Magyar Tudomány 179(2018)7, 1027–1037
DOI: 10.1556/2065.179.2018.7.10

GÉPI TANULÁS, PREDIKCIÓ ÉS OKSÁG A KÖZGAZDASÁG-TUDOMÁNYBAN

MACHINE LEARNING, PREDICTION AND CAUSALITY IN ECONOMICS

Muraközy Balázs
PhD, tudományos főmunkatárs, MTA Közgazdaság- és Regionális Tudományi Kutatóközpont Közgazdaságtudományi Intézet
murakozy.balazs@krtk.mta.hu

ÖSSZEFOGLALÁS

Az empirikus közgazdaság-tudomány módszertanának fókuszában az oksági hatások becslése áll, miközben a gépi tanulás fő erőssége az előrejelzés vagy predikció. Az eltérő célok miatt a két módszertan nem helyettesíti automatikusan egymást. A közelmúlt kutatásai rámutatnak, hogy legalább három területen inkább kiegészítő a két módszer: i) a gépi tanulás adatokat generálhat az oksági elemzés számára; ii) számos közgazdasági kérdés valójában prediktív és nem oksági; iii) az oksági elemzés bizonyos lépései prediktívek.

ABSTRACT

Empirical economics research focuses on estimating causal effects, while the main strength of machine learning is prediction. Because of the different aims the two approaches are unlikely to substitute each other automatically. Recent research has uncovered three promising areas when these two methods may strongly complement each other: i) machine learning can generate useful information for causal analysis; ii) many questions in economics are predictive in nature; and iii) some steps of causal analysis are, in fact, predictive.

Kulcsszavak: gépi tanulás, predikció, oksági elemzés, közgazdaságtan, ökonometria

Keywords: machine learning, prediction, causal analysis, economics, econometrics

ÖKONOMETRIA ÉS GÉPI TANULÁS

Mint ez a tematikus szám rámutat, a társadalomtudományokat alapvetően változtatja meg a rendelkezésre álló adatok és számítási kapacitások minőségi növekedése. Csábító azt gondolni, hogy a Big Data és a gépi tanulás automatikusan megoldja ezeknek a tudományoknak számos ügyét: segítenek előre jelezni a következő nagy válságot, megmutatják, hogy milyen szakpolitikai intézkedéseket kell bevezetni, és, végső soron, a gépek „kitanulják", mi egy-egy iparág vagy

ország gazdasági működésének modellje. A közgazdászok munka nélkül maradnak. Ez persze nem pont így van – ebben az esszében a célom az, hogy árnyaljam a képet.

Mivel is foglalkoznak a közgazdász kutatók? A fő cél a gazdaság működésének megértése, aminek legfőbb eszköze – gyakran formalizált – logikai modellek építése és tesztelése. Az előrejelzés vagy a gazdaságpolitikai tanácsadás ezeken a modelleken alapul. A teszteléshez kulcsfontosságú a változók közötti oksági kapcsolatok felderítése, amelyek általában egyben a modell paraméterei is. Például egy modellparaméter azt mutatja, hogy mennyivel többen tudnak elhelyezkedni egy újfajta foglalkoztatáspolitikai eszköz bevezetésének hatására, és nem azt, hogy mennyivel többen helyezkedtek el abban az évben, amikor bevezették az eszközt (és számos más dolog is történt). Az előbbi hasznos a modelltesztelés és a szakpolitika támogatásához, az utóbbi viszont önmagában kevéssé érdekes.

Az oksági hatások kiszámításának legfőbb nehézsége a társadalomtudományban az, hogy a jelenségek ritkán vizsgálhatók kísérletekkel; így a becsléseknek megfigyelési adatokra kell támaszkodniuk. Az empirikus társadalomkutatás – és annak közgazdasági ága, az ökonometria – alapproblémáját legtöbbször éppen ez jelenti. Az orvosi kísérletekkel ellentétben nem véletlenszerű, hogy kik részesülnek egy-egy „kezelésben" (például járnak jobb iskolába vagy nyernek támogatást), hanem nagyobb valószínűséggel veszik igénybe azok, akik több hasznot remélnek belőle – más szóval önszelekcióra kerül sor. (Az itt leírt szemlélet az elmúlt egy-két évtizedben jellemzi markánsan az empirikus közgazdaságtan főáramát: korábban sokkal kisebb hangsúly helyeződött arra, hogy a becsült kapcsolatok oksági jellegűek-e.)

