen p szignifikanciaszinten tekinthetők X_i kiválasztott kategóriapárjai és Y célváltozó kategóriái függetlennek egymástól. A következő lépésben kiválasztjuk azt a kontingenciatáblázatot, mely a legmagasabb p-értékkel rendelkezik. Ezt az értéket az eljárás összeveti egy, a modellkészítő által előre rögzített, $\alpha_{\text{egyesítés}}$ küszöbértékkel (a programcsomagok általában a szokásos 5 százalékos szignifikanciaszintet szokták felkínálni alapértelmezésként). Amennyiben $p > \alpha_{\text{egyesítés}}$ a kontingenciatáblázat X_i kategóriapárja, akkor egy új önálló kategóriába kerül egyesítésre. Ebben az esetben X_i eredeti kategóriáinak száma eggyel csökken, és az algoritmus újból indul az elejétől, azaz az „új" kategóriapárok kiválasztásától (amelyek között nyílván lehetnek olyanok is, melyek az előző ciklusban is kiválasztásra kerültek), az azokhoz rendelt kontingenciatáblázatokhoz tartozó p-értékek kiszámolásáig.

A kategóriák összevonásának ciklusa mindaddig folytatódik, míg a legmagasabb p-értékkel rendelkező kontingenciatáblázatokra igaz nem lesz a $p > \alpha_{\text{egyesítés}}$ feltétel. Ekkor a vizsgált magyarázóváltozó (X_i) esetében a ciklus leáll, és az algoritmus a következő lépésben már X teljes, lehetséges összevonások utáni, új kategóriastruktúrájára kiszámolja a p értékét. Az így létrehozott új változó már alkalmas arra, hogy a prediktív modell lehetséges magyarázóváltozójaként a modellépítés során felhasználjuk.

A 3. ábra példát mutat egy konkrét folytonos változó esetében az algoritmus végeredményére.[11]

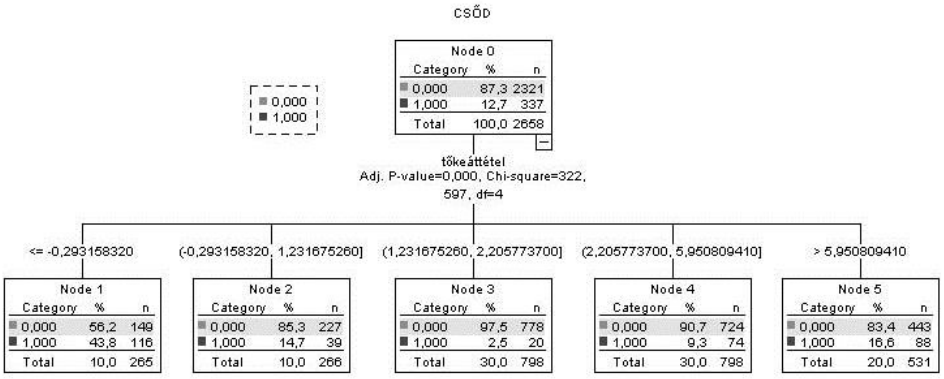
[8] Az algoritmus kategóriaváltozók esetén is alkalmas a változó meglevő kategóriáinak optimáló kritériumok mellett történő összevonására, ezáltal csökkentve a kategóriák számát.

[9] Folytonos változók esetén az algoritmus alapbeállításként a $K = 10$ értéket ajánlja fel. Ebben az esetben a változó decilisei jelennek meg, mint kategóriák.

[10] Logisztikus regresszió esetében $m = 2$.

[11] A vizsgálathoz magyarországi közepes méretű vállalatok pénzügyi mutatóit és esetleges csődeseményeit tartalmazó 3 800 megfigyelést tartalmazó adatbázist használtuk. Az esetek elegendően nagy száma itt lehetővé tette, hogy az adatbázist 70:30 arányban modellfejlesztésre és az eredmények tesztelésére alkalmas részekre osszuk fel. Az ábra az SPSS answer tree CHAID programmoduljának segítségével készült.

