A bizonyítéksúly és az információérték alkalmazása prediktív modellekben a folytonos változók kategorizálására

Nagy Gábor Szabolcs
Pécsi Tudományegyetem

A TANULMÁNY CÉLJA

A tanulmány elsődleges célja, hogy megismertesse a bizonyítéksúly (weight of evidence), és az információérték (information value) fogalmát, elméleti hátterét, alkalmazásának lehetőségeit, emellett népszerűsítse az R program használatát.

ALKALMAZOTT MÓDSZERTAN

A tanulmány elsődlegesen a bizonyítéksúlynak és az információértéknek a folytonos változók kategorizálására történő alkalmazásával foglalkozik. A bizonyítéksúly elméleti hátteréről, illetve az alkalmazási lehetőségekről a rendelkezésre álló szakirodalom alapján ad áttekintést a szerző. A változók kategorizálásának hatását egy olyan logisztikus regressziós modellben vizsgálja a tanulmány, amely a tőzsdei árfolyamok alakulásának trendjét jelzi előre.

LEGFONTOSABB EREDMÉNYEK

A WOE alapján történő kategorizálás viszonylag stabil, viszont a változók információértéke, illetve az információérték alapján meghatározott sorrendje a minta nagyságától függően eltérő. Az ún. információs kritériumok értékére is hatással van a kategorizálás, de nem minden esetben az a kategória változó eredményezi a legjobb AIC vagy BIC értéket, amelyik a WOE és az IV alapján optimális.

GYAKORLATI JAVASLATOK

A módszer nemcsak a folytonos változók kategorizálására alkalmazható, hanem exploratív eszközként is használható. Akkor is érdemes kiszámítani a bizonyítéksúlyokat és az információértéket, ha nem az optimális megoldást keressük, hanem az üzleti logika alapján szeretnénk kategorizálni, de nem egyértelmű, hogy két kategorizálás közül melyik a jobb. Emellett a WOE olyan esetekben is segítségünkre van, ha egy változó alapján szeretnénk szegmentálni az állományt. A logisztikus regressziós modellek kiértékelésénél mindenképpen több szempontot kell figyelembe venni, mivel nem minden esetben egyértelmű a változtatások hatása. Érdemes lenne alaposabban vizsgálni az információs kritériumok közötti összefüggéseket.

Kulcsszavak: bizonyítéksúly, információérték, uplift modell, logisztikus regresszió

BEVEZETÉS

Az üzleti életben használt prediktív modelleknél gyakran felmerül az igény, hogy a különböző magyarázó változók értékkészletét csoportosítsuk[1]. Csak akkor várhatjuk, hogy a modellünk megfelelően illeszkedjék, ha monoton kapcsolat van a magyarázóváltozók és a célváltozó között. Mivel ez nem mindig teljesül, a magyarázóerő növelése érdekében a folytonos változókat gyakran kategóriaváltozóvá alakítjuk át. Emellett az is előfordulhat, hogy egy magyarázóváltozó monoton kapcsolatot mutat a célváltozóval, de nem szignifikáns. Ebben az esetben megfelelő kategorizálással egy szignifikáns változót nyerhetünk (Hámori 2014, 2016). Ez az átalakítás több szempontból is előnyös: amellett, hogy megkönnyíti az összefüggések feltárását, csökkenti túltanulás veszélyét és a „zaj" hatását, növeli a modell pontosságát, és valamilyen szinten kezeli a hiányzó értékekkel, vagy az extrém szélsőértékekkel kapcsolatos problémákat. Az említett előnyök mellett ugyanakkor a hátránya is megvan a kategorizálásnak: a lehetséges értékek számának csökkenése mindig valamekkora információveszteséggel jár[2], ami egyrészt a modellek magyarázó erejének gyengüléséhez vezethet, másrészt bizonyos esetekben (például egy hisztogramnál) teljesen félrevezető lehet.

A folytonos változók értékkészletének csoportosítására többféle módszer létezik, az egyik legegyszerűbb megoldás, ha egyenlő hosszúságú sávokat képezünk (leggyakrabban pl. az életkornál vagy a jövedelemnél alkalmazzuk ezt a módszert), vagy ha úgy határozzuk meg a csoportokat, hogy mindegyikben azonos legyen az egyedek száma (pl. kvartilisek, percentilisek) [3]. Az első esetben gyakran előfordul, hogy az egyes csoportokban túl kevés az elemszám ahhoz, hogy bármiféle kalkulációt végezzünk, a kvantilisekkel pedig az a probléma, hogy gyakran nincs éles határ a különböző csoportok között, így ez a módszer alkalmatlan arra, hogy maximalizálja a kategóriák közötti különbségeket.

A gépi tanulási módszerek olyan megoldásokat is kínálnak, melynek segítségével úgy határozhatjuk meg a kategóriák közötti osztópontokat, hogy a célváltozó értékeinek tekintetében a lehető legnagyobb legyen a különbség a megképzett csoportok között. A bináris célváltozójú prediktív modelleknél alkalmazható például a bizonyítéksúly és az információérték. A módszer lényege, hogy minden egyes csoporthoz hozzárendeljük a bizonyítéksúlyt, s a csoportokat úgy képezzük meg, hogy a súlyok tekintetében a lehető legnagyobb legyen közöttük a különbség. A bizonyítéksúly a kategorizálás „erősségét" méri, abból a szempontból, hogy az adott csoportosítás mennyire képes elkülöníteni a bináris célváltozó értékeit. Az információérték az egyes csoportokhoz rendelt súlyok abszolút értékének az összege, s azt fejezi ki, hogy az adott kategóriaváltozó milyen „információ-értékkel" bír a bináris célváltozó szempontjából.

Az alábbi tanulmány két részből áll. Az első rész rövid áttekintést ad említett módszerek elméleti hátteréről, a jelenlegi alkalmazásokról, illetve a további alkalmazási lehetőségekről. A második rész egy példán keresztül mutatja be a bizonyítéksúly és az információérték használatát, illetve néhány gyakorlati szempontból érdekes kérdéssel foglalkozik. Ennek a résznek az is célja, hogy népszerűsítse a nyílt forráskódú statisztikai programcsomag, az R használatát, amelyben egyébként több csomagot is találunk a bizonyítéksúly és az információérték meghatározására.

---

[1] Az angol nyelvű szakirodalom több különböző kifejezést is használ a változók kategorizálására: coding and binning of predictors, chopping data, discretization, transformation of categorical predictors.

[2] A szakirodalomban egyébként komoly vita tárgya a folytonos változók kategorizálása, az előnyök mellett többen a hátrányokra is felhívják a figyelmet. Harrel (2015) 13, Dinerot (1996) pedig 7 érvet sorol fel amellett, hogy miért nem célszerű kategorizálni a folytonos változókat.

[3] Zeng (2014) két másik módszert is megemlít az egyenlő szélességű és egyenlő elemszámú csoportok létrehozása mellett. Az optimális kategorizálás (optimal binning) során először nagyobb számú, egyenlő szélességű csoportokba soroljuk az állományt. Ezeket a csoportokat úgy kezeljük, mint egy nominális változó kategóriáit, és besoroljuk őket az előre megadott számú szegmensek valamelyikébe. A több intervallumos módszer (multi-interval discretization binning) az entrópia minimalizálásának elvén alapul. Több intervallumot hozunk létre, majd ezekből rekurzív módon meghatározzuk a legjobb csoportosítást.

A BIZONYÍTÉKSÚLY ÉS AZ INFORMÁCIÓÉRTÉK ELMÉLETI HÁTTERE

Mielőtt belemennénk az alkalmazási lehetőségek tárgyalásába, érdemes kicsit közelebbről megismerkedni a módszer elméleti hátterével. Néhányan az információelméletből ismert entrópiával, illetve a kölcsönös információval (mutual information) hozzák összefüggésbe a bizonyítéksúly és az információérték fogalmát (Larsen 2015, Lin et al. 2015), de nem fejtik ki részletesen, hogy pontosan miben is áll ez a kapcsolatot. Valójában inkább csak egyfajta hasonlóság van a fogalmak között, mivel a kölcsönös információ is azt méri, hogy egy adott változó eloszlásának ismerete mennyire hasznos egy másik változó értékének az előrejelzésében.

A bizonyítéksúly elméleti gyökerei inkább Alan Turing munkásságára vezethetők vissza, bár ő maga még nem ezt a kifejezést használta[4]. A szintén kiváló statisztikus Irwing J. Good, aki Turing asszisztense volt a Bletchley Parkban a világháború alatt, a Biometrika című folyóiratban 1979-ben megjelent tanulmányában részletesen kifejti, hogy miként járult hozzá Turing a bayesi elmélet fejlődéséhez (Good 1979). Mivel a dokumentumok sokáig titkosítva voltak, csak az utóbbi időben vált ismertté, hogy az ENIGMA által küldött kódolt üzenetek feltörésére használt Banburismus algoritmus a Turing által kidolgozott hipotézisvizsgálaton alapult. Turing megközelítésében az volt az igazi újdonság, hogy nem közvetlenül a valószínűségeket használta, hanem az ún. esélyhányadosokat. Tulajdonképpen innen ered az az összefüggés is, amelyet ma bizonyítéksúlyként ismerünk (Larrañaga and Bielza 2012).