Ezért nem áll meg az a feltevés, hogy a kezeltek (például a támogatásban részesültek) és a kontrollcsoport csak a kezelés tényében különbözne egymástól, hanem általában más jellemzőikben is eltérnek. Például könnyen lehet, hogy a támogatást kapók jobban informáltak, és támogatás nélkül is jobban boldogulnának. E miatt a két csoport egyszerű összehasonlítása egyszerre tartalmazza a kezelés hatását és az önszelekciót. Az empirikus közgazdaságtan és az ökonometria kutatói az elmúlt években számos olyan módszert dolgoztak ki, amelyek a megfigyelési adatokban olyan mintázatokat keresnek, amelyek a kísérletekhez hasonlóan, természetes vagy kvázi-kísérletként értelmezhetőek.

Ezek a megközelítések a kezelésben részesült csoport mellé valamilyen ötlettel olyan kontrollcsoportot keresnek, amelyik minden fontos (akár megfigyelhető, akár nem megfigyelhető) jellemzőjében a lehető legjobban hasonlít a kezeltekre. Ez a szemlélet olyan súlyt kapott az elmúlt évtizedek empirikus közgazdasági kutatásában, hogy egyenesen egy „hihetőségi forradalomról" beszélhetünk (Angrist–Pischke, 2010).

Egy példa ezekre a kvázi-kísérleti módszerekre a szakadásos regresszió (regression discontinuity), amelyet egy példán keresztül a legkönnyebb megérteni.

Egy alkalmazás például azt vizsgálta, hogyan hat az emberek jövőbeli keresetére az, ha egy adott nagy presztízsű iskolába járhatnak. Természetesen, ha egyszerűen összehasonlítjuk az iskola diákjainak a bérét azokkal, akik jelentkeztek az iskolába, de nem vették fel őket, akkor túlbecsüljük az oksági hatást, mert akiket felvettek, várhatóan eleve jobb képességűek. A szakadásos regresszió abból becsli meg a hatást, hogy összehasonlítja azokat a diákokat, akiket éppen hogy felvettek (épp a ponthatár fölött voltak) azokkal, akiket pont nem vettek fel. Ezek a diákok feltehetőleg nagyon hasonló képességekkel rendelkeznek, de csak egy részük részesült kezelésben (Abdulkadiroğlu et al., 2014). Ugyanez a kérdés megbecsülhető akkor is, ha túljelentkezés esetén sorsolással választják ki a hallgatókat, vagyis tényleg egy természetes kísérletre kerül sor (például Angrist et al., 2012).

A modern empirikus közgazdaságtan fő célja tehát az oksági hatások becslése. A gépi tanulás célja ezzel szemben a predikció, vagyis hogy a rendelkezésre álló magyarázó (vagy prediktor) változók segítségével minél pontosabban előre tudja jelezni egy eredményváltozó alakulását. A siker kritériuma az, hogy az előrejelzésre használt mintán kívül is minél pontosabban tudjon előre jelezni a modell. Az optimális modell megtalálja az egyensúlyt az alulillesztés (túl kevés változó bevonása a modellbe) és a túlillesztés (a túl sok változó nemcsak a mintát, hanem a zajt is leképezi) között. A cél a gyakorlatban általában az, hogy az eljárás a sok lehetőség közül kiválassza azokat a változókat (és a megfelelő függvényformáikat), amelyek javítják az előrejelző erőt; ezt az eljárást nevezzük regularizációnak (például Hastie et al., 2009).

A gépi tanulás különösen hatékony a Big Data környezetben, vagyis olyankor, amikor nagyon sok megfigyelés és változó szerepel az adatbázisban. Egy másik lényeges jellemzője az, hogy maguk a modellek nagyon eltérő szerkezetűek, amelyek a döntési fáktól a regressziókon át a neurális hálókig terjednek. A különböző módszerek eltérő körülmények között hatékonyak, viszont ha ezekben a nagyon eltérő struktúrákban hasonló eredmény születik, az igencsak meggyőző.