3. ábra. CHAID-alapú kategorizálás



A kategorizálandó pénzügyi mutató: *tőkáttétel = összes eszköz/saját tőke*. Az ábra téglalapjaiban láthatók a célváltozó (Csőd) lehetséges értékei (0,1) szerinti megoszlások a pénzügyi mutató kategóriáinak megfelelően. A legfelső téglalapban (a fa csúcsán) látható, hogy a fejlesztési adatbázis 2 321 fizetőképes és 337 csődös vállalatot tartalmazott. Az ábráról könnyen leolvasható a kategorizálási szabály. Az új öt kategóriával rendelkező mutató kategóriahatárai rendre a következők:

– 0,29316 az első és második kategória,
– 1,23167 a második és harmadik kategória,
– 2,20577 a harmadik és negyedik kategória,
– 5,95081 a negyedik és ötödik kategória esetén.

Az ábrát tovább tanulmányozva kibontakozik az eredeti folytonos változó és a csődesemény várható bekövetkezésének fordított, közel U alakú kapcsolata (lásd az egyes alsó téglalapokban szereplő relatív gyakoriságokat). Ennek megfelelően a mutatót eredeti folytonos alakjában szerepeltetve a logisztikus regressziós modellben a *tőkeáttétel* nem bizonyul szignifikánsnak (p = 0,126) a szokásos szinteken, amint azt a 4. táblázat mutatja.

4. táblázat

Paraméterbecslés a tőkeáttétel szerint

Változó	β	Standard hiba	Wald-statisztika értéke	Szabadságfok	Szignifikancia-szint	Exp(β)
Tőkeáttétel	0,005	0,003	2,345	1	0,126	1,005
Konstans	–1,961	0,062	986,363	1	0,000	0,141

Nézzük meg ezek után a CHAID-alapú kategorizálással előállított változóval (CHAIDKAT) készült logisztikus regressziós modell paraméterbecslését.

5. táblázat

Paraméterbecslés a kategorizált CHAID- (CHAIDKAT-) változó bevonásakor

Változó	β	Standard hiba	Wald-statisztika értéke	Szabadságfok	Szignifikancia-szint	Exp(β)
CHAIDKAT			231,308	4	0,000	
CHAIDKAT(1)	1,357	0,170	63,457	1	0,000	3,885
CHAIDKAT(2)	–0,115	0,207	0,309	1	0,008	0,891
CHAIDKAT(3)	–2,045	0,255	64,417	1	0,000	0,129
CHAIDKAT(4)	–0,666	0,169	15,550	1	0,000	0,514
Konstans	–1,616	0,117	191,778	1	0,000	0,199

Az új kategorizált mutató minden szokásos szinten szignifikánsnak bizonyul.

5. Alternatív kategorizálások vizsgálata

Felmerülhet a kérdés, hogy vajon más kategorizálási logikával milyen eredményt lehet elérni. Ehhez nézzük meg, mi történik akkor, ha más, könnyen kialakítható kategorizálási logika szerint alakítunk ki kategóriahatárokat. Ehhez induljunk ki a folytonos alapváltozó eloszlását leíró fontosabb statisztikákból. A 6. táblázat tartalmazza az tőkáttétel mutatót jellemző leíró statisztikákat:

6. táblázat

A tőkáttétel mutató statisztikái

N	
Érvényben levő	2658
Hiányzó	0
Átlag	5,5495
Median	2,2068
Tapasztalati szórás	15,0414
Variancia	226,2444
Ferdeség	4,2978
Ferdeség standard hibája	0,0475
Kurtózis	23,1797
Kurtózis standard hibája	0,0949
Mintaterjedelem	127,8471
Minimum	–25,0491
Maximum	102,7980
Percentilis	
10	–0,2870
20	1,2318
25	1,3392
30	1,4837
40	1,7703
50	2,2068
60	2,8255
70	3,8971
75	4,6157
80	5,9510
90	11,7640

Az alapstatisztikák felhasználásával definiáljunk további négyféle alternatív kategorizálási változatot a következő módon:

– A 20, 40, 60, 80 percentilisek (kvintilisek), mint kategóriahatárok által meghatározott öt kategóriájú változó (ÖTKAT).

