

Gépi tanulás, predikció és okság a társadalomtudományokban

Muraközy Balázs (MTA KRTK)

Bemutatkozik a Számítógépes Társadalomtudomány
témacsoport, MTA, 2017

Empirikus közgazdasági kérdések

- ▶ Felváltja-e automatikusan a gépi tanulás a "hagyományos" közgazdasági, ökonometriai modellezést?
- ▶ Modern empirikus közgazdaságtan
 - ▶ Fókusza elsősorban oksági
 - ▶ Fő kihívás az, hogy nem kísérleti adatokból oksági hatásokat becsüljön
- ▶ Gépi tanulás
 - ▶ Célja a mintán kívülre előrejelzés
- ▶ Az eltérő cél miatt a két szemlélet nem automatikusan helyettesítő, sokkal inkább sokszor kiegészíthetik egymást

Vázlat

Miért nem ad jó oksági becslést az a modell, ami jól jelez előre?

Amikor a két módszer kiegészíti egymás

- Adatok az internetről és a műholdakról

- Sok kérdés valójában prediktív

- Az oksági elemzésekben is vannak prediktív lépések

Alkalmazás: vállalati EU-támogatások hatása

Következtetések

Miért nem ad jó oksági becslést az a modell, ami jól jelez előre?

Amikor a két módszer kiegészíti egymás

Adatok az internetről és a műholdakról

Sok kérdés valójában prediktív

Az oksági elemzésekben is vannak prediktív lépések

Alkalmazás: vállalati EU-támogatások hatása

Következtetések

Oksági kérdések

- ▶ Oksági kérdések
 - ▶ Növeli a béreket az, ha valaki jobb iskolába jár?
 - ▶ Gyorsabban nőnek-e azok a vállalatok egy támogatás hatására?
 - ▶ Az árfolyam leértékelődése befolyásolja-e az inflációt?
- ▶ Cél egy együttható becslése, ami két változó közötti oksági kapcsolatot mutat
- ▶ Alapvető módszere a regresszió, ami kiszűri a függő változóra ható egyéb tényezőket

”Természetes kísérletek”

- ▶ Kvázi-kísérleti megközelítés
 - ▶ Bizonyos egyedek valamifajta ”kezelésben” részesültek (jobb iskolába jártak, támogatásban részesültek, leértékelődött a valutájuk)
 - ▶ A kezelés és a kimenet közötti korreláció egyszerre tartalmazza az oksági hatást és a ”szelektív torzítást” - más típusú egyének ”választódnak ki” a kezelésre
- ▶ ”Hitelességi forradalom”
 - ▶ Keressünk a kezeltnek olyan kontrollcsoportot, ami a leginkább hasonlít hozzájuk
 - ▶ A becslés lényege, hogy a kezelt csoport kimeneteit (pl jövedelem) összehasonlítjuk a kontrollcsoportéval
 - ▶ Minden olyan változó hatását ki kell szűrni, ami egyszerre befolyásolja a kezelés valószínűségét és a kimenetet

Két kvázi-kísérleti módszer

- ▶ Szakadós regresszió
 - ▶ Akiket pontszámuk alapján pont nem vettek fel az egyetemre szinte pont olyanok, mint akiket épp felvettek - jó kontrollcsoport
 - ▶ Megfigyelhető és nem megfigyelhető jellemzőikben hasonló
- ▶ Párosítás
 - ▶ A rendelkezésünkre álló információk alapján megbecsüljük, hogy ki milyen valószínűséggel részesül kezelésben
 - ▶ Minden kezeltnek ez alapján keresünk párt, ezek kerülnek a kontrollcsoportba
 - ▶ Ha csak a megfigyelt változók befolyásolják a kezelést, akkor jó becslést ad

Gépi tanulás

- ▶ Célja: minél pontosabb előrejelzés a becslésre használt mintán kívül
- ▶ Fő átváltás
 - ▶ Alulillesztés: nem fogja meg az adatban lévő mintákat
 - ▶ Túlillesztés: A zajra is ráilleszti a modellt, így a mintán kívül nem jól jelez előre
- ▶ Sokszínű módszertan: fák, regressziók, neurális hálók stb. és ezek kombinációi
- ▶ Akkor jönnek ki különösen az előnyei, ha nagyon sok változó van

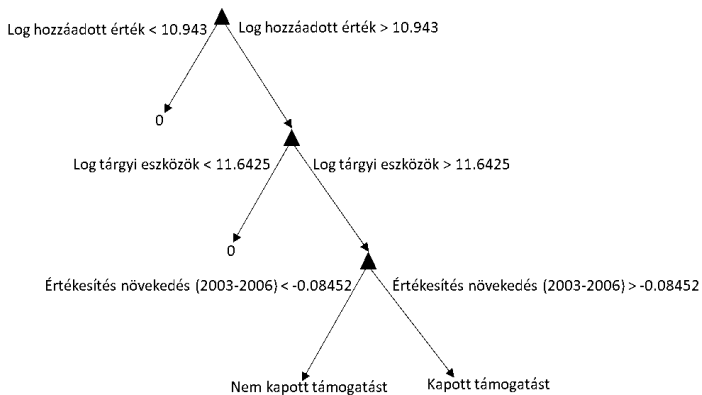
Miért nem ad automatikusan jó oksági becslést a jó prediktív modell?