Az alapvető célok és megközelítések eltérése több lényeges különbséghez vezet az ökonometria és a gépi tanulás között. Miközben az ökonometriai modellek alapvetően olyan egyenleteket becsülnek, amelyek együtthatói közvetlen kapcsolatban vannak az elméleti modellekkel, ez a gépi tanulásnál nem így van. Először is, sok gépi tanulásos modell nem is néhány egyenletből áll, hanem fákból vagy egymásra épülő egyenletek neurális hálójából. Közel sem világos, hogy ezek hogyan feleltethetők meg a közgazdasági szereplők viselkedését leíró egyenletek paramétereinek. Másodszor, ha ki is lehet számolni ilyen paramétereket, akkor sincs garancia arra, hogy ezek oksági kapcsolatokat írnának le. Lehet, hogy éppen egy olyan változó rendelkezik a legnagyobb előrejelző erővel, amely valójában egy harmadik változó hatását mutatja. Harmadszor, a gépi tanulásos modellekből kinyert együtthatók, még ha léteznek is, gyakran instabilak. A nagyon hasonló előrejelző tulajdonságú (és mély struktúrájú) modellek is gyakran eltérő változó-

kat tartalmaznak, így nagyon eltérő közgazdasági modellek lennének felírhatóak belőlük (Mullainathan–Spiess, 2017).

Ezek a különbségek vezetnek ahhoz, hogy a Big Data és a gépi tanulás nem veszi majd át automatikusan az ökonometria szerepét. Bizonyos alkalmazásokban a két módszer közötti helyettesítés erősebb, másokban viszont ezek a módszerek erősen kiegészítik egymást. Erre következik néhány példa.

ADATOK AZ INTERNETRŐL ÉS MŰHOLDAKRÓL

A Big Data és a gépi tanulás eszközeivel számos esetben valódi társadalomtudományi kísérlet végezhető el, ami miatt kevésbé kell tartani az önszelekció torzító hatásaitól – vagyis olyan empirikus módszertan alkalmazható az adatok elemzésekor, mint akár a kísérleti természettudományokban. Például Ali Ahmed és Mats Hammarstedt (2008) a diszkriminációt vizsgálta a svéd bérlakások piacán, olyan módon, hogy három eltérő fiktív személy nevében kerestek véletlenszerűen albérletet. Az arab nevű férfi lényegesen kevesebb visszahívást kapott, mint a svéd nevű albérletkeresők, míg a svéd nevű nő több visszahívást kapott, mint a svéd nevű férfi. Az internetes platformon pontosan, alacsony költséggel és teljesen kísérleti módszertannal vizsgálható a diszkrimináció szerepe. Azonban, mint látható, ez az adat nem magától jött létre, hanem a kutatók – egy nagyon alkalmas platformon – maguk hozták létre az információkat. Nagyon hasonló kísérletet korábban is el lehetett végezni úgy, hogy a kutatók levélben válaszoltak álláshirdetésekre, amelyben ugyanazt a választ írták különböző bőrszínű emberek nevében (Lavergne–Mullainathan, 2004).

Miközben az előző példában is lényegesen egyszerűbb az interneten kísérletezni, az interneten sok esetben nagyon egyszerűen lehet nagy mennyiségű kísérleti adatot generálni különösen az e-kereskedelemmel kapcsolatos kérdések esetében. Könnyen megbecsülhető például az, hogy mennyivel vesznek többet az emberek egy-egy termékből, ha 10 százalékkal csökken az ára, ha több millió véletlenszerűen kiválasztott potenciális vásárlónak véletlenszerűen 10 százalékkal alacsonyabb árat ajánl fel egy e-kereskedő. Ezeknek a kísérleteknek az egyszerűsége és alacsony költsége ahhoz vezetett, hogy a legnagyobb internetes cégek folyamatosan ehhez hasonló kísérleteket végeznek termékválasztékuk, áraik vagy oldaluk kinézetének optimalizálása érdekében. Az így létrejött adatok sok közgazdasági paraméter becslését könnyítik meg, és számos üzleti alkalmazásuk is van.

