DOKTORI (PhD) ÉRTEKEZÉS TÉZISEI

HÁMORI GÁBOR

Kaposvári Egyetem

2014

KAPOSVÁRI EGYETEM

Gazdálkodás- és Szervezéstudományok Doktori Iskola

Doktori Iskola vezetője:

PROF. DR. Kerekes Sándor

MTA Doktora

Témavezető:

PROF. DR. SZÁZ JÁNOS

Egyetemi Tanár

PREDIKCIÓS CÉLÚ KLASSZIFIKÁLÓ STATISZTIKAI MODELLEK GYAKORLATI KÉRDÉSEI

Készítette:

Hámori Gábor

KAPOSVÁR

2014

DOI: 10.17166/KE.2015.007

Tartalomjegyzék

Kutatás előzményei, célkitűzések

Az előrejelzési célú (predikciós) regressziós modellezés gyakorlatához kapcsolható technikai részletkérdések külön-külön nagyon jól feltérképezett területnek tekinthetők, az egyes problémákra fokuszálva nagyon sok mérvadó tanulmány született az elmúlt évtizedekben. A szerzők általában a kérdések mögött meghúzódó matematikai háttér igényes bemutatásával és igénybevételével izoláltan, a modellkészítés folyamatából kiragadva, elemezik az egyes részletterületeket, problémákat. Az alkalmazott statisztikai modellezés, a szélesebb értelemben vett adatbányászati megoldások rohamos és iparszerű elterjedésével azonban megjelent egy új igény az egyes, a modellfejlesztés során felmerülő kérdéseket a modellezés folyamatában vizsgáló és elemző megközelítés iránt. A disszertáció a predikciós célú bináris klasszifikációk részterületén, a logisztikus regressziós modellt a vizsgálódás középpontjába helyezve, tesz kísérletet ennek a statisztika területén újszerű szemléletmódnak az érvényesítésére. A tanulmány segítségével átfogó képet kapunk a teljes modellezési folyamatról, az egyes lépéseknél jelentkező tipikus döntési helyzetekről, és az azokra adható válaszokról, legjobb gyakorlatokról. A hivatkozások segítségével az egyes részletkérdések iránt mélyebb ismeretre vágyók is útmutatást kapnak A disszertáció így mind a gyakorló, mind az elméleti szakemberek érdeklődésére is igényt tarthat.

A predikciós célú klasszifikáló modellek gyakorlati alkalmazása esetében azonban kitüntetett jelentőségű az elkészült modell illeszkedése, klasszifikáló ereje. A disszertációban ezért megvizsgálunk három olyan hipotézist, melyek segítségével a logisztikus regressziós modell illeszkedése jelentősen javítható.

- ***Hipotézis 1***: A modellezési adatbázisban a szélsőséges, kiugró (outlier) értékek megfelelő kezelésével a modell illeszkedése javul.
- ***Hipotézis 2***: A folytonos változók kategorizálása még abban az esetben is növelheti a modell prediktív erejét, amennyiben célváltozóval az eredeti folytonos változó monoton kapcsolatban van
- ***Hipotézis 3***: Ismeretes, hogy a folytonos változók kategorizálása esetén az egyes kategóriákat meghatározó kategóriahatárok számának és elhelyezkedésének megválasztása befolyásolja a változó prediktív erejét a modellben. A CHAID-

algoritmus, disszertációban ismertetésre kerülő, részalgoritmusa alkalmas az illeszkedés maximálása szempontjából optimális kategóriahatárok kialakítására.

Anyag és módszer

A vizsgálandó hipotézisek igazolásához egy-egy valós, banki illetve kórházi adatbázist használtunk fel. Az adatvédelmi okok miatt az adatállományok pontos származási helyét nem tüntethettük fel. Az elemzéshez az SPSS.20 statisztikai szoftver került alkalmazásra.