- ▶ Sok gépi tanulós módszerben egyáltalán nem világos, hogy mi felel meg a becsülni szándékozott együtthatónak (pl. fák vagy neurális hálók)
- ▶ A túlillesztés kiküszöbölésére kidob olyan változókat, amelyek hatását ki kellene szűrni az oksági becsléshez
- ▶ Ha kiszámolható is a szükséges paraméter, az nagyon instabil lehet attól függően, hogy éppen milyen más változók kerültek be a modellbe (Mullainathan and Spiess, 2017)

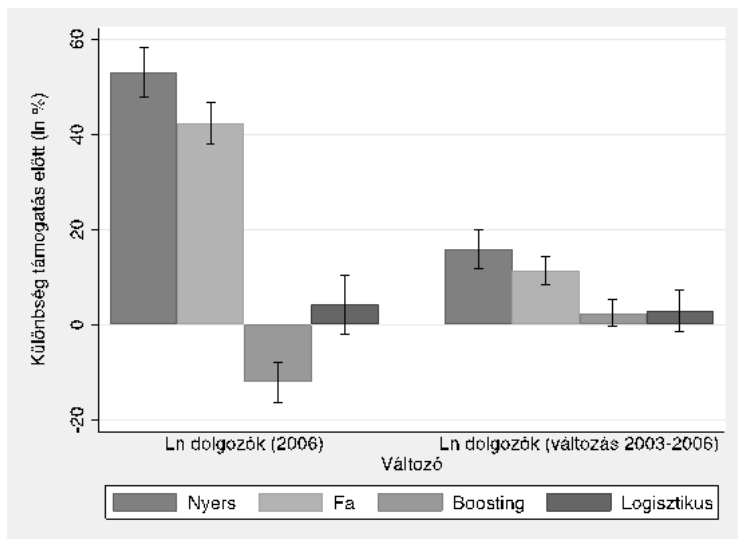
Eu támogatások

- ▶ Az EU-támogatások számottevő részét költötte Magyarország vállalati támogatásokra pl. gépvásárlás, ISO-tanúsítvány bevezetése
- ▶ Ezek fő célja a foglalkoztatás és a versenyképesség növelése
- ▶ Az adatok vállalati szinten tartalmazzák a támogatásokat és a vállalatok mérlegeit
- ▶ A kérdés az, hogy a 2007-ben támogatott vállalatok gyorsabban nőttek-e a következő 3 évben, mint a nem támogatottak
- ▶ Párosításos becslés, különféle gépi tanulási módszerek felhasználásával a propensity score becslésben
 - ▶ Döntési fa: a kontrollcsoport az lesz, ami a fának ugyanazon a levelén van
 - ▶ Boosting: sok döntési fa eredményéből számolható a propensity score
 - ▶ Logisztikus regresszió, a változók gépi tanulóval választva

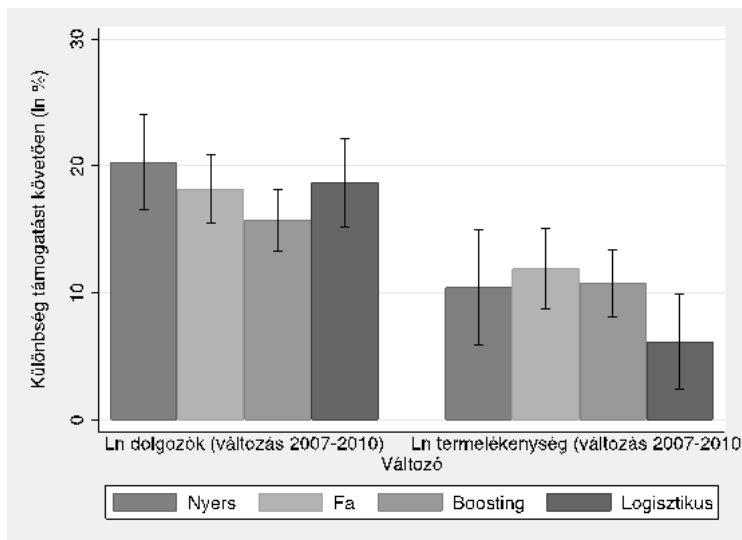
A becsült döntési fa



Kiegyensúlyozottsági teszt



Becsült hatás



Miért nem ad jó oksági becslést az a modell, ami jól jelez előre?

Amikor a két módszer kiegészíti egymás

Adatok az internetről és a műholdakról

Sok kérdés valójában prediktív

Az oksági elemzésekben is vannak prediktív lépések

Alkalmazás: vállalati EU-támogatások hatása

Következtetések

Következtetések

- ▶ A gépi tanulás és az oksági elemzés célja eltérő, nem fogja kiszorítani az egyik a másikat
- ▶ A kettő viszont több területen is kiegészítheti egymást, különösen a big data térnyerésével