Hozzá kell azonban tenni, hogy az interneten létrejött Big Data nagy része nem tudatos kísérletezés eredménye, hanem típusát tekintve ugyanúgy megfigyelési adat, mint amelyeket korábban az ökonometria eszközeivel elemeztek a közgazdászok. Vagyis elemzésükhöz szükség van ökonometriai módszerekre, de egyben tömöríteni is kell az információt elemzés előtt. Például az ökono-

metria alapvetően arra a feltevésre épít, hogy eleve adott valamennyi – nem túl sok – magyarázó változó, és azokból épít a kutató valamilyen elméleti megfontolásokból modellt. Amennyiben a potenciális magyarázó változók száma több ezerre nő, akkor ez az elképzelés nem áll meg, és szükség van arra, hogy a változók kiválasztása vagy tömörítése, valamilyen automatikus eljárással – vagyis gépi tanulással – történjen.

A gépi tanulás segíthet olyan változókat létrehozni, amelyeket magyarázó vagy függő változóként lehet felhasználni az ökonometriai elemzésben olyan kérdések megválaszolására, amelyeket korábban nem vagy csak nagyon korlátozottan lehetett vizsgálni. Például a műholdas információkból rendkívüli földrajzi részletességgel mérhető, hogy mennyi fényt bocsát ki éjszaka az adott térség. Ez az információ – gépi tanulásra támaszkodva – elég pontosan közelítheti a gazdasági növekedést. Különösen fontosak az ilyen mérőszámok a fejlődő országok esetében, ahol a gazdasági aktivitás nagy része rejtve maradhat a hivatalos statisztikák elől. Ezeknek az újfajta változóknak a segítségével sokkal pontosabban vizsgálható, hogy milyen tényezők hatnak egy-egy térség gazdasági növekedésére (Henderson et al., 2012; Donaldson–Storeygard, 2016).

Talán még sokrétűbb közgazdasági és társadalomtudományi lehetőségeket rejt magában a Google Trends-adatok felhasználása. Ezek azt mutatják meg, hogy mennyien kerestek rá bizonyos kifejezésekre, továbbá gazdag területi és időbeli bontásban érhetők el (Varian, 2014; Stephen-Davidowitz, 2017).

SOK KÉRDÉS VALÓJÁBAN PREDIKTÍV

A bevezetőben kiemeltem, hogy a legtöbb közgazdaság-tudományi empirikus elemzés célja az, hogy oksági hatásokat, valamiféle empirikus paramétereket becsüljön meg. Vannak azonban olyan mély közgazdasági kérdések, amelyek lényegüket tekintve prediktívek: például olyan elméleteket tesztelnek, amelyek azt állítják, hogy egy bizonyos változó *nem* rendelkezhet prediktív erővel. Ezeket korábban ökonometriai megközelítésben vizsgálták, de igazából a gépi tanulás sokkal jobban illik az ilyen hipotézisek teszteléséhez. Ha a gépi tanulásos modellek legnagyobb erőfeszítésük dacára sem tudnak érdemben előre jelezni, a változónak gyaníthatóan tényleg nincs prediktív ereje.

Például a pénzügyi közgazdaságtan egyik alapmodellje a hatékony piacok elmélete, amelynek különböző formái azt állítják, hogy múltbeli információk alapján nem lehet előre jelezni a jövőbeli részvényhozamokat. Benjamin Moritz és Tom Zimmermann (2016) amellett érvel, hogy – mivel rendkívül sok változó jöhet szóba egy ilyen vizsgálatnál – érdemes gépi tanulást használni ennek az elméletnek a teszteléséhez. Megmutatják, hogy az amerikai tőzsdeindexet előre lehet jelezni múltbeli adatokból, ami ellentmond a hatékony piacok elméletének.

Miközben a gazdaságpolitika közgazdasági támogatásában legtöbbször fontos az oksági becslés (Hogyan hat a diákok tudására, ha még egy tanárt felvesznek?), vannak olyan területek, ahol a szakpolitikai kérdés prediktív választ igényel (Melyik tanárt érdemes felvenni a rendelkezésre álló információk alapján?) (Chalfin et al., 2016). Ezek az elemzések erősen kiegészítik az oksági vizsgálatokat: a tanárok kiválasztásának módszere befolyásolja a hatást, és a becsült hatás nagysága mutatja meg, hogy mennyi tanárt érdemes felvenni.