A bizonyítéksúly elméleti hátterének bemutatásához Good egy későbbi írását használom fel (Good 1985: 250-2), némileg átszerkesztve, illetve kiegészítve, hogy a téma szempontjából fontos összefüggések nagyobb hangsúlyt kapjanak. Jelöljön *H* egy tetszőleges hipotézist, *E* pedig egy empirikus bizonyítékot, ami lehet akár egy esemény bekövetkezése, vagy egy kísérlet eredménye is. A bayesi statisztikából jól ismert az alábbi összefüggés[5]:

$$P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E)} \quad (1)$$

ahol *P(H)* a *H* hipotézis valószínűségére vonatkozó kezdeti (priori) vélekedés, *P(E)* az *E* bizonyíték bekövetkezésének valószínűsége, a *P(E|H)* feltételes valószínűség azt jelöli, hogy a *H* hipotézis fennállása esetén milyen valószínűséggel kapjuk meg az *E* bizonyítékot, végül a *P(H|E)* az utólagos (posteriori) valószínűsége annak, hogy a bizonyíték megtörténte után a H hipotézist fenntartjuk. Az (1) összefüggésnek két különböző olvasata is van (Hunyadi 2011): Az egyik értelmezés szerint valamilyen kezdeti tudásból indulunk ki, majd további információkat, tapasztalatokat, bizonyítékokat gyűjtünk, melyek ismeretében végül a tudás egy magasabb szintjére jutunk. A másik értelmezés szerint bizonytalan környezetben kell döntést hoznunk arról, hogy a *H* hipotézisünk igaz, vagy sem. A kapott bizonyítékok alapján felülvizsgáljuk eredeti állításunkat, így végső döntésünket a kezdeti vélekedésünk, a tapasztalatok, illetve ezek valószínűségei határozzák meg.

Jelölje *W(H:E)* a bizonyítéksúlyt (weight of evidence), ami azt fejezi ki, hogy *E* milyen mértékben szolgáltat bizonyítékot a H hipotézisre. Teljesen logikus, ha azt feltételezzük, hogy *W(E:H)* valamilyen módon a *P(E|H)* és a P(H|E) feltételes valószínűségek függvénye, azaz

$$W(H:E) = f[P(E|H), P(E|\bar{H})] \quad (2)$$

A bizonyítéksúly, és a P(H) kezdeti (priori) vélekedés együttesen határozza meg a *P(H|E)* utólagos (posteriori) valószínűséget:

$$P(H|E) = g[f[P(E|H), P(E|\bar{H})], P(H)] \quad (3)$$

A könnyebb áttekinthetőség kedvéért vezessük be az alábbi jelöléseket: *P(H)=x*, *P(E)=y*, illetve *P(H|E)=z*. Felhasználva a bayesi összefüggéseket, a (3) kifejezés a következőképpen írható fel:

$$z = g\left[f\left[\frac{zy}{x}, \frac{y(1-z)}{1-x}\right], x\right] \quad (4)$$

[4] A weight of evidence mellett a szakirodalomban előfordulnak még a „degree of corrobaton" és a „degree of confirmation" kifejezések is. Turing a 'score' és a 'decibannage' mellett a „factor in favour of h" kifejezést használta, majd Good a bayesi gyökerekre utalva „Bayes factor in favour of h" néven hivatkozott rá (Gillies 1990).

[5] A dolgozat keretei nem teszik lehetővé, hogy részletesen foglalkozzunk a hagyományos statisztika és a bayesi statisztika közötti különbségekkel. Magyar nyelven Hunyadi (2011) ad nagyon jó áttekintést a témáról.

Kimutatható, hogy f a $P(E|H)/P(E|\bar{H})$ hányados monoton függvénye (Good, 1968, p. 141). Ha H $\bar{H}$ és statisztikai hipotéziseket jelöl, E pedig az ezek vizsgálatára vonatkozó kísérlet eredménye, akkor $P(E|H)/P(E|\bar{H})$ az ún. likelihood hányados, vagy egész pontosan a bayesi likelihoodok hányadosa.

A bayesi statisztikában a valószínűség és a likelihood mellett egy másik fontos fogalom az esélyhányados (odds), amely hasonlóan a valószínűségekhez a priori és a posteriori is meghatározható. A kezdeti esélyhányados $O(H) = P(H)/P(\bar{H}) = P(H)/(1 - P(H))$, a végső esélyhányados pedig $O(H|E) = P(H|E)/P(\bar{H}|E) = P(H|E)/(1 - P(H|E))$. Felhasználva a jól ismert bayesi összefüggéseket, $P(E|H)/P(E|\bar{H})$ a hányados a következőképpen is felírható:

$$W(H:E) = \frac{P(E|H)}{P(E|\bar{H})} = \frac{P(H|E) \cdot \frac{P(E)}{P(H)}}{P(\bar{H}|E) \cdot \frac{P(E)}{P(\bar{H})}} = \frac{O(H|E)}{O(H)} \quad (5)$$

Az (5) kifejezést átrendezve a következő összefüggést kapjuk:

$$O(H|E) = \frac{P(E|H)}{P(E|\bar{H})} \cdot O(H) \quad (6)$$

ami azt jelenti, hogy az a posteriori esélyhányados a likelihood hányados és az a priori esélyhányados szorzata.

Bár Good egyértelműen bayesi gyökereket hangsúlyozza, Gillies (1990) szerint a bizonyítéksúly inkább a Popper által használt korroboráció fogalmához áll közelebb (Popper 1935). Popper szerint a tapasztalat nem nyújthat tudást, a kísérletekből nem feltétlenül következik, hogy az elméletünk igaz, az igazolhatóság helyett a falszifikálhatóság a megfelelő tudományos kritérium. A tudományban nem létezik igazolás, csak korroboráció, ha egy elmélet kiállja a próbát, akkor Popper szóhasználatában „korroborálódik", ami kevesebb, mint az igazolás. Ez utóbbi esetben jó okunk van azt hinni, hogy az elmélet igaz, míg a korroboráció csak annyit jelent, hogy igaznak tartjuk az elméletet (Popper, 1935).

Ha *h* jelöli a hipotézist, *e* a bizonyítékot, *k* pedig a háttértudást, akkor az igazolás (support) a három paraméteres *S(h,e,k)* függvénnyel írható le, amely azt fejezi ki, hogy az *e* bizonyíték, milyen értékben járul hozzá a *k* tudáshoz, vagy másként fogalmazva milyen mértékben nő a *h* hipotézis korroborációja az *e* bizonyíték ismeretében. A korroboráció ezzel szemben a két paraméteres *C(h, e & k)* függvénnyel írható le, amely a *h* hipotézis korroborációjának új szintjét méri, miután a bizonyítékot „hozzáadjuk" a háttértudáshoz. Ez egyben kifejezi a *h* hipotézissel kapcsolatos új vélekedést is, a kezdeti tudás és az új bizonyíték ismeretében. A korroboráció (C) és az igazolás (S) a következőképpen viszonyul egymáshoz (Gillies 1990):

$$C(h, e \,\&\, k) = S(h, e, k) + C(h, k) \quad (7)$$

A bayesi gondolkodás szerint C = P, így (7) a következőképpen írható fel:

$$S(h, e, k) = P(h, e \,\&\, k) - P(h, k) \quad (8)$$

A Turing-Good-féle bizonyítéksúly viszont a következőképpen néz ki:

$$W(h, e, k) = \log P(e, h \,\&\, k) - logP(e, h \,\&\, k) \quad (9)$$

A (8) és a (9) kifejezéseket összevetve látható, hogy két különböző dologról van szó. A (9) inkább arra hasonlít, ahogy Popper a tesztek megbízhatóságát méri (Gillies 1990):

$$Q = P(e, h \,\&\, k) - P(e, k) \quad (10)$$

A (10) kifejezésben szereplő Q érték annálnagyobb, minél nagyobb *P(e, h & k)* és minél kisebb *P(e, k)*, ami azt jelenti, hogy az *e* bizonyíték nagyon valószínű a *h* hipotézis fennállása mellett, és nagyon valószínűtlen a *h* hipotézis fennállása nélkül.

Egy speciális esetben kimutatható, hogy a Turing-féle bizonyítéksúly megegyezik a Popper-féle Q értékkel. Mivel a tudomány területén gyakran általános hipotéziseket fogalmazunk meg, s Poppernél ezek nem verifikálhatók, csak falszifikálhatók, a hipotézis bekövetkezésének valószínűsége P(h)=0. Tegyük fel, hogy az a hipotézisünk, hogy minden holló fekete. Ha fogadnánk erre a hipotézisre, akkor soha nem nyernénk, mivel nem lehet teljes bizonyossággal bebizonyítani, hogy minden holló fekete. Elegendő lenne mindössze egyetlen fehér hollót találni ahhoz, hogy elveszítsük a fogadást. Ha tehát *P(h)=0*, akkor

$$P(e) = P(e, h)P(h) + P(e, \bar{h})P(\bar{h}) = P(e, \bar{h}) \quad (11)$$

A Turing-Good-féle bizonyítéksúly ebben az esetben

$$W = \log P(e, h) - \log P(e) \quad (12)$$

a Popper-féle Q érték pedig

$$Q = P(e,h) = P(e) \tag{13}$$

Látható, hogy (8) és (9) között az egyetlen különbség, hogy a Turing-Good-féle képletben a valószínűségek logaritmusa szerepel.