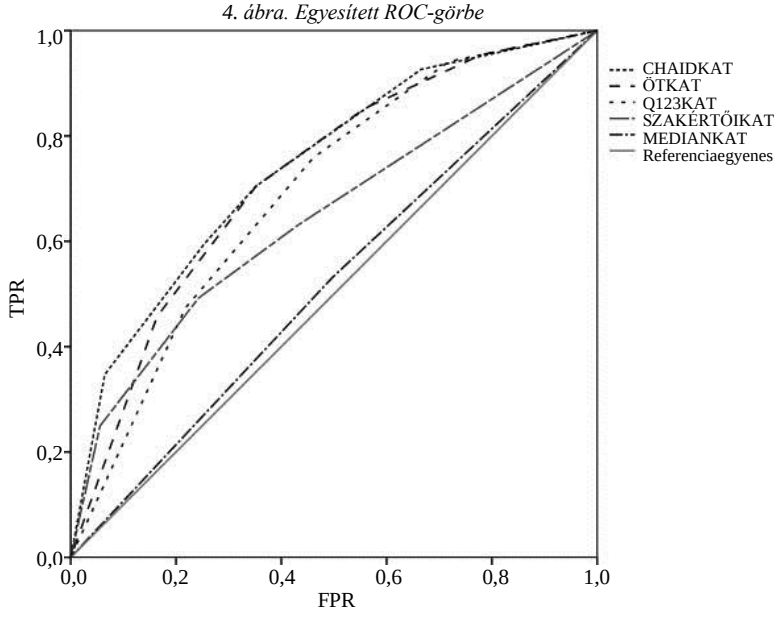
– A három kvartilis segítségével kialakított négykategóriás változó (Q123KAT).

– Az eloszlást valamilyen szakértői megfontolás alapján önkényesen négy részre felosztó pontok[12] segítségével kijelölt négykategóriájú változó (SZAKÉRTŐIKAT).

– Az eloszlás mediánja által definiált bináris változó (MEDIANKAT).

A vizsgálatunk célja az alternatív kategorizálási módszerek által előállított változók prediktív erejének értékelése az CHAID-alapú kategorizálási módszer eredményeképpen előálló CHAIDKAT-változóval szemben. A feladat végrehajtására egyváltozós logisztikus regressziók futtatására kerül sor az egyes változókon, majd a leválasztott tesztadatbázison ROC-görbével és Gini-együtthatóval értékeljük a modellek illeszkedését.[13] A 4. ábra egyesítve tartalmazza az öt alternatív modell ROC-görbéjét.

4. ábra. Egyesített ROC-görbe



[12] Az osztópontok rendre: 0, 1, 3, 6.

[13] Az alternatív kategorizálások paraméterbecslései a Függelékben találhatók. Mivel a modellek illeszkedésének mérése nem azon az adatbázison történik, amelyiken a regressziós modell kialakításra került, ezért ebben az esetben a Nagelkerke R^2 számítása hagyományos (szoftverek által támogatott) módon nem lehetséges.

Az ábrán látható, hogy a legnagyobb területet befoglaló görbe, ezáltal a legnagyobb prediktív erő az optimális kategorizálási eljáráshoz tartozó CHAIDKAT-változóhoz tartozik. Az egyes modellekhez tartozó Gini-értékeket a 7. táblázatban foglaljuk össze.

7. táblázat

A Gini-értékek összefoglaló táblázata
(százalék)

Mutató	CHAIDKAT	ÖTKAT	Q123KAT	SZAKÉRTŐIKAT	MEDIANKAT
Gini	48,4	44,5	38,7	29,2	3,4

A független tesztmintán történő visszamérés eredményeképpen, várakozásunkkal összhangban, a CHAID-alapú kategorizálás segítségével kialakított kategóriaváltozó mutatta a legjobb illeszkedést (Gini = 48,4 százalék) az alkalmazott alternatív módszerekkel szemben.

A CHAID-alapú kategorizálás további előnye, hogy segítségével azonnali képet kaphatunk a folytonos változó és a célváltozó közötti kapcsolatról, annak monoton vagy nem monoton jellegéről, valamint arról, hogy CHAID-alapú kategorizálást alkalmazva milyen erősségű kapcsolat várható a célváltozó vonatkozásában (lásd a 3. ábrán a *p*-értéket). Alacsony EPV-érték esetén az előzetes változószelekció során az egyes magyarázóváltozókat jellemző empirikus szignifikanciák (*p*-értékek) CHAID-del is számolhatók a bemutatott khi-négyzet-alapú technika gyors és kényelmes alternatívájaként.