Eredmények, következtetések

Az alábbiakban bemutatjuk az egyes hipotézisek vizsgálatát és az abból nyert következtetéseket:

Hipotézis 1 vizsgálata

Mivel a folytonos változók szélsőséges, kiugró értékei az úgynevezett outlierek általában rontják az alkalmazott modell illeszkedését, ezért a modellezés során külön gondot kell fordítani a szélsőséges, kiugró értékek felismerésére és megfelelő kezelésére. Különösen fontos ez a lineáris modellek esetén, ahol a kovariancia mátrix számítása során felhasznált folytonos változók átlagai jelentősen módosulhatnak a szélsőséges értékek következtében, nagy hatást gyakorolva ez által magának a mátrixnak egyes elemeire is. A hipotézis vizsgálatban a „KOR” folytonos életkor változóval szeretnénk prediktálni a műtét utána szövődményes állapot bekövetkezését. Az előrejelzéshez dichotóm logisztikus regressziós modellt alkalmazunk. Egy fővárosi kórház adatbázisa 350 beteget tartalmazott, melyből 98 esetben következett be szövődmény a 2008-2011 megfigyelési időszakban. A legfiatalabb beteg 15, a legidősebb 87 éves volt. A normális eloszláshoz hasonló eloszlás átlaga 54,86 mediánja 56 év. A „KOR” magyarázó változó felhasználásával futtatott logisztikus regresszió paraméterbecsléseit foglalja össze a következő táblázat.

1. Táblázat. Paraméterbecslés.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1	KOR	,032	,009	13,120	1	,000	1,032
	Constant	-2,735	,521	27,558	1	,000	,065

Látható, hogy a „KOR" szignifikáns magyarázó változó (*P=0,000*) a modellben. Következő lépésben szövődménymentes 77 éves beteget 577 és egy szövődményes 84 éves beteget 384 évesre „változtattunk". Ezáltal létrehoztunk két outlier megfigyelést. A regressziót újra futtatva a következő eredményeket kaptuk:

2. Táblázat. Paraméterbecslés.Outlier

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1	KOR	,002	,003	,524	1	,469	1,002
	Constant	-1,072	,213	25,335	1	,000	,342

Látható, hogy nemcsak a paraméterek értéke változott meg, hanem a magyarázó változó szignifikancia értéke (*P=0,469*) is . A „KOR" változó semmilyen szokásos szinten nem tekinthető szignifikánsnak ebben a modellben. A vizsgálat megmutatta, hogy a szélsőséges megfigyelések megfelelő kezelése nélkül modelljeink illeszkedése, és ez által magyarázó ereje jelentősen romolhat.

Hipotézis 2 vizsgálata

Közismert, hogy mind a lineáris, mind a logisztikus regressziós modell esetén, az alkalmazott függvénytípus következtében, csak a célváltozó tekintetében monoton kapcsolatot mutató folytonos változók esetében várhatunk megfelelő illeszkedést. Ezért a nem monoton kapcsolatot mutató változók 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, hogy a folytonos változó nem szignifikáns, míg a kategorizált párja szignifikáns magyarázó változónak bizonyul modellünkben. A hipotézis vizsgálatához az előző kórházi adatbázist használtuk, ahol az életkorral (KOR) és a testtömegindex (BMI) folytonos változókkal szeretnénk előrejelezni a műtét utáni szövődmények előfordulását. Az alábbiakban láthatjuk a két változót egyszerre a

modellbe bevonó (enter) logisztikus regressziós modell paraméterbecslésével kapcsolatos statisztikákat feltüntető táblázatot.

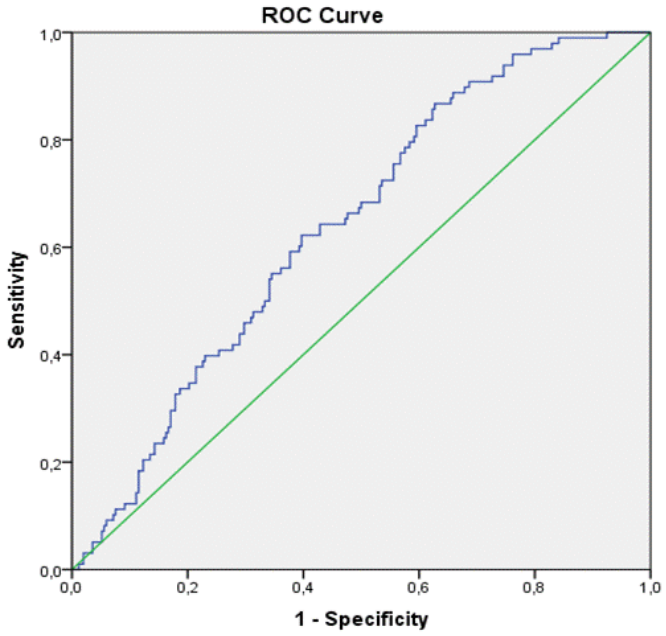
3. Táblázat. Paraméterbecslés (KOR, BMI)