A WOE ÉS AZ IV ALKALMAZÁSÁNAK LEHETŐSÉGEI A STATISZTIKAI MODELLEZÉSBEN

1. Az üzleti alkalmazásokban elsősorban a folytonos változók kategorizálásának egyik lehetséges eszközeként használjuk a bizonyítéksúlyt és az információértéket. A szakirodalom ennek kapcsán elsősorban a scoring, és az uplift modellekről tesz említést.
2. Egy másik lehetséges alkalmazási terület, az ún. weights of evidence modeling, amikor nem csupán a változók kategorizálására használjuk a bizonyítéksúlyt, hanem kifejezetten a prediktív modellezésre. Erre az alkalmazásra más szakterületen, leginkább a geostatisztikában találunk példákat (Agterberg et al. 1993, Barbieri and Cambuli 2009, Hartley 2014). Üzleti területen is vannak persze hasonló módszertant használó modellek, amelyek odds modellként ismertek. A weights-of-evidence módszert először orvosi diagnózisra fejlesztették ki, majd a 80-as években kezdték el alkalmazni más területeken is (pl. mineral prospectivity modeling). A módszer matematikai hátterét Bonham-Carter (1994) és Carranza (2009) foglalja össze.
3. Végül a harmadik terület, a bizonyítéksúly exploratív eszközként történő alkalmazása (Larsen 2015), inkább csak nyomokban lelhető fel a szakirodalomban, külön kifejezetten ezzel foglalkozó tanulmány nem készült. Exploratív eszközként a bizonyítéksúly alkalmas lehet (Larsen 2015):

- annak vizsgálatára, hogy az egyes változók önmagukban miként járulnak hozzá a kimenethez;
- a változók közötti lineáris és nem lineáris kapcsolatok feltárására;
- a változók rangsorolására a magyarázóerő alapján;
- a bináris célváltozó és a prediktív változók közötti korreláció vizualizálására;
- egy folytonos és egy kategória változó erősségének összehasonlítására anélkül, hogy dummy változókat kellene létrehoznunk;
- a hiányzó értékekkel kapcsolatos problémák kezelésére anélkül, hogy ezeket más értékkel kellene helyettesítenünk;
- a hiányzó értékek magyarázó erejének kiértékelésére.

A folytonos változók kategorizálása a bizonyítéksúly és az információérték segítségével

Az üzleti életben használt prediktív modellek többsége valamilyen bináris célváltozóra épített klasszifikációs modell, aminek egyszerűen az az oka, hogy az megfigyelt jelenségek többnyire bináris természetűek, vagy legalábbis leegyszerűsíthetők egy bináris probléma szintjére. Leginkább olyan kérdésekre keressük a választ, hogy mely ügyfelek hitelképesek, kik morzsolódhatnak le egy adott időtávon, kiket érdemes megkeresni valamilyen kampány keretében, vagy kik mutatják a legnagyobb hajlandóságot egy adott termék megvásárlására. Az alkalmazott modellek különböző módszereket használnak, melyek közül a logisztikus regressziós modellek, a döntési fák és a neurális hálók a leginkább népszerűek.

A modellezés során gyakran felmerülő kérdés, hogy miként lehet megfelelően kategorizálni az egyes változókat, hogyan lehet például csoportosítani egy folytonos változó értékkészletét, majd több lehetséges csoportosítás közül hogyan lehet kiválasztani a modellezés szempontjából leginkább megfelelőt. A cél az, hogy végül egy olyan adatállományt állítsunk, amely nem veszít a magyarázó erejéből, és tartalmazza a célváltozó előrejelzése szempontjából legfontosabb változókat.

Folytonos változók esetén a legismertebb módszerek közé tartozik az egyenlő szélességű osztályközök létrehozása, illetve az egyenlő elemszámú osztályközök létrehozása, de gyakran előfordul, hogy egyszerűen az üzleti logika alapján kategorizálunk. A bizonyítéksúly felhasználásával történő kategorizálás az említett módszerekhez képest alapvetően abban más, hogy figyelembe veszi a célváltozó értékeinek a megoszlását, és olyan csoportokat ad eredményül, amelyek a lehető legnagyobb mértékben különböznek egymástól.

A kategorizálás igénye egyébként nemcsak a folytonos változóknál merül fel (pl. korsávok, jövedelemsávok meghatározása), gyakran előfordul,

hogy a nagy értékkészlettel rendelkező diszkrét változókat is átalakíthatunk (tipikus példa erre az irányítószámok, települések kistérségek vagy régiók szerinti kategorizálása).

a) A bizonyítéksúly meghatározása (WOE – Weight of Evidence)

Mivel a továbbiakban inkább a gyakorlatban felmerülő kérdésekkel foglalkozunk, célszerű bevezetni egy olyan jelölést, amely közelebb áll az üzleti életben használt modellekhez. Jelölje Y a bináris célváltozót , melynek Y = 1 értéke lesz most a számunkra fontos kimenetet, X pedig jelölje azt a magyarázó változót, amelyet kategória változóvá szeretnénk alakítani oly módon, hogy értékkészletét N különböző csoportba soroljuk. Az egyes csoportok bizonyítéksúlya a következőképpen határozható meg (Lee et al. 2013):

$$WOE_i = log\frac{P(X = x_i|Y = 1)}{P(X = x_i|Y = 0)} \qquad i = 1, 2, \ldots N \tag{14}$$

ahol $P(X=x_i|Y=1)$ annak valószínűsége, hogy $Y=1$ az x_i csoportba kerül. Ezt a gyakorlatban úgy határozzuk meg, hogy az adott csoportban előforduló $Y=1$ értékek számát elosztjuk a teljes mintában fellelhető $Y=1$ értékek összegével.

A bizonyítéksúlyt az esélyhányadosból (odds ratio) számítjuk, egyszerűen vesszük annak logaritmusát. Ha az esélyhányados értéke 1, akkor a bizonyítéksúly 0, ha a számláló kisebb, mint a nevező, akkor a WOE negatív értéket vesz fel, ellenkező esetben pedig pozitívot. Ez üzletileg úgy értelmezhető, hogy egy nullánál nagyobb WOE értéknél az adott csoportban nagyobb az esély a számlálóban megjelenő esemény bekövetkezésére, illetve minél nagyobb a WOE értéke, annál nagyobb az esély a célváltozó azon értékének bekövetkezésére, amelyet előre kívánunk jelezni.

A kategóriák meghatározásához nincsenek szigorú előírások, az alábbi hüvelykujjszabályokat érdemes használni (Siddiqi 2006: 80): (1) Ahhoz, hogy a változó kellően robusztus legyen, úgy kell meghatározni a határokat, hogy minden csoportba a vizsgált adatállomány legalább 5 százaléka kerüljön. (2) Arra kell törekedni, hogy lehetőleg minden csoportban legyen kedvező és kedvezőtlen kimenet, azaz a bináris változó minden egyes csoporton belül vegye fel a 0 és 1 értékeket is. Ha ez nem így lenne, akkor a 0-val való osztás miatt probléma lépne fel, bár ezt is lehet kezelni. (3) A hiányzó értékeket célszerű külön kategóriába tenni. (4) Ha két csoportnak azonos, vagy közeli WOE értéke van, akkor ezeket össze kell vonni, a cél az, hogy az egyes kategóriáknál különböző WOE értékeket kapjunk. (5) Az a jó, ha a kiszámított WOE értékek sorozata monoton növekvő vagy csökkenő, de a gyakorlatban nem mindig sikerül elérni ezt az ideális állapotot. Annak ellenőrzésére, hogy megfelelő-e a kategorizálás, különböző eszközöket alkalmazhatunk, de a legegyszerűbben egy ábra segítségével ellenőrizhetjük a WOE értékek monotonitását. (6) Minél nagyobb a WOE értékek közötti különbség, annál nagyobb a változó magyarázó ereje.

A kategória változóknál nem feltétlenül adódik első ránézésre, hogy miként lehet összevonni az egyes értékeket. Ezekben az esetekben az üzleti logikát, vagy a hétköznapi logikát is figyelembe kell venni. Ha például foglalkozásokat kategorizálunk, akkor nyilvánvaló nem célszerű egy csoportban tenni olyan foglalkozásokat, amelyek teljesen más ágazatban tevékenykednek, még akkor sem, ha alapból azonos WOE értéket számoltunk. A lényeg, hogy a kategória változók újracsoportosításánál is érdemes kiszámítani a bizonyítéksúlyt és az információértéket, és ezek alapján összehasonlítani a különböző csoportosításokat.

b) Az információérték meghatározása (IV – Information Value)

Arra a kérdésre, hogy több lehetséges kategorizálás közül melyiket célszerű választani, az információérték segítségével adhatunk választ. Azt ugyan nem mutatja meg, hogy a változók különböző kombinációinak milyen extra hozzáadott értéke van, vagy hogy milyen korreláció lehet az egyes változók között, de általánosságban mégis jó indikátornak tűnik ahhoz, hogy ez alapján hasonlítsuk össze egymással a magyarázó változókat.