6. További következtetések, kutatási irányok

A bemutatott példák mind valós adatszerkezeteken készültek. Bár a példák kialakítása során regressziós eszközként a logisztikus regressziós modellt alkalmaztuk, mindazon által intuitív megfontolások alapján azt gondoljuk, hogy a bemutatottak nagymértékben általánosíthatók minden olyan regressziós modell típusnál, ahol az illesztendő függvény (link) monoton jellegű. Ennek igazolása további kutatást igényel. Jövőbeli vizsgálatok tárgya lehet az is, hogy vajon miképpen változik meg a kategorizált változók illeszkedése, ha CHAID helyett más algoritmusokat alkalmazunk a kategóriahatárok kijelölésére.

Függelék

Teljesítményértékelés ROC-görbével és Gini-együtthatóval

A prediktív modellek globális illeszkedésének jellemzésére a gyakorlatban széleskörűen alkalmazott ROC-görbe és a Gini-mutató szoros rokonságban áll az elemi statisztikában közismert Lorenz-görbével és az arra épülő koncentráció méréssel, mint ezt a továbbiakban látni fogjuk. A módszer jobb megértéséhez felhasználjuk a klasszifikációs célú prediktív modellek másik népszerű teljesítménymérési eszközének, a konfúziós mátrixnak (más néven klasszifikációs táblának) a fogalmát.

Az egyszerűség kedvéért tegyük fel, hogy elkészült modellünket hitelezési döntéshez szeretnénk felhasználni, ahol azt kívánjuk előre jelezni, hogy egy adott vállalat fizetőképesség szempontjából kialakított „túlélő" és „csődös" kategória melyikébe fog kerülni. A döntés egy adott vállalat esetében úgy történik, hogy a vállalat modell által szolgáltatott függvényértékét viszonyítjuk egy előre definiált döntési küszöbértékhez (cut-off point). Amennyiben ez az érték nagyobb egyenlő, mint a cut-off, akkor csődősnek tekintjük, egyéb esetekben pedig túlélőnek. Az adott mintán és cut-off mellett az összes értékelendő vállalaton elvégezve a leírt besorolást, majd összevetve a valóságos kategóriába való tartozással kapjuk a konfúziós mátrixot.

F1. ábra. Konfúziós mátrix

		Tényleges csoport	
		csődös	túlélő
Elrőrejelzett csoport	csődös	TP helyes besorolás (csőd)	FP elsőfajú hiba
	túlélő	FN másodfajú hiba	TN helyes besorolás (túlélő)

Megjegyzés. TP = true positive (igaz pozitív), TN = true negative (igaz negatív), FP = false positive (hamis pozitív), *FN* = false negative (hamis negatív).

Az F1. ábrán látható, hogy a helyes besorolások száma TP + TN, míg a téves besorolásoké FP + FN. Elsőfajú hibának azt tekintjük, amikor egy ténylegesen túlélő vállalatot tévesen csődösnek minősítünk (FP), míg másodfajú hiba esetén egy csődös vállalatot minősítünk tévesen túlélőnek (amennyiben H_0: a vállalat túlélő). A mátrix elemeiből számos mutatószám képezhető, melyekből fontosságuk miatt kettőt emelnénk ki. Az első az ún. TPR (true positive rate – igaz pozitív arány), melyet találati érzékenységnek (sensitivity) is neveznek:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \qquad /1/$$