AZ OKSÁGI ELEMZÉSEKBEN IS VANNAK PREDIKTÍV LÉPÉSEK

Az utóbbi évek egy fontos gondolata az, hogy az ökonometriai elemzés során számos esetben szerepelnek prediktív lépések. A becslés hatásosabbá tehető akkor, ha felismerjük ezeket a lépéseket, és prediktív módszereket használunk becslésükre. Az, hogy melyek ezek a lépések, és pontosan milyen módszert kell használni becslésükhöz, egy nagyon aktív kutatási terület.

Például a kontrollcsoport létrehozásának egyik gyakran használt módszere a párosítás (*matching,* például Imbens–Rubin, 2015, Part III). Ennek alapötlete az, hogy minden kezelt egyénnek keresünk egy (vagy több) olyan nem kezelt párt, aki a legjobban hasonlít rá megfigyelhető, kezelés előtti jellemzői (például kora, neme, betegségének típusa, lakóhelye) szempontjából. Amennyiben sok megfigyelt jellemző van, akkor a legtöbb kezelt egyénnek nehéz volna olyan párt találni, amely minden egyes változójában eléggé hasonlít rá – túl sok dimenziós a probléma. Ilyenkor információtömörítésre van szükség: a sok magyarázó változóból egyet (vagy keveset) készítünk, és ez alapján végezzük a párosítást. A legtöbbször használt módszer az, hogy a megfigyelt változók alapján megbecsüljük, hogy melyik egyént milyen valószínűséggel kezelik (ezt a valószínűséget nevezzük *propensity score*-nak), és olyan párt keresünk, amelyik a legjobban hasonlít a kezelés valószínűségében. A kezelés hatását úgy kaphatjuk meg, ha összehasonlítjuk minden kezelt egyén kezelés utáni egészségi állapotát a párjáéval – vagyis egy olyan betegével, aki a kezelés előtt a legjobban hasonlított hozzá.

Ez a módszer tehát a következő lépésekből áll – egy orvosi kezelés példáján. Első lépésben megbecsüljük a megfigyelhető jellemzők (kor, nem, korábbi betegségek) alapján annak valószínűségét, hogy az adott ember részesül-e a kezelésben. A második lépésben minden kezelt egyénnek keresünk egy olyan nem kezelt párt, aki a legjobban hasonlít hozzá – vagyis esetében ugyanakkora a kezelés becsült valószínűsége. A kontrollcsoport ezekből a párokból áll. A harmadik lépésben pedig (általában regressziós elemzéssel) összehasonlítjuk a kezelt csoport kimenetét (mennyi idő alatt gyógyult meg) a kontrollcsoportéval. Ez adja az oksági becslést.

Nagyon lényeges meglátás az, hogy az első lépés valójában tisztán prediktív: potenciálisan nagyon sok változó alapján szeretnénk előre jelezni a kezelés valószínűségét. Nincs szó ebben a lépésben oksági hatásokról, és nem fontos az sem, hogy melyik változónak mi az együtthatója. A gépi tanulás tökéletesen alkalmas ennek a lépésnek a végrehajtására. A legtöbb közgazdász azonban – már csak megszokásból is – a „hagyományos" módszereket használja, vagyis végiggondolja, hogy mely változók befolyásolják a kezelés valószínűségét, és azokkal futtat egy regressziót.