Az információértéket kiszámolva eldönthetjük, hogy egyáltalán van-e értelme egy adott változó kategorizálásának, vagy hogy több kategorizálás közül melyiket érdemes választani. Az egyes kategóriákhoz rendelt bizonyítéksúlyok segítségével az információérték a következőképpen határozható meg:

$$IV = \sum_{i=1}^{N} (P(X = x_i|Y = 1) - P(X = x_i|Y = 0)) \cdot WOE_i \tag{15}$$

Az információérték az egyes csoportokhoz rendelt bizonyítéksúly értékek súlyozott összege, ahol a súlyok a WOE képletében számlálóként és nevezőként szereplő értékek közötti abszolút különbségek. Az IV értékeit a következőképpen szokás értékelni (Siddiqi 2006: 81):

1. táblázat:
Az információérték lehetséges értékeinek értelmezése

Érték	Értelmezés
0,02 alatt	a változó nem alkalmas magyarázó változónak
0,02 és 0,1 között	gyenge kapcsolat a magyarázó változó és a célváltozó között
0,1 és 0,3 között	közepesen erős kapcsolat a magyarázó változó és a célváltozó között
0,3 és 0,5 között	erős kapcsolat a magyarázó változó és a célváltozó között
0,5 felett	a túl magas érték valamilyen hibára is utalhat, célszerű újból ellenőrizni a számítást

Forrás: Siddiqi (2006)

Mivel az információérték az adott változón létrehozott csoportok számával együtt növekszik, ezért ha a kevesebb és a több csoport mellett is egyaránt monoton növekvő vagy csökkenő a WOE, akkor azt a kategorizálást érdemes választani, amelynél több csoport képződik.

c) Az információérték és a lift kapcsolata

A WOE-hez hasonlóan az ún. Lift érték is az egyes osztályok előrejelző képességét méri a célváltozóra vonatkozóan (Csicsman – Szabó Sipos 2011), egész pontosan azt adja meg, hogy egy adott változó egyes osztályaiban hányszor nagyobb a célesemény bekövetkezésének aránya, mint a teljes állományban. A WOE és a Lift közötti kapcsolat az alábbiak szerint határozható meg:

$$WoE = log\frac{1-p}{\frac{1}{lift}-p} \tag{16}$$

amelyből átrendezve a képletet a következőt kapjuk:

$$lift = \frac{1}{(1-p)e^{-WoE}+p} \tag{17}$$

ahol p = (azon objektumok száma, melyekre a célváltozó értéke pozitív) / (összes objektum száma). Értéke 0 <= p <= 1.

d) A WOE és IV módszer kiterjesztései

A WOE és az IV ebben a formában a bináris célváltozójú, hagyományos prediktív modelleknél használható. Létezik azonban a módszernek két további kiterjesztése is, az egyik alkalmassá teszi a folytonos – vagy legalább kettőnél több értéket felvevő – célváltozók melletti használatra (Lin & Hsie 2014, Lin 2015), a másik pedig az uplift modelleknél történő alkalmazásra (Lee *et al.* 2013, Larsen 2015).

Alkalmazás folytonos célváltozók esetén

Ha egy olyan célváltozónk van, amely a jó adósokat különböző kategóriákba rangsorolja, és csak egy érték jelöli a rossz adósokat, akkor a bizonyítéksúly és az információérték a következőképpen határozható meg (Lin & Hsie 2014, Lin 2015):

$$WOE_i = log\frac{P(X=x_i|Y=1)}{P(X=x_i)} \tag{18}$$

$$IV = \sum_{i=1}^{N}(P(X=x_i|Y=1) - P(X=x_i)) \cdot WOE_i \tag{19}$$

ahol $P(X=x_i)$ az egyes csoportok előfordulásának valószínűsége, azaz a csoport elemszámának és a minta elemszámának a hányadosa. Ebben az alternatív képletben nem a kedvezőtlen kimenetek megoszlásához viszonyítunk, hanem a vizsgált adatállomány megoszlásához.

Alkalmazás uplift modelleknél

A scoring modellek mellett a bizonyítéksúly alkalmazásának egy speciális területe a Customer Relationship Management által használt uplift modellezés[6]. Erről viszonylag kevés szó esik még a hazai szakirodalomban, de angol nyelven több tanulmány is elérhető (Rzepakowski & Jaroszewicz 2002, Lee et al. 2013, Strickland 2015). Általános probléma, hogy a magas marketing költségekhez viszonyítva a legtöbb kampányban alacsony a válaszadási hajlandóság, s ebből kifolyólag az értékesítés is. A marketing kampányok gondosabb előkészítésével, meg lehet határozni úgy a célcsoportot, hogy azok kerüljenek bele, akik nagyobb valószínűséggel fognak üzletet kötni. A modellek segítségével készült célcsoport leválogatások jobb eredményt hoznak, mint a véletlenszerű csoportmeghatározás.

A hagyományos prediktív modelleket használó kampányok célcsoportjába mindenki belekerül, aki a modell alapján vélhetően érdeklődik a termék iránt. Így a célcsoportban olyanok is lehetnek, akikkel a megkeresés nélkül is lehetne üzletet kötni, mert alapból érdeklődnek a termék iránt. Az ő esetükben a kampányra fordított összeg kidobott pénz. A kampány járulékos hatását mérő modellek (incremental response model) általában négy csoportba sorolják az ügyfeleket (Lee et al. 2013): (i) akik csak a marketing kampány hatására válaszolnak, de maguktól nem mutatnának hajlandóságot a válaszadásra, (ii) akik a marketing kampánytól függetlenül is válaszolnak, (iii) akik egyáltalán nem válaszolnak, függetlenül attól, hogy megkeressük-e őket vagy sem, végül (iv) akiknél a marketing kampány negatív hatást ér el, azaz még kevesebb hajlandóságot mutatnak a válaszadásra, mint a kampány nélkül. A modellek célja, hogy az (i) csoportba sorolható ügyfeleket azonosítsuk, ők lesznek azok, akik a célváltozóban szerepelnek. A (ii) és (iii) csoportba tartozó ügyfelek megkeresése felesleges pénzkidobás, a (iv) csoport esetében pedig kifejezetten negatív hatással jár a megkeresés. A modellek segítségével tehát egyrészt hatékonyabban lehet elkölteni a kampányokra rendelkezésre álló összegeket, másrészt mérsékelni lehet a marketing megkeresések negatív hatását.

A célcsoport és a kontrollcsoport szétválasztása rendszerint véletlenszerűen történik. A modellnek azokat az ügyfeleket kell azonosítania, akik kifejezetten a kampány hatására vásárolnak, de egyébként nem tennék. A kampány lift értéke a célcsoportra és a kontrollcsoportra kiszámított sikerességi mutatók különbsége, a kampány akkor tekinthető sikeresnek, ha a célcsoport felülteljesíti a kontrollcsoportot. Az uplift modellek pont a marketing kampányok hatékonyságának további növelését szolgálják, hogy minél nagyobb legyen a különbség a célcsoport és a kontrollcsoport eredménye között. Míg a hagyományos sales modellek annak feltételes valószínűségét próbálják meghatározni, hogy a célcsoporton belül milyen lesz a válaszadási arány P(response|treatment), addig az uplift modellek inkább a kampány hatásának növelésére irányulnak, ami a célcsoport válaszadási hajlandósága és a kontrollcsoport válaszadási hajlandósága közötti különbséggel mérhető: P(response|treatment) – P(response|no treatment) (Radcliffe 2007, Radcliffe and Surry 2011).

Az uplift modelleknél a WOE és az IV értékeket a célcsoportra és a kontrollcsoportra is meghatározzuk, majd kiszámítjuk ezek különbségét, az ún. nettó bizonyítéksúlyt (NWOE – net weight of evidence), majd a nettó információértéket (NIV – Net Information Value) (Lee et al. 2013, illetve Larsen 2015):

$$NWOE = log\frac{P(X = x_i|Y = 1)_T/P(X = x_i|Y = 0)_T}{P(X = x_i|Y = 1)_C/P(X = x_i|Y = 0)_C} \quad (20)$$

ahol a T index a célcsoportot (target group vagy treated group), a C index pedig a kontrollcsoportot (control group) jelöli; és a célcsoportra, illetve a kontrolcsoportra számított feltételes valószínűségek. AZ NWOE lényegében a célcsoport és a kontrolcsoport odds rátáit hasonlítja össze.

[6] Az uplift modeling mellett a szakirodalom használja még az incremental modeling, true lift modeling, vagy a net-lift modeling kifejezéseket is. (Larsen 2016) Az „uplift" kifejezés értelmezése egyébként nem teljesen egységes a szakirodalomban. Általában a kampány célcsoport (target group vagy treated group) és a véletlenszerűen elkülönített kontrollcsoport (control group) eredményessége közötti különbséget értjük alatta, ami a válaszadási rátákban (response rate) vagy a konverziós rátákban (sales conversion ratio) is mérhető. Ebben az értelmezésben a két csoport rátája közötti különbség, az „uplift" a marketing hozzáadott értéke. Más szerzők viszont egyszerűen csak „lift"-ként hivatkoznak az említett különbségre, s az „uplift" alatt a két csoport lift értéke közötti különbség növelésére irányuló törekvést értik („upping lift").

A nettó információérték (Net Informatkon Value) a következőképpen határozható meg:

$$\sum_{i=1}^{N}(P(X = x_i|Y = 1)_T P(X = x_i|Y = 0)_C - P(X = x_i|Y = 0)_T P(X = x_i|Y = 1)_C) \cdot NWOE_i \quad (21)$$

Ha a változók magyarázóereje lényegesen alacsonyabb a validálásra használt állományon, mint a tanuló állományon, akkor a változó nem elég robusztus. Ilyen esetekben az NIV számításánál egy „büntető" korrekciós tényezőt (penalty) kell alkalmazni, amely a tanuló állományra számított NWOE és a tesztelésre használt állományra számított NWOE különbségét méri.

$$\omega = |NWOE_{train} - NWOE_{valid}| \quad (22)$$

A korrekciós tényező a következőképpen határozható meg:

$$p = \sum_{i=1}^{N}(P(X = x_i|Y = 1)_T P(X = x_i|Y = 0)_C - P(X = x_i|Y = 0)_T P(X = x_i|Y = 1)_C) \cdot \omega \quad (23)$$

A (11) és (13) összefüggések felhasználásával a korrigált nettó információérték az alábbiak szerint adódik:

$$PNIV = NIV - p \quad (24)$$

A korrekciós tényezőnek fontos szerepe van az NIV értékelésében. Az NWOE értékek rendszerint jobban különböznek egymástól, mint a WOE értékek, mivel az NWOE „aggregálja" a célcsoport és a kontrollcsoport WOE értékeit is. A büntető faktorral korrigált PNIV robusztusabb mérőszám, mint a sima NIV.