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1	BMI	,010	,009	1,187	1	,276	1,010
	KOR	,032	,009	13,088	1	,000	1,032
	Constant	-3,024	,585	26,709	1	,000	,049

Látható, hogy a két változó közül csak a KOR változó szignifikáns a szokásos szinteken. Ennek megfelelően a modell illeszkedése is nagyon gyenge (GINI=28%) az alábbi ROC-görbének megfelelően

1. ábra. ROC görbe (KOR, BMI)



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

4. Táblázat. BMI kategorizálása

BMI	*0-28*	*28-34*	*34-*
BMIKAT2	*1*	*2*	*3*
SZÖVŐDMÉNYRÁTA	*12,70%*	*31,40%*	*44,60%*
WOE	*1,927748*	*0,781485*	*0,216846*

Látható, hogy kategóriák mentén monoton csökken a WOE értéke, tehát a BMI folytonos változó a célváltozóval való kapcsolata monoton, azaz a magasabb BMI értékek nagyobb szövődmény rátákat vonzanak. A változó ennek ellenére nem bizonyult szignifikánsnak eredeti modellünkben. A regressziót újra futtatva az életkor folytonos és a testtömegindex kategorizált változóval a következő eredményeket kapjuk.

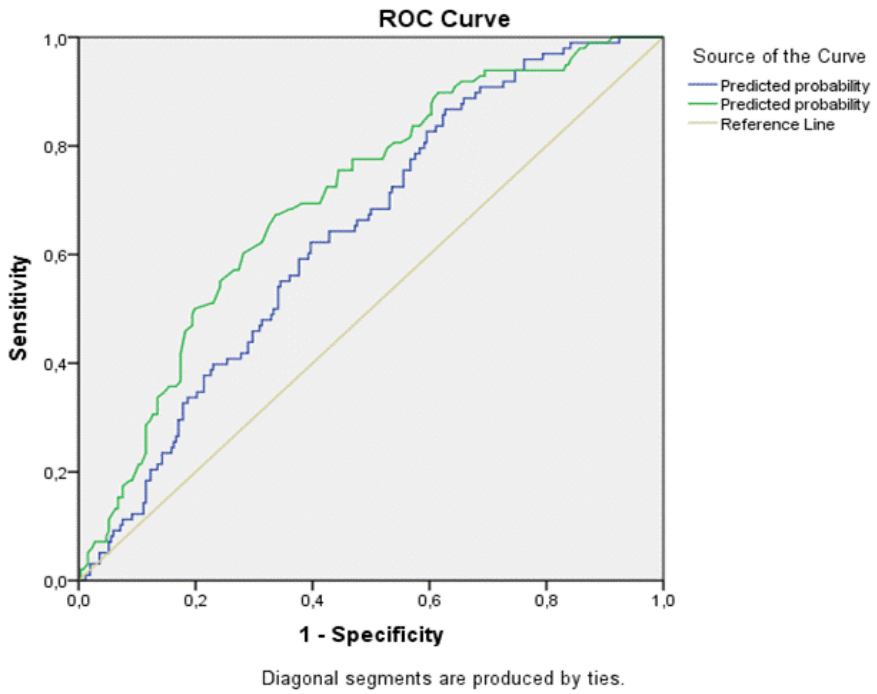
5. Táblázat. Paraméterbecslés (KOR, BMIKAT2)

Variables in the Equation

		B	*S.E.*	*Wald*	*df*	*Sig.*	*Exp(B)*
Step 1	*KOR*	*,034*	*,009*	*13,767*	*1*	*,000*	*1,035*
	BMIKAT2	*,879*	*,189*	*21,561*	*1*	*,000*	*2,409*
	Constant	*-4,615*	*,711*	*42,113*	*1*	*,000*	*,010*

A testtömegindex itt már szignifikáns és ennek megfelelően alakul a modell illeszkedése is, amely most már sokkal jobb (GINI=40%), mint az eredeti modellünk esetében. Az alábbi ábrán a két modellhez tartozó ROC-görbéket egyszerre tanulmányozhatjuk.