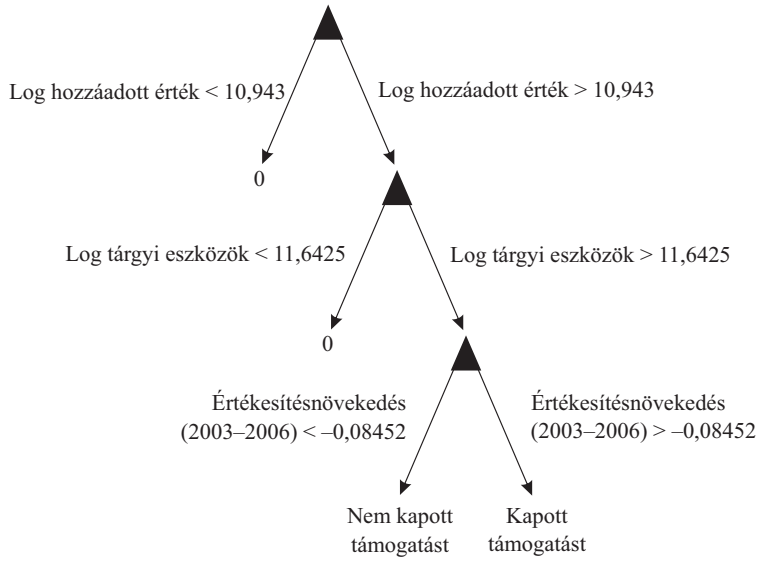
ALKALMAZÁS: A VÁLLALATI EU-TÁMOGATÁSOK HATÁSA

Ebben a fejezetben az előzőekben leírtakra mutatok egy leegyszerűsített példát. A kutatási kérdés az, hogy milyen módon befolyásolta a vállalatok teljesítményét (a dolgozók számát és termelékenységét) az, hogy a vállalat részesült-e valamiféle EU-forrásból származó vállalati támogatásban. Az egyszerűség kedvéért itt csak egy év (2007) támogatásait vizsgáltam.

Az adatok jellege miatt itt párosításos módszert érdemes alkalmazni. Ennek az első lépése, amely alapvetően prediktív, abból áll, hogy minden vállalatra különböző gépi tanulásos módszerekkel megbecsüljük a támogatás megnyerésének valószínűségét. Az előrejelzéshez csak a 2007 előtti információkat kell felhasználni, hogy a támogatás hatása ne jelenjen meg a magyarázó változók között. Ezt követően minden támogatott vállalatnak párokat kell választani (a legközelebbi szomszéd módszerrel), majd pedig kiszámolni, hogy mennyiben viselkedett eltérően a kezelt csoport a párokból álló kontrollcsoporttól a támogatást követő időszakban (2007 és 2010 között).

Az első módszer egy döntési fa volt, eredményét az *1. ábra* mutatja. E szerint azok a vállalatok kapnak támogatást, amelyeknek nagy a hozzáadott értékük és a tőkeállományuk, valamint a múltban is növekedtek. Ez megegyezik a közgazdasági intuícióval is: a nagyobb vállalatok inkább hajlandóak kifizetni a pályázás költségeit, és a múltban is növekvő vállalatoknak inkább vannak olyan növekedési terveik, amelyekhez fel tudják használni az ingyenes forrásokat. Ez a modell 71 százalékban jelzi helyesen előre, hogy melyik vállalat pályázik.

A *boosting* eljárás (lásd például Hastie et al., 2009) több tucat fát épít fel, amelyek „szavazással" döntik el, hogy mi legyen az előrejelzés: minden megfigyelés esetében megvizsgálják, hogy melyik fa milyen előrejelzést ad, és a boosting modell előrejelzése az, amit a fák többsége jelez előre. Ezek a módszerek gyakran hatékonyabban jeleznek előre, mint egy-egy fa. Ez esetünkben is így van: ez a modell 72,1 százalékban jelez előre helyesen. Összehasonlításképp ugyanezekkel a változókkal egy logit modellt is lefuttattam, amelyben a változókat a gépi tanulás algoritmusa választotta ki. A logit modell egy olyan nemlineáris regresszió,



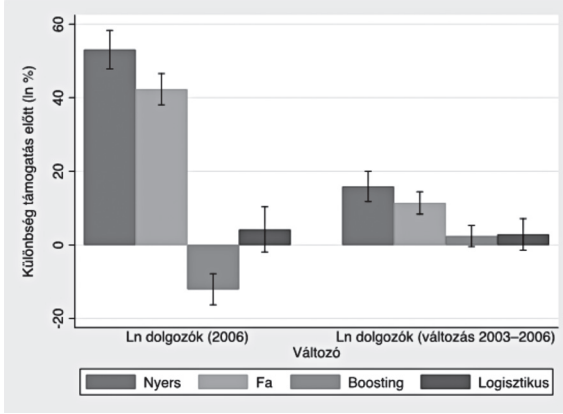
1. ábra. Döntési fa

amely egy fektetett S alakú görbét illeszt a pontokra, így különösen alkalmas az olyan elemzésekre, amikor a kimenet kétértékű (akkor vesz fel nulla értéket, ha a vállalat nem kapott támogatást, és akkor 1-et, ha kapott). Ennek előrejelző ereje 70,8%.