R CSOMAGOK A WOE ÉS AZ IV MEGHATÁROZÁSÁRA

R csomagok a bizonyítéksúly és az információérték számítására

Mind az akadémiai szektorban, mind a versenyszférában egyre szélesebb körben elterjedt a nyílt forráskódú statisztikai program, az R, melyben több olyan csomagot is találunk, amelyek segítségével elvégezhetjük a változók kategorizálását, illetve meghatározhatjuk az információértéket. A programnak alapvetően két változata van, az egyszerűbb felületű R, illetve az RStudio.

2. táblázat
R csomagok a WOE és IV számítására

R Csomag	Leírás
Woe	Weight of Evidence és Information Value értékeket számol egy függő és egy független változóra. Szerző: Sudarson Mothilal Thoppay (2015) **Dokumentáció:** https://cran.r-project.org/web/packages/woe/woe.pdf https://cran.r-project.org/web/packages/woe/index.html
InformationValue	A WOE és IV számítás mellett több olyan függvényt is tartalmaz, amelyek segítségével elemezni lehet az összefüggéseket. Szerző: Selva Prabhakaran (2015) **Dokumentáció:** https://cran.r-project.org/web/packages/InformationValue/InformationValue.pdf https://cran.r-project.org/web/packages/InformationValue/index.html

Information	A WOE és IV értékek mellett NWOE és NIV számításra is alkalmas, így uplift modelleknél is kiválóan használható. Szerző: Kim Larsen (2016) Dokumentáció: https://cran.r-project.org/web/packages/Information/Information.pdf https://cran.r-project.org/web/packages/Information/index.html
smbinning	Scoring modellekhez ajánlott csomag, numerikus változók kategorizálására. Szerző: Herman Jopia (2015) Dokumentáció: https://cran.r-project.org/web/packages/smbinning/smbinning.pdf https://cran.r-project.org/web/packages/smbinning/index.html

Forrás: saját szerkesztés

ESETTANULMÁNY - A WOE ÉS AZ IV SZÁMÍTÁS R-BEN

Az alábbiakban egy rövid példa segítségével szeretném bemutatni a WOE és IV számítás menetét, illetve megvizsgálni néhány gyakorlati kérdést. Mivel a WOE és IV alkalmazásáról szóló publikációk általában scoring vagy uplift modellekről tesznek említést, most egy másik kérdéskört választottam, mégpedig a tőzsdei trendek logisztikus regresszióval történő előrejelzését (Zaidi and Amirat 2016).

Mivel a dolgozatban a bizonyítéksúlyra, és az információértékre, illetve a modellek információtartalmára koncentrálunk, és nem a modellezésen van a hangsúly, ezért néhány egyszerűsítéssel élünk. Nem a tökéletes előrejelző modell alkotása a célunk, hanem az, hogy egy életből vett példán nézzük meg a WOE és IV számítás alkalmazását, illetve megvizsgáljuk a kategorizálás. Számunkra most nem a modellek minősége, vagy az előrejelzés pontossága a legfontosabb, hanem hogy miként változik a modell információtartalma, illetve hogy a kategóriaváltozók használata hogyan hat a modell teljesítményére. Így eltekintünk a minőség és az illeszkedés tesztelésére használt módszerek részletes ismertetésétől, illetve alkalmazásától, s csak olyan mértékben foglalkozunk ezzel, amennyire a téma szempontjából feltétlenül szükséges.

a) A modellezésre használt adatállomány

A modellezésre az OTP részvény napi árfolyam adatait használjuk, a 2015/01/01 és 2017/01/31 közötti időszakból. Összesen 522 kereskedési nap adata áll rendelkezésünkre, melyből 148 rekordot a modell tesztelésére használunk.

Zaidi és Amirat (2016) modelljéből indulunk ki, melyben magyarázó változóként az olaj ára mellett a nyitó ár, a maximum ár, a minimum ár és a forgalom szerepel. Ezt az alapmodellt némileg átalakítjuk, egyrészt kihagyjuk az olajárat, másrészt kiegészítjük a modellt olyan statisztikai indikátorokkal, amelyek az idősorban fellelhető rejtett információkat is képesek feltárni (Matur, 2012), így többek között különböző időtávra vonatkozó mozgóátlagokat, illetve az ezekből számítható MACD indikátort építjük be a modellbe, valamint egy olyan változót, amely jelzi, hogy a megelőző napokon hány napig volt változatlan a trend[7].

b) A példában használt regressziós modell

Ahhoz, hogy logisztikus regressziós modellt tudjunk építeni, először képeznünk kell a hozamokból egy bináris kimenetű célváltozót. Mivel a hozamok változását vizsgáljuk, a célváltozó

[7] A modellezés során használt magyarázó változók: *open* – nyitó árfolyam, *maximum* – napi legmagasabb érték, *minimum* – napi legalacsonyabb érték, *turnover* – forgalom a részvények darabszámában megadva, *trend_in_days* – az adott napot megelőzően hány napig volt ugyanaz a trend, *range* – a napi minimum és a napi maximum közötti különbség, *rel_range* – a napi maximum és a napi minimum közötti különbség osztva a napi átlagárral, *ma_8* – 8 napos mozgóátlag, *ma_17* – 17 napos mozgóátlag, *ma_12* – 12 napos mozgóátlag, *ma_26* – 26 napos mozgóátlag, *macd* – macd index, *signal* – az macd 9 napos mozgóátlaga.

értéke akkor lesz 1, ha a T napon számított hozam magasabb, mint T-1 napon számított hozam. A hozamokat a következőképpen határozzuk meg (Zaidi and Amirat 2016):

$$return = \frac{p_j - p_{j-1}}{p_{j-1}} \cdot 100 \quad (25)$$

ahol p_j a j nap záróárfolyama, p_{j-1} pedig a j-1 nap záróárfolyama. Fontos megjegyezni, hogy a trendet próbáljuk megbecsülni, nem a hozamok előjelét. Ha negatív a hozam, de két egymást követő napon csökken, akkor ez ugyanúgy emelkedő trendre utal, mintha a pozitív előjelű hozam növekszik. A regressziós modell annak valószínűségét fogja megbecsülni, hogy T napon növekvő a trend.

$$z_i = c + \beta_1 \beta x_{i1} + \beta_2 \beta x_{i2} + \cdots + \beta_n \beta x_{in} \quad (26)$$

ahol z az odds ráta, $z_i = \log\left(\frac{p_i}{1-p_i}\right)$, x_{ij} a j változó értéke az i esetben, β_j a regressziós együttható, n pedig a magyarázó változók száma. A β értékeket a maximum likelihood módszerrel becsüljük.

c) A WOE és az IV számítása

Mielőtt elkezdenénk modellezni, az első lépésben megvizsgáljuk, hogy van-e olyan változó, amelyet érdemes kategorizálni, egyben teszteljük, hogy a kategorizálás mennyire tekinthető stabilnak különböző minták esetén. Ezt a legegyszerűbben úgy tehetjük meg, ha a WOE és IV számítást először a teljes állományra alkalmazzuk[8], majd egy kisebb állományon is elvégezzük ugyanezt. Az egyszerűség kedvéért most az amúgy is rendelkezésre álló, 70-30 arányban megosztott tanuló és validáló állományt fogjuk felhasználni erre a célra.

3. táblázat:
A változók információértéke három különböző mintán

Dataset (N = 522)		TrainSet (N = 374)		ValidationSet (N = 148)	
Variable	**IV**	**Variable**	**IV**	**Variable**	**IV**
rel_range	0,14705655	rel_range	0,1852636	rel_range	0.4015835
macd1	0,08290204	macd1	0,1733338	macd2	0,3896391
range	0,07753752	turnover_huf	0,1199885	signal3	0,3011544
macd2	0,07045966	open	0,1155685	ma_12	0,2502117
turnover	0,06409273	range	0,1139165	ma_17	0,2424295

Forrás: saját szerkesztés

Minél kisebb a minta elemszáma, annál magasabb az egyes változók információértéke, illetve megfigyelhető a változók sorrendjének átrendeződése is. Ahogy csökken a vizsgált minta elemszáma, úgy a változók egy része „kicserélődik", másrészt egyre magasabb IV értékeket kapunk.

A példában a relatív terjedelemnek van a legmagasabb információértéke, ezt változót fogjuk kategorizálni, egyrészt az üzleti logika alapján nagyjából egyenlő szélességű sávokra osztva, másrészt a bizonyítéksúlyok alapján. A WOE és az IV értékeket az informationValue csomag segítségével határozzuk meg. Az IV értékek alapján eldönthetjük, hogy érdemes-e egyáltalán foglalkoznunk a változók kategorizálásával. Ha viszont a WOE értékek alapján meghatározott kategóriák határait vizsgáljuk, akkor azt tapasztaljuk, hogy ez nagyjából stabilan alakul a különböző mintákon.

---

maximum közötti különbség, rel_range – a napi maximum és a napi minimum közötti különbség osztva a napi átlagárral, ma_8 – 8 napos mozgóátlag, ma_17 – 17 napos mozgóátlag, ma_12 – 12 napos mozgóátlag, ma_26 – 26 napos mozgóátlag, macd – macd index, signal – az macd 9 napos mozgóátlaga.