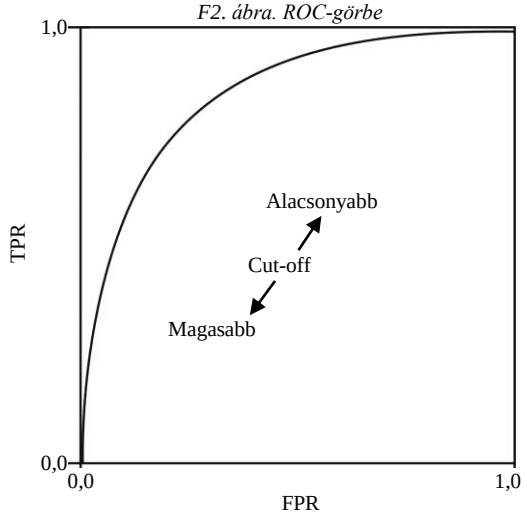
A kifejezés megmutatja, hogy adott cut-off mellett modellünk hány százalékát képes helyesen felismerni (besorolni) a csődős vállalkozásoknak. A másik fontos mutatószám az FPR (false positive rate – hamis pozitív arány), melyet a következőképpen írhatunk fel

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}, \qquad /2/$$

és megmutatja, hogy adott cut-off mellett a túlélő vállalatok hány százaléka lesz tévesen csődösnek minősítve.

Ezek után hozzáfoghatunk a ROC-görbe definiálásához.[14] A módszer ötlete abból a felismerésből ered, hogy a konfúziós mátrix elemei egy adott cut-off mellett értelmezhetők, a cut-off-érték megváltoztatásával a mátrix elemei is új értékeket vesznek fel. Úgyis mondhatnánk, hogy ahány cut-off-értéket veszünk, annyi konfúziós mátrixot kapunk. A ROC-görbe egy ábrába sűrítve mutatja meg számunkra, hogyan változik a TPR- és FPR-mutató, amint a cut-off-értéket nulla és egy között mozgatjuk.

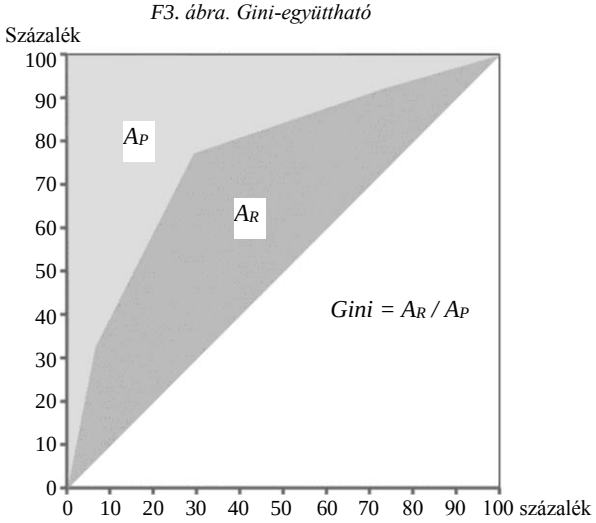
F2. ábra. ROC-görbe



Az F2. ábrát tanulmányozva könnyű belátni, hogy a tökéletesen klasszifikáló (előrejelző) modell esetében a ROC-görbe az ábra téglalapjának bal felső sarkába simulva végigköveti a bal oldali függőleges és a felső vízszintes szakaszt, míg a klasszifikáló erővel egyáltalán nem rendelkező modell görbéje a téglalap (0,0) pontból kiinduló átlójára illeszkedő egyenes. A valóságos modellek jellemző ROC-görbék természetesen valahol a két szélsőséges helyzet között szoktak elhelyezkedni. A ROC-görbe segítségével az elkészült modellről egy globális képet kapunk, hiszen bármely cut-off-érték esetén az ábráról könnyen leolvasható és számolható a konfúziós mátrix valamennyi eleme. Felmerül a kérdés, hogy a lineáris regressziónál megszokott R^2 mintájára alkotható-e egy olyan mutatószám a ROC-görbe segítségével, melynél a tökéletesen illeszkedő modell esetén a mutató értéke egy, illetve az előrejelző erővel nem bíró modell esetén ez az érték nulla. A leginkább elterjedt megoldás az ún. Gini-együttható használata, melynél a görbe és az átló közötti területet (A_R) hasonlítjuk a bennfoglaló derékszögű háromszög területéhez (A_P) az F3. ábrán látható módon.

[14] A ROC-görbét először a második világháborúban használták radarjelzések elemzésére.