A párosításos eljárásban a kontrollcsoport minőségének egyik fő tesztje a kiegyensúlyozottsági (balancing) teszt, amely egyszerűen megvizsgálja, hogy mennyiben tért el a múltban, kulcsjellemzőit tekintve, a kezeltek csoportja a kontrollcsoporttól. Ha már a múltban is eltérően alakult a teljesítményük, akkor feltehetőleg támogatás nélkül is eltérő mértékben növekedtek volna. A kiegyensúlyozottsági teszt nagyon jó példa a gépi tanulás és az ökonometria közötti kiegészítésre: ez nem a gépi tanulás prediktív erejét méri, hanem azt, hogy közgazdasági szempontból hiteles kontrollcsoport jött-e létre.

Ezt a tesztet szemlélteti a *2. ábra,* amelyben a számok azt mutatják, hogy mennyiben tért el a támogatást megelőzően a kezelt és a kontrollcsoport átlaga.

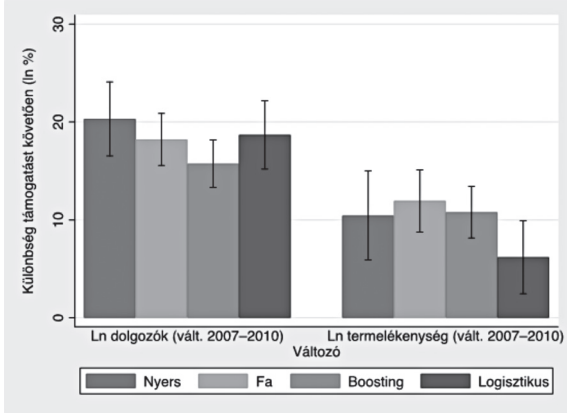
Viszonyítási pontként az első oszlop azt mutatja, hogy mennyiben tért el az eredeti (párosítatlan) mintán a kezelt és a nem kezelt csoport. A kezelt vállalatok lényegesen nagyobbak (több mint 50 százalékkal több dolgozóval rendelkeznek), és gyorsabban nőttek (15 százalékponttal gyorsabban nőtt a dolgozóik száma 2003–2006 között). Ezek a különbségek mindegyik párosítási eljárás után csökkennek. Azonban az eltérő módszerek különböző kontrollcsoportokat generálnak. A döntési fa által generált kontrollcsoport lényegesen nagyobb volt, és gyorsabban is nőtt, mint a kezelt csoport, míg a boosting és a logisztikus regresszió



2. ábra. Kiegyensúlyozottsági teszt: a kezelt és kontrollcsoport közötti különbség a kezelés előtt

kontrollcsoportja nagyon hasonlít a kezelt csoporthoz. Fontos következtetés, hogy nem feltétlenül a legnagyobb prediktív erejű eljárással készült kontrollcsoport hasonlít legjobban a kezeltre.

Végül, a *3. ábra* tartalmazza a hatás becslését, vagyis azt, hogy átlagosan mennyivel növekedtek gyorsabban a támogatott vállalatok 2007–2010 között, mint a kontrollcsoport cégei. Az látható, hogy a különböző kontrollcsoportok alapján kapott eredmények nem térnek el egymástól 5 százalékos szignifikanciaszinten. A kép nagy vonalakban azonban az, hogy a támogatás segítette a vállalatok növekedését, és bizonyos mértékben a termelékenységük fellendítését is. Ez az eredmény mindegyik kontrollcsoport mellett megmarad, nem függ attól, hogyan jelezzük előre a kezelés valószínűségét.



3. ábra. Becsült hatás: a kezelt és kontrollcsoport közötti különbség a kezelést követően

A példa erősen leegyszerűsített. Valójában figyelembe kellene venni azt is, hogy melyik vállalat kapott máskor is támogatást, illetve jobban kellene modellezni a vállalati heterogenitást is. Kiterjedt ökonometriai irodalom – elsősorban a panelökonometria – foglalkozik ezeknek a problémáknak a kezelésével. Ezekből általában azt az eredményt kapjuk, hogy a támogatás hatására a vállalatoknak legfeljebb a méretük nő, de a termelékenységük nem.

KÖVETKEZTETÉSEK

Ennek a tanulmánynak a fő üzenete az, hogy a gépi tanulás és a Big Data nem teszi szükségtelenné az ökonometriát, hanem a két módszer erősen kiegészíti egymást. A megfigyelési adatok elemzésekor szükség van az ökonometria alapszemléletére, de a nagy adatbázisok hatékony kezeléséhez alapvető fontosságú a gépi tanulás eszközeinek alkalmazása.