[8] Miután az read.xlsx() segítségével beolvastuk a fájlt, a következőképpen határozhatjuk meg az információértéket, illetve jeleníthetjük meg a kiválasztott változó kategorizálását:

```
IV <- Information: create_infotables(data=datafile, y ="target", parallel=FALSE)
print(head(IV$Summary), row.names=FALSE)
print(IV$Tables$variable, row.names = FALSE)
```

Ahol a datafile az adatállományt, a target a célváltozót, a variable pedig a megjeleníteni kívánt változót jelöli.

A továbbiakban két különböző kategorizálást használunk, az egyik az üzleti logikát követi, a másik a WOE és IV algoritmus alapján történő kategorizálás.

4. táblázat:
WOE és IV értékek egyenlő hosszúságú osztályközökre

Category1	N	Percent	Nr(Y = 1)	Nr(Y = 0)	Distr(Y=1)	Distr(Y=0)	WOE	IV
[0 - 2,0]	324	0,6207	151	173	0,5898	0,6504	-0,0977	0,0059
[2,1 - 4,0]	169	0,3238	87	82	0,3398	0,3083	0,0975	0,0031
[4,1 - 6,0]	23	0,0441	15	8	0,0586	0,0301	0,6669	0,0190
[6,1 - 8,0]	5	0,0096	2	3	0,0078	0,0113	-0,3671	0,0013
[8,1 - 10,0	0	0,0000	0	0	0,0000	0,0000	0,0000	0,0000
[10,1 - 12,	1	0,0019	1	0	0,0039	0,0000	0,0000	0,0000
					Information Value			**0,0234**

Forrás: saját szerkesztés

5. táblázat:
Az Information package segítségével meghatározott WOE és IV értékek

Category2	N	Percent	Nr(Y = 1)	Nr(Y = 0)	Distr(Y=1)	Distr(Y=0)	WOE	IV
[0,4 - 0,8]	32	0,068376	10	22	0,0391	0,0827	-0,7501	0,0327
[0,9 - 1,1]	61	0,130342	32	29	0,1250	0,1090	0,1368	0,0022
[1,2 - 1,3]	44	0,094017	27	17	0,1055	0,0639	0,5009	0,0208
[1,4 - 1,5]	69	0,147436	29	40	0,1133	0,1504	-0,2833	0,0105
[1,6 - 1,7]	48	0,102564	22	26	0,0859	0,0977	-0,1287	0,0015
[1,8 - 1,9]	40	0,085470	18	22	0,0703	0,0827	-0,1624	0,0020
[2,0 - 2,2]	65	0,138889	31	34	0,1211	0,1278	-0,0541	0,0004
[2,3 - 2,5]	51	0,108974	32	19	0,1250	0,0714	0,5596	0,0300
[2,6 - 3,3]	58	0,123932	22	36	0,0859	0,1353	-0,4542	0,0224
[3,4 - 10,3	54	0,115385	33	21	0,1289	0,0789	0,4903	0,0245
					Information Value			**0,1471**

Forrás: saját szerkesztés

Ahol: Cagegory1: az üzleti logika alapján meghatározott kategorizálás; **Category2:** a WOE algoritmus alapján meghatározott kategorizálás; **N:** az adott kategóriába tartozó elemek száma; **Percent:** az egyes kategóriák aránya a mintán belül; **Nr(Y=1):** a célváltozó Y=1 értékének előfordulása az adott kategóriában; **Nr(Y=0):** a célváltozó Y=0 értékének előfordulása az adott kategóriában; **Distr(Y=1):** a célváltozó Y=1 értékének eloszlása, adott kategóriába tartozó elemszám / Y=1 értékek elemszáma a mintában; **Distr(Y=0):** a célváltozó Y=1 értékének eloszlása, adott kategóriába tartozó elemszám / Y=1 értékek elemszáma a mintában; **WOE:** bizonyítéksúly; **IV:** információérték.

A fenti táblázatokat összehasonlítva rögtön látható az algoritmus alapján történő kategorizálás egyik gyakorlati haszna: elkerülhetjük azt a hibát, hogy egy kategóriának túl kevés az elemszáma. Ez a modellezésnél okozhat problémát, ha a teszt állományban szerepel egy olyan érték, ami a tanuló állományból hiányzik, így a modell nincs felkészítve rá.

Általában olyan változókat érdemes kategorizálni, amelyek nem szignifikánsak a modell szempontjából, de a kategorizálással szignifikánssá válhatnak, vagy amelyeknél alapból nem nincs monoton kapcsolat a célváltozóval.

Az eredményeket összehasonlítva látható, hogy a WOE és IV algoritmus alapján készült csopor-

tosításnak nagyobb az információértéke, úgyhogy elvileg inkább ez a csoportosítás tűnik megfelelőnek arra, hogy a modellben használjuk.

Az értékek összehasonlítása és ábrázolása mellett létezik egy másik módszer is az eredmények tesztelésére. Egy egyszerű logisztikus regressziós illesztés segítségével ellenőrizhető, hogy megfelelő-e a WOE transzformáció (Zeng 2014). Ha mindössze a WOE transzformáción átesett változót építjük be a modellbe, akkor a maximum likelihood becslésnek van egy explicit megoldása, mégpedig a (ahol b és g a „bad" és a „good" kimenet) és a . Ha jól végeztük el a WOE transzformációt, akkor a modellnek ezt az eredményt kell adnia. Ha a tengelymetszet nem és a meredekség nem 1, akkor nem jó az algoritmus.

d) A lefuttatott regressziós modellek

A modellezéshez két részre osztjuk az állományt, 374 rekordot használunk a modell építésére, 148 rekordot pedig a modell validálására[9]. Különböző regressziós modelleket vizsgálunk, hogy a változók beépítésének hatását nyomon követhessük[10]:

- **reg0:** az alapmodellben csak az alapvető változókat használjuk (nyitó ár, legalacsonyabb ár, legmagasabb ár, illetve forgalom a részvények számában megadva).
- **reg1:** kibővítjük az alapmodellt további változókkal, de a kategória változókat egyelőre nem használjuk. A minimum vagy a maximum árat kicseréljük a *terjedelemmel*, ami azt mutatja, hogy adott napon milyen sávban mozgott az árfolyam. Mivel ez egy származtatott változó, lineárisan függ a minimumtól és a maximumtól, ahhoz, hogy használni tudjuk, ki kell hagynunk ez utóbbiak közül az egyik változót. Ugyanez igaz a mozgóátlagokra, és a belőlük számított MACD indexre is, vagy a két mozgóátlagot használjuk, vagy ezek közül az egyiket és az MACD indexet.
- **reg2:** a *reg1* modellt annyiban módosítjuk, hogy a *terjedelem* helyett az üzleti logikának megfelelő kategorizálást használjuk.
- **reg3:** szintén a *reg1* modellből indulunk ki, de most a *terjedelem* helyett a WOE és IV módszerrel meghatározott kategorizálást használjuk.
- **reg4:** a *reg3* modellt leszűkítjük a szignifikáns magyarázó változókra.

Az eredményekkel kapcsolatos feltételezéseink a következők: (i) Feltételezzük, hogy az alapmodell kibővítése javítja a modell teljesítményét. (ii) Feltételezzük, hogy a kategóriaváltozók alkalmazása szintén javítja a modell teljesítményét, és arra számítunk, hogy a WOE és IV módszer segítségével meghatározott kategorizálással jobb eredményt érünk el. (iii) Végül korábbi eredmények alapján feltételezzük, hogy a nem szignifikáns változók elhagyása is növeli a modell teljesítményét.

e) Az eredmények értékelésének szempontjai

A logisztikus regressziós modellek kiértékelése egy nagyobb témakör, amivel most nem foglalkozunk részletesen, csupán néhány mutatószámokat nézünk meg, ezek közül is elsősorban az ún. információs kritériumokra koncentrálunk. A logisztikus regressziós modellek értékelésére egyébként nincsen általánosan elfogadott szabály,

---

[9] Egy lehetséges megoldás R-ben az állomány 70-30 arányban történő megosztására:

```
set.seed(123)
index <- sample(2,nrow(dataset),replace=TRUE,prob=c(0.7,0.3))
train <- dataset[index==1,]
validation <- dataset[index==2,]
```

ahol a *dataset* a felosztani kívánt állomány neve. A *set.seed()* függvény segítségével úgy generálhatunk véletlen számsorozatot, hogy az bármikor ugyanúgy megismételhető.

[10] A regressziós modell a következőképpen tanítható:

```
reg0 <- glm(target ~ open  + minimum + maximum + turnover, family = binomial,data = train)
```

A többi modellnél a következő magyarázó változókat használjuk:

```
reg1: open + maximum + turnover + rel_range + trend_in_days + ma_8 + ma_17 + signal1
reg2: open + maximum + turnover + rel_range_cat1 + trend_in_days + ma_8 + ma_17 + signal1
reg3: open + maximum + turnover + rel_range_cat2 + trend_in_days + ma_8 + ma_17 + signal1
reg4: open + minimum + turnover + rel_range_cat2 + ma_8
```

Az eredményeket a summary(reg0) parancs segítségével lehet kiírni.

de a modellek rangsorolásánál általában három szempontot szokás figyelembe venni: (i) hogy mennyire jó a modell illeszkedése, (ii) hogy az egyes változók statisztikai szempontból mennyire szignifikánsak, végül (iii) hogy mennyire pontos a modell által adott előrejelzés[11].

Ahogy a lineáris regressziós modelleknél az R^2 az általánosan elfogadott mérőszám, úgy a nem lineáris modelleknél léteznek úgynevezett pszeudo R^2 mutatószámok, ilyen például a MacFadden-féle R^2. Minél közelebb van a mutató értéke a 1-hez, annál erősebb kapcsolatra utal.