F3. ábra. Gini-együttható



A területhányados alapján könnyen belátható, hogy a Gini-mutató értéke 1 a perfekt-, 0 a véletlenszerűen előrejelző modell esetén.[15] A Gini-együttható széles körben elterjedt mérőszáma a modell illeszkedésének. Ennek az oka, hogy a mutató minden további nélkül számolható a fejlesztési mintától különböző tesztmintára is. Ezzel szemben a többi, hagyományos illeszkedési mérőszámot a programcsomagok alapértelmezésben csak a fejlesztő mintára számolják ki. Ezért a grafikus tulajdonságai miatt könnyen interpretálható ROC-görbe és a belőle számolható Gini-együttható a gyakorlati modellezés egyik legkedveltebb mérőszáma.

Alternatív modellek paraméterbecslései

Változó	β	Standard hiba	Wald-statisztika értéke	Szabadságfok	Szignifikanciaszint	Exp(β)
ÖTKAT			171,592	4	0,000	
ÖTKAT(1)	0,730	0,151	23,447	1	0,000	2,075
ÖTKAT(2)	–2,071	0,304	46,370	1	0,000	0,126
ÖTKAT(3)	–1,201	0,221	29,477	1	0,000	0,301
ÖTKAT(4)	–0,628	0,188	11,163	1	0,001	0,534
Konstans	–1,616	0,117	191,778	1	0,000	0,199
Q123KAT			117,961	3	0,000	
Q123KAT(1)	0,563	0,141	15,863	1	0,000	1,755
Q123KAT(2)	–1,985	0,275	52,022	1	0,000	0,137
Q123KAT(3)	–0,575	0,172	11,105	1	0,001	0,563
Konstans	–1,718	0,108	252,807	1	0,000	0,179

(*A táblázat folytatása a következő oldalon*.)

[15] Tökéletes modell esetén $A_R = A_P$, a véletlenszerűen előrejelző modell esetén $A_R = 0$.

(*Folytatás.*)

Változó	β	Standard hiba	Wald-statisztika értéke	Szabadságfok	Szignifikanciaszint	Exp(β)
SZAKÉRTŐIKAT			149,628	3	0,000	
SZAKÉRTŐIKAT(1)	1,219	0,182	44,747	1	0,000	3,385
SZAKÉRTŐIKAT(2)	–0,774	0,152	25,884	1	0,000	0,461
SZAKÉRTŐIKAT(3)	–0,609	0,193	9,897	1	0,002	0,544
Konstans	–1,632	0,118	191,705	1	0,000	0,195
MEDIANKAT(1)	0,092	0,117	0,620	1	0,431	1,096
Konstans	–1,976	0,084	555,726	1	0,000	0,139

Irodalom

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

ORAVECZ B. [2008]: Hiányzó adatok és kezelésük a statisztikai elemzésekben. *Statisztikai Szemle.* 86. évf. 4. sz. 366–384. old.

PEDUZZI, P. – CONCATO, J. – KEMPER, E. – HOLFORD, T. R. – FEINSTEIN, A. R. [1996]: A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*. Vol. 49. No. 12. pp. 1372–1379. DOI: http://dx.doi.org/10.1016/S0895-4356(96)00236-3

YITZHAKI, S. – SCHECHTMAN, E. [2012]: Identifying monotonic and non-monotonic relationships. *Economics Letters*. Vol. 116. Issue 1. pp. 23–25. DOI: http://dx.doi.org/10.1016/j.econlet.2011.12.123

VARGHA A. – BERGMAN, L. R. [2012]: A method to maximize the information of a continuous variable in relation to a dichotomous grouping variable: cutpoint analysis. *Hungarian Statistical Review.* Special number 16. pp. 101–122.

Summary

When complex databases are analysed, the relatively low number of observations compared to the number of possible explanatory variables is a common problem, which may necessitate the preliminary selection of explanatory variables. As a solution for this problem, the author presents a preselection technique based on chi-square statistics and draws attention to the associated interpretation risks.

It is also typical that one should decide about the preliminary categorization of continuous variables that can increase the predictive power of the model even if the target variable and the original continuous variable are monotonically related. The study presents that the well-known CHAID algorithm can lead to a continuous-variable-related category structure, which provides better fit than other simple categorization rules.