2. ábra. ROC görbe (KOR, BMIKAT2)



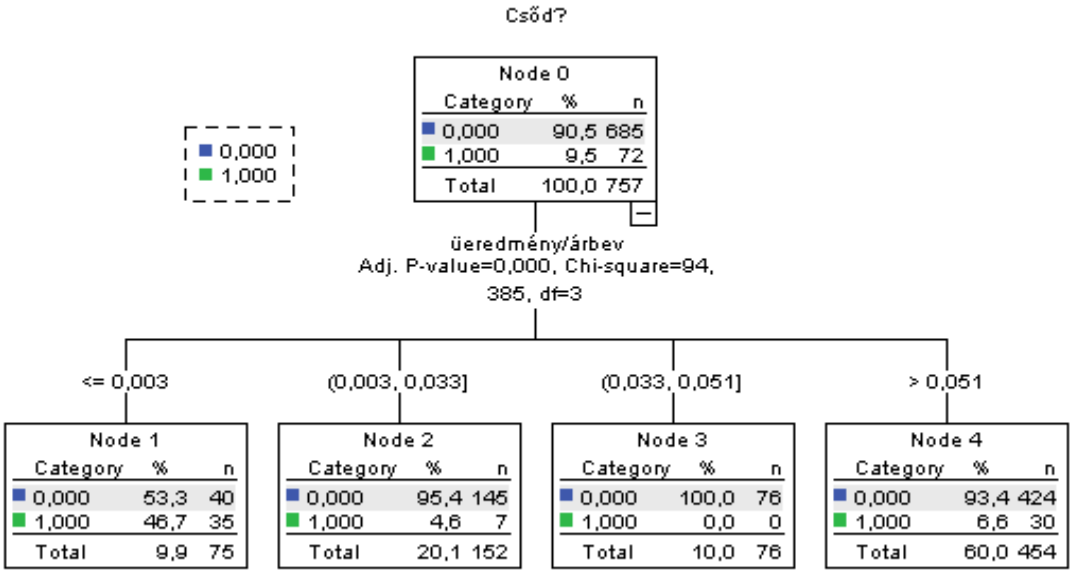
Hipotézis 3 vizsgálata

A harmadik hipotézis vizsgálatánál az alapvető kérdés az, hogy hány kategóriával rendelkezzen a folytonos változóból kategorizálással létrehozott új változó, illetve hol legyenek a kategória határok. Mivel a kategóriák száma és elhelyezkedése hatást gyakorol a végső változó perdikciós erejére a modellben, nem mindegy, hogy a folytonos skála felosztását milyen módon hajtjuk végre. Az optimális felosztás megtalálására alkalmas algoritmus jellemzője, hogy a célváltozó tekintetében a felosztást úgy hajtja végre, hogy a létrejövő kategóriákon belül a homogenitás, míg a kategóriák között heterogenitás a legnagyobb legyen. Ilyen algoritmus, egy a döntési fák családjába tartozó rekurzív klasszifikáló eljárás, az úgynevezett CHAID (Chi-squared Automatic Interaction Detector) algoritmus részalgoritmusa. Az algoritmus célja, hogy a *K* különböző kategóriával rendelkező változó[1] esetében összevonásra kerüljenek azok a kategóriák, melyek legkevésbé különböznek egymástól az m különféle kategóriával rendelkező célváltozó tekintetében [2]. Ehhez X_i kategorizált folytonos változó kategóriái közül az összes lehetséges módon kiválaszt kettőt. Amennyiben a vizsgált magyarázó változó K különböző kategóriával rendelkezik, a kiválasztás *K*(K-1)/2* féleképpen történhet. Ezt követően *K*(K-1)/2* különböző, (2 x m) méretű, kontingenciatáblára Pearson féle khi-négyzet teszt segítségével kiszámolja, hogy milyen „*p*" szignifikancia szinten tekinthetők X_i kiválasztott kategória párjai és Y célváltozó kategóriái függetlennek egymástól. A következő lépésben kiválasztásra kerül az a kontingenciatábla, mely a legmagasabb „*p*" értékkel rendelkezik. Ezt az értéket az eljárás összeveti egy, a modellkészítő által előre lerögzített, $\alpha_{\text{egyesítés}}$ küszöbértékkel (a programcsomagok általában a szokásos 5%-os szignifikancia szintet szokták felkínálni alapértelmezésként). Amennyiben $p > \alpha_{\text{egyesítés}}$ a kontingenciatáblázat X_i kategóriapárja egy új önálló kategóriába kerül egyesítésre. Ebben az esetben X_i eredeti kategóriáinak száma eggyel csökkent, 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 kontingenciatáblákhoz tartozó „*p*" értékek kiszámolásáig.