A modellek által adott előrejelzés pontossága többféleképpen is mérhető. Az egyik lehetséges megoldás a klasszifikációs ráta, ami azt fejezi ki, hogy az esetek hány százalékában adott pontos előrejelzést a modellünk. Emellett meghatározható a ROC-görbe is, illetve a görbe alatti terület mérőszáma, az area under curve (AUC) mutató, ez utóbbi alapján fogjuk most összehasonlítani a modelleket.

Ami számunkra a téma szempontjából a leginkább érdekes, az az információs kritériumok viselkedése. Ezek a mérőszámok a modell egészének információértékét próbálják számszerűsíteni. Általános szabály, hogy több modell közül azt kell választani, amelynél alacsonyabb az információs kritérium értéke. A szakirodalom több információs kritériumról tesz említést.

Az AIC (Akaike Information Criterion) egy „realatív" mérőszám[12], amely az információelméletre vezethető vissza, és az adott modell alkalmazásával járó információveszteséget próbálja megbecsülni (Akaike 1973). Hátránya, hogy abszolút értelemben semmit sem mond a modell minőségéről, nem ad jelzést arra vonatkozóan, hogy miként illeszkedik a modell, így önmagában nem elégséges kritérium a kiértékelésre. Az AIC a következőképpen határozható meg:

$$AIC = 2k - kln(L) \qquad (27)$$

ahol L a modell likelihood függvényének maximuma,k pedig a becsült paraméterek száma. Az információs kritérium értéke tehát egyrészt a modell devianciájától[13], másrészt a paraméterek számától függ. Mivel egy jól illeszkedő modellnél az AIC értéke kicsi, az általános szabály az, hogy azt a modellt célszerű választani, amelynél a legkisebb az AIC érték. A mutatószámban két egymással ellentétes hatást ragad meg: a jól illeszkedő modell devianciája kicsi, ha viszont túl sok paraméter kerül a modellbe, akkor a modell egyre közelebb kerül a telített modellhez, és csökken a lényegkiemelő szerepe. Általános szabály, hogy ha két egymásba ágyazott modell között 2-nél kisebb az AIC értékek eltérése, akkor a két modell nem különbözik lényegesen.

A Schwarz-féle (SBC, SBIC) információs kritérium, amely bayesi információs kritériumként (BIC) is ismert, szintén a likelihood függvényen alapul, és szoros kapcsolatban van az AIC-vel (Schwarz 1978). A BIC értéke a következőképpen határozható meg:

$$BIC = (\mathrm{n})k - 2\ln(\hat{L}) \qquad (28)$$

ahol az modell likelihood függvényének maximum értéke, azaz , ahol a likelihood függvényt maximalizáló paraméter értékeit jelöli, pedig az adatokat; a minta elemszáma, pedig a becsült paraméterek száma. A kritérium elméleti és matematikai háttere meglehetősen komplex, itt elegendő annyit megjegyeznünk, hogy a kiválasztásnál azt a modellt preferáljuk, amelyiknél alacsonyabb az BIC érték (Dziak *et al.* 2012).

---

[11] A modell illeszkedésének tesztelésére használható módszerek: Likelihood Ratio Test, Pseudo R2, Hosmer-Lemeshow Test. Az egyes változók vizsgálatára használható módszerek: Wald teszt, variable importance. A modell által adott előrejelzés kiértékelése: klasszifikációs mátrix, klasszifikációs ráta, ROC Curve, K-Fold Cross Validation.

[12] Látezik két kiterjesztése: CAIC, CAICF. CAIC (Consistent Akaikes Information Criterion)

[13] A modell devianciája a loglikelihood-függvény maximumának mínusz kétszerese. Az adott modell devianciájának, illetve a telített modell devianciájának különbsége az ún. reziduális deviancia, a nullmodel devianciája és a telített modell devianciája közötti különbség pedig a nulldeviancia. A reziduális deviancia és a nulldeviancia nem lehet negatív, és a reziduális deviancia pedig nem nagyobb a nulldevianciánál. Az igazán lényeges kérdés az, hogy egy szűkebb modell devianciája mennyivel nagyobb egy bővebb modell devianciájához képest.

A harmadik lehetséges módszer a Hannan-Quinn információs kritérium (HQC) alkalmazása (Hannan & Quinn 1979):

$$HQC = -2L_{max} + 2k\ln(\ln(n)) \quad (29)$$

ahol a log-likelihood, a paraméterek száma, pedig a megfigyelések száma.

Példánkban két dolgot vizsgálunk, egyrészt miként változik a modell információtartalma, az egyes módosítások hatására, másrészt miként befolyásolja a kategória változó alkalmazása az eredményeket.

f) Az eredmények értékelése

Az alábbi táblázat a különböző regressziós modelleknél mért eredményeket foglalja össze.

6. táblázat:
A vizsgált regressziós modellek eredményei[14]

Modell	McFadden R2	AUC	AIC	BIC
reg0	0, 1867239	0,7446886	431,6539	451,2752
reg1	0, 2062494	0,7824176	429,5307	464,849
reg2	0, 2300604	-	421,1855	464,3523
reg3	0, 2231327	0,7954212	436,7773	503,4896
reg4	0, 2339786	0,7749084	425,154	480,0936

Forrás: saját szerkesztés

Az eredeti modell további változókkal történő bővítése, majd a kategorizálás javítja a modell illeszkedését, illetve pontosságát, ami abból látszik, hogy magasabb pszeudo R^2 és AUC értéket kapunk, mint az alapmodell esetén.

Az információs kritériumokra gyakorolt hatás azonban nem ilyen egyértelmű. Egyrészt látható, hogy az AIC és a BIC nem feltétlenül egy irányba változik. Az alapmodell bővítése, illetve a végső modellben csak a szignifikáns változók használata láthatóan csökkenti az AIC értékét, a BIC értéke viszont növekszik. Ez azért érdekes, mert mindkét mutató esetében az alacsonyabb értékű modellt kell választani. Másrészt az is érdekes, hogy amíg az üzleti logika szerinti kategorizálás csökkenti az AIC értékét (reg2), addig az elvileg magasabb információértékkel rendelkező kategorizálásnál magasabb AIC értéket kapunk (reg3), tehát ez utóbbi modell valamivel gyengébben teljesít ebből a szempontból.

A reg2 modell esetén nem tudjuk meghatározni az AUC értéket, ennek az az oka, hogy az üzleti logika szerint készített kategória változó esetében néhány kategóriánál olyan kicsi az elemszám, hogy a tanuló a tanuló állományba nem is került bele, így a modell nincs „felkészítve" bizonyos értékekre.

Látható, hogy a magasabb információértékű kategória változó alkalmazása nem feltétlenül javítja az információs kritérium értékét. A reg2 modell esetén az AIC és a BIC is magasabb, mint a reg1 modell esetében. Más szempontból viszont hasznos a kategorizálás, hiszen a Pszeudo R^2 érték és az AUC is magasabb, mint a reg1 modell esetén.

Összességében megállapíthatjuk, hogy a kategorizálást elsősorban olyan változóknál érdemes alkalmazni, amelyek eredeti állapotukban nem szignifikánsak, de átalakítva szignifikáns változóvá tehetők. A WOE szerinti kategorizálás mellett szól, hogy ebben az esetben nem futhatunk olyan hibába, mint az üzleti logika alapján történő kategorizálásnál, hogy egyes sávokban túl kicsi az elemszám, emiatt előfordulhat, hogy a kategorizált változónak nem ugyanaz lesz az értékkészlete a teszt állományban, mint a tanuló állományban. Mindezek ellenére látszik, hogy a modellre gyakorolt hatás nem egyértelmű, nagyban függ attól, hogy milyen szempontok alapján értékeljük ki a modellt.

[14] A McFadden Pszeudo R^2 kiszámításához a *pscl* csomagot kell betölteni, és ebből a *pR2()* függvényt használni. A ROC görbe megrajzolásához, illetve az AUC számításához a *ROC* csomagot kell betölteni. Az információs kritériumok értékét az *AIC()* és a *BIC()* függvényekkel lehet kiszámítani.

ÖSSZEGZÉS

A tanulmány alapvető célja az volt, hogy megismertesse a bizonyítéksúly, és az információérték fogalmát, elméleti hátterét, illetve alkalmazásának lehetőségeit, másrészt népszerűsítse az R program használatát.

Következtetések:

- Az említett módszerek nem minden típusú modellnél alkalmazhatók jól. Valószínűleg nem véletlen, hogy elsősorban scoring és uplift modelleknél terjedt el a bizonyítéksúly és az információérték használata. Példaként egy elég speciális témakört vizsgáltunk meg, a tőzsdei trendek előrejelzését, de bizonyos változóknál itt is lehet létjogosultsága a módszer alkalmazásának.
- A kategorizálás viszonylag stabil, viszont a változók információértéke, illetve az információérték alapján meghatározott sorrendje a minta nagyságától függően eltérő.
- Az információs kritériumra hatással van a kategorizálás, de nem minden esetben az a kategorizálás eredményezi a legjobb AIC vagy BIC értéket, amelyik a WOE és az IV alapján optimális. Érdemes lenne megvizsgálni mélyebben az IV és az AIC, illetve a többi információs kritérium közötti összefüggéseket, mert ez nem feltétlenül egyértelmű. Ugyanígy a modell illeszkedésére gyakorolt hatás sem egyértelmű.
- Az alapmodell teljesítménye jelentősen javítható, ha olyan változókkal bővítjük, amelyek az idősorból kinyerhető információkat is megjelenítik, illetve ha a végső modellt leszűkítjük a szignifikáns változókra. Mindkét hatás kimutatható az információs kritériumok értékének változásában.