A rendelkezésre álló új adatok és a gépi tanulás prediktív szemlélete befolyásolhatja a közgazdászok kérdésválasztásait is. Nagyobb szerepet kaphatnak a prediktív elméletek és hipotézisek, illetve az olyan vizsgálatok, amelyek az új adatok nélkül nem is lettek volna megvalósíthatóak. Aktív kutatások folynak a két módszertant egyszerre alkalmazó empirikus eszközök továbbfejlesztésére.

IRODALOM

Abdulkadiroğlu, A. – Angrist, J. – Pathak, P. (2014): The Elite Illusion: Achievement Effects at Boston and New York Exam Schools. *Econometrica*, 82, 1, 137–196. DOI: 10.3386/w17264, http://www.nber.org/papers/w17264

Ahmed, A. M. – Hammarstedt, M. (2008): Discrimination in the Rental Housing Market: A Field Experiment on the Internet. *Journal of Urban Economics*, 64, 2, 362–372. DOI: 10.1016/j.jue.2008.02.004, https://www.researchgate.net/publication/222011648_Discrimination_in_the_Rental_Housing_Market_A_Field_Experiment_on_the_Internet

Angrist, J. D. – Dynarski, S. M. – Kane, T. J. et al. (2012): Who Benefits from KIPP? *Journal of Policy Analysis and Management*, 31, 4, 837–860. DOI: 10.3386/w15740, http://www.nber.org/papers/w15740

Angrist, J. D. – Pischke, J. S. (2010): The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics. *The Journal of Economic Perspectives*, 24, 2, 3–30. DOI: 10.3386/w15794, http://www.nber.org/papers/w15794

Chalfin, A. – Danieli, O. – Hillis, A. et al. (2016): Productivity and Selection of Human Capital with Machine Learning. *The American Economic Review*, 106, 5, 124–127. http://www.hbs.edu/faculty/Publication%20Files/PredictiveHiring_4a3d9f9f-62e9-4dad-ac58-e3c3571d3995.pdf

Donaldson, D. – Storeygard, A. (2016): The View from Above: Applications of Satellite Data in Economics. *The Journal of Economic Perspectives*, 30, 4, 171–198. DOI: 0.1257/jep.30.4.171, http://dave-donaldson.com/wp-content/uploads/2016/10/Donaldson_Storeygard_JEP.pdf

Hastie, T. – Tibshirani, R. – Friedman, J. (2009): Overview of Supervised Learning. In: *The Elements of Statistical Learning*. New York: Springer, 9–41. DOI: 10.1007/b94608_2, https://www.springer.com/cda/content/document/cda_downloaddocument/9780387848570-c1.pdf?SGWID=0-0-45-733855-p173883504

Henderson, J. V. – Storeygard, A. – Weil, D. N. (2012): Measuring Economic Growth from Outer Space. *The American Economic Review*, 102, 2, 994–1028. DOI: 10.3386/w15199, http://www.nber.org/papers/w15199

Imbens, G. W. – Rubin, D. B. (2015): *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press

Lavergne, M. – Mullainathan, S. (2004): Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *The American Economic Review*, 94, 4, 991–1013. DOI: 10.3386/w9873, http://www.nber.org/papers/w9873

Moritz, B. – Zimmermann, T. (2016): *Tree-based Conditional Portfolio Sorts: The Relation between Past and Future Stock Returns*. SSRN. DOI: 10.2139/ssrn.2740751, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2740751

Mullainathan, S. – Spiess, J. (2017): Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, 31, 2, 87–106. DOI: .1257/jep.31.2.87, https://pubs.aeaweb.org/doi/pdfplus/10.1257/jep.31.2.87

Stephens-Davidowitz, S. (2017): *Everybody Lies: Big Data, New Data, and What the Internet Can Tell Us About Who We Really Are*. New York: HarperCollins

Varian, H. R. (2014): Big Data: New Tricks for Econometrics. *The Journal of Economic Perspectives*, 28, 2, 3–27. DOI: 10.1257/jep.28.2.3, https://pubs.aeaweb.org/doi/pdf/10.1257/jep.28.2.3