[1] Folytonos változók esetén az algoritmus alapbeállításként a K=10 értéket ajánlja fel. Ebben az esetben a változó decilisei jelennek meg, mint kategóriák.
[2] Logisztikus regresszió esetében m=2

A kategóriák összevonásának ciklusa mindaddig folytatódik, míg a legmagasabb *„p”* értékkel rendelkező kontingenciatáblára igaza nem lesz a $p < \alpha_{egyesítés}$ feltétel. Ekkor a vizsgált magyarázó változó *(X_i)* esetében a ciklus leáll, és az algoritmus a következő lépésben most már *X* teljes, lehetséges összevonások utáni**,** új kategória-struktúrájára kiszámolja a *„p”* értékét. Az így létrehozott új változó most már alkalmas arra, hogy a prediktív modell lehetséges magyarázó változójaként a modellépítés során felhasználásra kerüljön. A következő ábrán egy konkrét folytonos változó esetében az algoritmus végeredményét[3] látjuk.

3. ábra. CHAID alapú kategorizálás



Egy hazai középvállalati ügyfeleket tartalmazó banki adatbázis esetében a kategorizálandó pénzügyi mutató az *üzemieredmény/árbevétel*. Az ábra téglalapjaiban láthatók a célváltozó (Csőd) lehetséges értékei (0,1) szerinti megoszlások a pénzügyi mutató kategóriái mentén[4]. A legfelső téglalapban (a fa csúcsán) látható, hogy a kiindulási adatbázis 685 fizetőképes és 72 csődös vállalatot tartalmazott. Az ábráról leolvashatók a kategóriahatárok, melyek rendre a következők:

- 0,003 az első és második kategória esetében
- 0,033 a második és harmadik kategória esetében
- 0,051 a harmadik és negyedik kategória esetében

[3] Az ábra az SPSS answer tree CHAID programmoduljának segítségével készült.
[4] A felső téglalapban a teljes adatbázis megoszlását mutatja a „Csőd” változó kategóriái mentén.

Ezek után megnéztük, hogy mi történik akkor, ha más kategorizálási logika mentén alakítjuk ki a kategóriahatárokat. A folytonos alapváltozó eloszlását leíró fontosabb statisztikákból indultunk ki. A következő táblázat tartalmazza az *üzemieredmény/árbevétel* mutatót jellemző leíró statisztikákat:

6. Táblázat. Üzemi eredmény/árbevétel mutató statisztikái

Statistics

N	**Valid**	***757***
	Missing	***0***
Mean		***,0538***
Median		***,0645***
Mode		***,05***
Std. Deviation		***,40277***
Range		***9,17***
Minimum		***-7,58***
Maximum		***1,59***
Percentiles	**10**	***,0031***
	20	***,0196***
	25	***,0252***
	30	***,0328***
	40	***,0508***
	50	***,0645***
	60	***,0809***
	70	***,1044***
	75	***,1195***
	80	***,1378***
	90	***,1961***

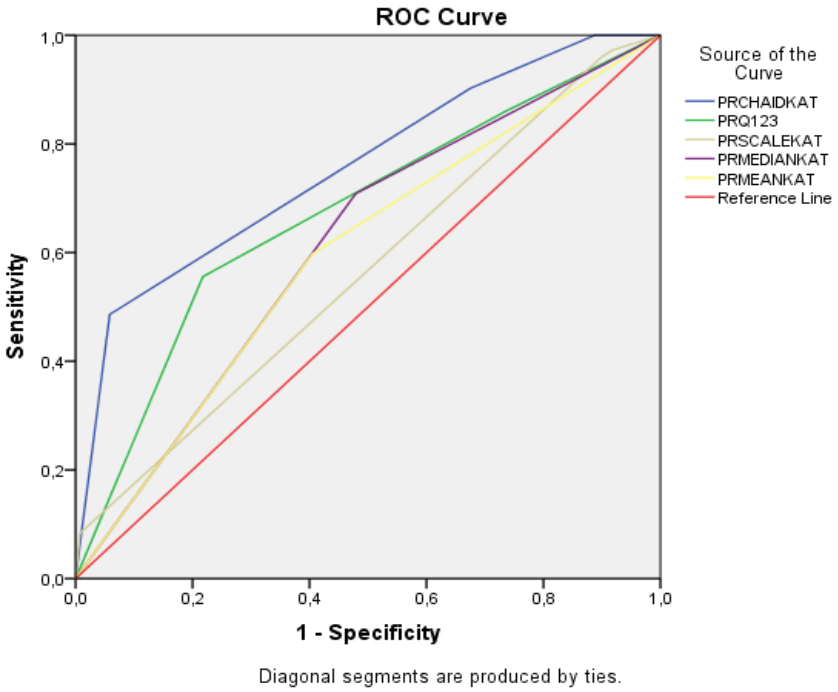
Az alapstatisztikák felhasználásával definiáltunk további négyféle kategorizálási módszert a következő módon:

1. A három kvartilis segítségével meghatározott négykategóriás változó, melyet nevezzünk „Q123KAT” módon
2. Az eloszlást önkényesen négy részre felosztó pontok segítségével meghatározott „SCALEKAT” változó
3. A medián által meghatározott bináris változó, melyet jelöljük „MEDIÁNKAT” jelöléssel
4. Az átlag (mean) segítségével hozzuk létre a „MEANKAT” bináris kategóriaváltozót

Az CHAID-alapú optimális kategorizálás eredményeképpen létrejött változót elneveztük „CHAIDKAT” módon, melynél a kategóriahatárok rendre a már ismert 0,003; 0,033 és 0,051. Vizsgálatunk célja az alternatív kategorizálási módszerek által előállított változók prediktív erejének értékelése az optimális kategorizálási módszer eredményeképpen előálló „CHAIDKAT” változóval szemben. A feladat végrehajtására egyváltozós logisztikus regressziókat futtatunk az egyes változókon, majd ROC görbével és GINI koefficienssel értékeltük az illeszkedést.

Az egyes modellfuttatások egyesített és ROC görbéi és a hozzájuk tartozó GINI mutatók látható az alábbi ábrán és táblázatban:

4. ábra. Egyesített ROC görbe.



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 az alábbi táblázatban foglaljuk össze.

7. Táblázat. GINI értékek összefoglaló táblázata

MUTATÓ	*CHAIDKAT*	*Q123KAT*	*MEDIANKAT*	*MEANKAT*	*SCALEKAT*
GINI	51,4%	34,7%	23%	19,3%	13%

A vizsgálat az adott adatmintán egyértelműen igazolta a CHAID-alapú optimális kategorizálás eredményeképpen létrejövő változó nagyobb prediktív erejét a logisztikus regressziós modellben.

Új kutatási eredmények

Sikerült igazolni, hogy a szélsőséges értékkel rendelkező megfigyelések megfelelő kezelésével a regresszió klasszifikációs ereje növelhető ***(hipotézis 1)***. Ugyanígy kimutatásra került, hogy a folytonos változók esetében a változó kategorizálása még abban az esetben is javíthatja a modell illeszkedését, ha egyébként a a célváltozó és a folytonos prediktor közötti kapcsolat jellege monoton természetű ***(hipotézis 2)***. Ugyanitt bemutattunk egy az illeszkedést befolyásoló, és a disszertációban bemutatott vizsgálat által az adatmintán igazoltan optimális kategóriahatárok meghatározására alkalmas módszert ***(hipotézis 3).***

Az értekezés témaköréből írt, vagy megjelenés alatt álló tudományos közlemények

Magyar nyelvű közlemények

Hámori Gábor: „Fizetésképtelenség előrejelzése logit-modellel”. Bankszemle. 2001 / 1-2. p.65-87

Hámori Gábor: „Chaid-alapú döntési fák jellemzői”. Statisztikai Szemle 79. évfolyam 2001 / 8 p.703-710

Hámori Gábor-Csákány Tibor: „Szofisztikált kockázatkezelési módszerek egészségügyi alkalmazásokban/kórházi környezetben”, Kórház 2012/12 p.18-19

Hámori Gábor: „Érvényesség és korlátok az algoritmikus döntéshozatalban”. Gazdaság és Pénzügy 2015/4-es számába megjelenésre befogadásra került. p.1-11

Hámori Gábor:"Magyarázó változók kezelésének egyes kérdései regressziós modellezés során". Statisztikai Szemle 2016 januári számába megjelenésre befogadásra került. p.1-19

Idegen nyelven közlemények

Gábor Hámori: „Regression based classification models and expert judgement in predictive situations”, Regional and Business Studies 2015 Vol 7 No 1, 51-60

Egyéb

Best Papers, Global Spine Congress organized by AOSpine 2013, Hong Kong, April 4-6, 2013: Tibor Csákány, Gábor Hámori, P.P Varga: "Risk factors for surgical site infection following thoracolumbar spinal operations and a novel risk stratification model using predictive analytics".

Hámori Gábor: „Információszerzés nagyméretű adatbázisokból”, HVG Big Data konferencia, Budapest 2014. október. 28 (konferenciakötet megjelenés alatt)