Gyakorlati jelentőség:

- A módszer nemcsak a folytonos változók kategorizálására alkalmazható, hanem exploratív eszközként is használható.
- Akkor is érdemes kiszámítani a bizonyítéksúlyokat és az információértéket, ha az üzleti logika alapján szeretnénk kategorizálni, de nem egyértelmű, hogy két kategorizálás közül melyik a jobb.
- Olyan változóknál is alkalmazható a módszer, amelyek szerint szegmentálni szeretnénk az állományt.
- A logisztikus regressziós modellek kiértékelésénél mindenképpen több szempontot kell figyelembe venni, mivel nem minden esetben egyértelmű a változtatások hatása.
- Érdemes lenne alaposabban vizsgálni az információs kritériumok közötti összefüggéseket.

HIVATKOZÁSOK

Agterberg, F. P., Bonham-Carter, G. F., Cheng, Q., Wright, D. F. (1993), „Weights of evidence modeling and weighted logistic regression for mineral potential mapping", in: Davis, J. C. and Herzfeld, U. C. (eds.), *Computers in Geology, 25 Years of Progress*, Oxford University Press, Oxford, pp.13-22

Akaike, H. (1973), „Information theory and an extension of the maximum likelihood principle", in: B. N. Petrov & F. Csaki (eds.), *Second international symposium on information theory*, Budapest: Akadémiai Kiadó, pp.267-81

Barbieri, G. and Cambuli, P. (2009), „The weight of evidence statistical method in landslide susceptiblity mapping of the Rio Pardu Valley (Sardinia, Italy)", 18th World IMACS / MODSIM Congress, Cairns, Australia, 13-17 July 2009

Bonham-Carter, G. F. (1994), *Geographic Information Systems for Geoscientists*, Oxford, UK: Elsevier

Carranza, J. M. (2009), *Geochemical Anomaly and Mineral Prospectivity Mapping in GIS*, Vol. 11, Amsterdam: Elsevier

Csicsman J. – Sipos Szabó E. (2011), *Matematikai alapok az adatbányászati szoftverek első megismeréséhez*, Egyetemi jegyzet, Downloadable [2016-08-21]: http://www.inf.u-szeged.hu/~csicsman/oktatas/statprog/konyv/stat_book.pdf

Dinerot, T. (1996), "Seven Reasons Why You Should Not Categorize Continuous Data", *Journal of Health & Social Policy* **8** 1, pp.63-72

Dziak, J., Coffman, D., Lanza, S., Li, R. (2012), *Sensitivity and Specificity of Information*, Criteria, The Methodology Center, Pennsylvania State University, Technical Report Series #12-119. Downloadable [2017-02-25]: https://methodology.psu.edu/media/techreports/12-119.pdf

Gillies, D. (1990), "The Turing-Good weight of evidence function and Popper's measure of severity of a test", *British Journal for the Philosophy of Science*, **41** 1, pp.143-6

Good, I. J. (1979), "Studies in the history of probability and statistics. XXXVII A. M. Turing's statistical work in the World War II.", *Biometrika*, **66** 2, pp.393-6

Good, I. J. (1985), "Weight of Evidence: A Brief Survey" in: J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith (eds.): *Bayesian Statistics: Proceedings of the Second International Meeting in Valencia*, New York: North-Holland, pp.249-70

Hámori G. (2014), *Predikciós célú klasszifikációs statisztikai modellek gyakorlati kérdései*, Doktori (PhD) értekezés, Kaposvári Egyetem, Gazdálkodás- és Szervezéstudományok Doktori Iskola

Hámori G. (2016), „A magyarázóváltozók kezelésének egyes kérdései regressziós modellezés során", *Statisztikai Szemle*, **94** 1, 5-21. old.

Hannan, E. J., Quinn, B. G. (1979), „The Determination of the Order of an Autoregression", *Journal of the Royal Statistical Society, Series B*, **41** pp.190-5

Harrel, F. E. (2015), „Problems Caused by Categorizing Continuous Variables", Online [2016-08-20]: http://biostat.mc.vanderbilt.edu/wiki/Main/CatContinuous

Hartley, B. K. (2014), *Evaluation of Weights of Evidence to Predict Gold Occurences in Northern Minnesota's Archean Greenstone Belts*, Master of Science Thesis, Faculty of USC Graduate School, University of Southern California

Hunyadi L. (2011), „Bayesi gondolkodás a statisztikában" *Statisztikai Szemle*, **89** 10-11, 1150-71. old.

Jopia, H. (2015), „R Package 'smbinning': Optimal Binning for Scoring Modeling", Online [2016-08-21]: http://blog.revolutionanalytics.com/2015/03/r-package-smbinning-optimal-binning-for-scoring-modeling.html

Larrañaga, P., Bielza, C. (2012), „Alan Turing and Bayesian statistics" *Mathware & Soft Computing Magazine*, **19** 2, pp.23-4

Larsen, K. (2015), „Data Exploration with Weight of Evidence and Information Value in R", Online [2016-08-21]: http://multithreaded.stitchfix.com/blog/2015/08/13/weight-of-evidence/

Larsen, K. (2016), „Information Package Vignette", Online [2016-08-21]: http://127.0.0.1:29474/library/Information/doc/Information-vignette.html

Lee, T., Zhang, R., Meng, X., Ryan, L. (2013), „Incremental Response Modeling Using SAS Enterprise Miner", SAS Global forum 2013. Downloadable [2016-08-21]: https://support.sas.com/resources/papers/proceedings13/096-2013.pdf

Lin, A. Z. (2015), „Entrophy-based Measures of Weight of Evidence and Information Value for Variable Reduction and Segmentation for Continuous Dependent Variables", Downloadable [2016-08-21]: http://support.sas.com/resources/papers/proceedings15/3242-2015.pdf

Lin, A. Z. and Hsieh, T. Y. (2014), „Expanding the Use of Weight of Evidence and Information Value to Continuous Dependent Variables for Variable Reduction and Scorecard Development", SESUG. Paper SD-84. Downloadable [2016-08-21]: http://www.lexjansen.com/sesug/2014/84_file_final.pdf

Popper, K. R. (1934), The Logic of Scientific Discovery, 6th Impression, Hutchinson, 1972, Magyarul megjelent: A tudományos kutatás logikája, Európa Könyvkiadó, Budapest, 1997

Radcliffe, N. J. (2007), „Using Control Groups to Target on Predicted Lift: Building and Assessing Uplift Models", Direct Marketing Analytics Journal, pp.14-21

Radcliffe, N. J. and Surry, P. D. (2011), Real-World Uplift Modelling with Significance-Based Uplift Trees, Stochastic Solutions White Paper

Rzepakowski, P. and Jaroszewicz, S. (2012), „Uplift Modeling in Direct Marketing", *Journal of Telecommunications and Information Technology*, 2, Downloadable: http://www.nit.eu/czasopisma/JTIT/2012/2/43.pdf

Siddiqi, N. (2006), *Credit Risk Scorecards. Developing and Implementing Intelligent Credit Scoring*, John Wiley & Sons, New York

Strickland, J. (2015), „What are Uplift Models?" Online: http://www.analyticbridge.com/profiles/blogs/what-are-uplift-models

Schwarz, G. (1978), „Estimating the Dimension of a Model", *Annals of Statistics*, 6 2, pp.461-4

Zaidi, M. and Amirat, A. (2016): „Forecasting Stock Market Trends by Logistic Regression and Neural Networks. Evidence from KSA Stock Market", *International Journal of Economics, Commerce and Management*, **IV** 6, Downloadable [2016-08-21]: http://ijecm.co.uk/wp-content/uploads/2016/06/4614.pdf

Zeng, G. (2014), „A Necessary Condition for a Good Binning Algorithm in Credit Scoring", *Applied Mathematical Sciences*, **8** 65, pp.3229-42

Nagy Gábor Szabolcs, PhD hallgató
nagy.gabor.szabolcs@web.de
Pécsi Tudományegyetem
Közgazdaságtudományi Kar

The application of the weight of evidence and information value approach in predictive modeling for binning continuous variables

AIM OF THE PAPER

The aim of this study is to discuss the concept of weight of evidence and information value, to give an overview on their theoretical background and their possible practical applications, as well as to propagate the usage of R, which is a free software environment for statistical computing and graphics.

METHODOLOGY

The study deals with the use of weight of evidence (WOE) and information value (IV) for the binning of continuous predictors in logistic regression. The overview on the theoretical background and the practical applications basesd on a literature study. The effect of the binning is examined with a logistic regression model, which predicts the stock exchange trends.

RESULTS

The result of the WOE binning is relatively steady, but the information value of the binned predictors, and the ranking of the predictors based on their information value differs by sample size. The value of the information criterions are influenced by the usage of the binned predictors, but not necessarily the optimal WOE binning gives the best AIC or BIC result.

IMPLICATIONS

The WOE and IV method can be used in the exploratory data analysis too. It's worth to calculate WOE and IV even if we want to use a binning which based on the business logic, but it is not clear, which categorization is better. WOE also can be used for categorization of those variables which we want to use for segmentation. By evaluation of the logistic regression models several factors has to be considered, because the effect of the changes in the predictors is not always clear. It would be worth to examine the relationships between the information criterions.

Keywords: *weight of evidence, information value, uplift modeling, logistic regression*