

Lakásárak becslése regressziós modellezéssel

Molnár Marcell, CEZCRR - marcell.molnar@stud.uni-corvinus.hu

Ebben a projektfeladatban lakásárak előrejelzésére építünk különböző komplexitású regressziós modelleket. A modellek célja az előrejelzés, ezért minden modellt értékelési kritériumban valamilyen formában negatívan fogjuk szülyozni a komplexitást a túltanulás elkerülése érdekében. A kiválasztott modelleket a Test halmazon fogjuk összehasonlítani.

1 Előkészületek, adattisztítás

A magyarázó változók kombinációinak segítségével 8 eltérő komplexitású modell kerül definiálásra. A legjobb prediktív modell megtalálása érdekében 3 különböző módszer alapján választjuk ki külön-külön a legjobb előrejelző képességeket (direkt módszer, indirekt módszer, Lasso regularizáció). A minta (545 megfigyelés, 12+1 magyarázó változó) magyarázó változói között találhatók számszerű adatok (alapterület, hálószobák száma, fürdőszobák száma, emeletek száma, parkolóhelyek száma), bináris változók (fűút melletti, vendégszoba, légkondicionálás, vízfűtés, népszerű fekvés) és egy bináris kódolású változó is a bútorozásra.

Alkotott odellek a magyarázó változóik alapján:

	Model ID	Size	Independent variables
0	1	1	area
1	2	1	parking
2	3	2	stories+prefarea
3	4	3	area+stories+prefarea
4	5	4	area+stories+bathrooms+prefarea
5	6	4	area+bedrooms+mainroad+guestroom
6	7	5	area+hotwaterheating+airconditioning+parking+prefarea
7	8	13	összes magyarázó változó

Table 1: Modeldefiníciók

A modellek között szerepel 2 darab egyváltozós modell, 5 darab "alacsony komplexitású" (2-4 magyarázó változó), egy 5 változós és az összes magyarázó változót tartalmazó is.

2 Módszertan

Az adatokat 80-20 százalékban véletlenül felbontjuk Train és Test halmazokra, ahol kibecsüljük a coefficiensek értékeit, és összehasonlítjuk a modellek teljesítményét.

Az összehasonlítást 3 módon végezzük:

- Direkt módszer: Az adathalmazt tovább osztjuk 80-20 százalékban Train és Validation halmazokra, ahol a Train halmazon végezzük a paraméterbecslést, és a Validation halmazon számított RMSE alapján választjuk ki a legjobb modellt
- Indirekt módszer: A modellek kiválasztása az információs kritériumok alapján történik, külön az AIC és külön a BIC szerint is
- Lasso regularizáció: A kiválasztás a paraméterbecsléssel egyidőben történik, ahol különböző lambda paraméterrel állítjuk be a modell komplexitásának büntetését

Először csak egyszer végezzük el az adatok szétválasztását, és értelmezzük a kapott eredményeket, majd megismételjük 100-szor szimuláció segítségével is.

3 Elsőkörös eredmények

3.1 Direkt módszer

Egyszeri szétosztás után a direkt módszer a 8-adik, legkomplexebb modellt értékeli a Validation RMSE alapján a legjobbnak, viszont látható, hogy akár az 5-ös, akár a 6-os számú modell hasonló hibával becsül a Validation halmazon. Ez ellentmondhat a túltanulás állításainak, de az is lehet, hogy csak az adatok véletlen szétosztásának műve.

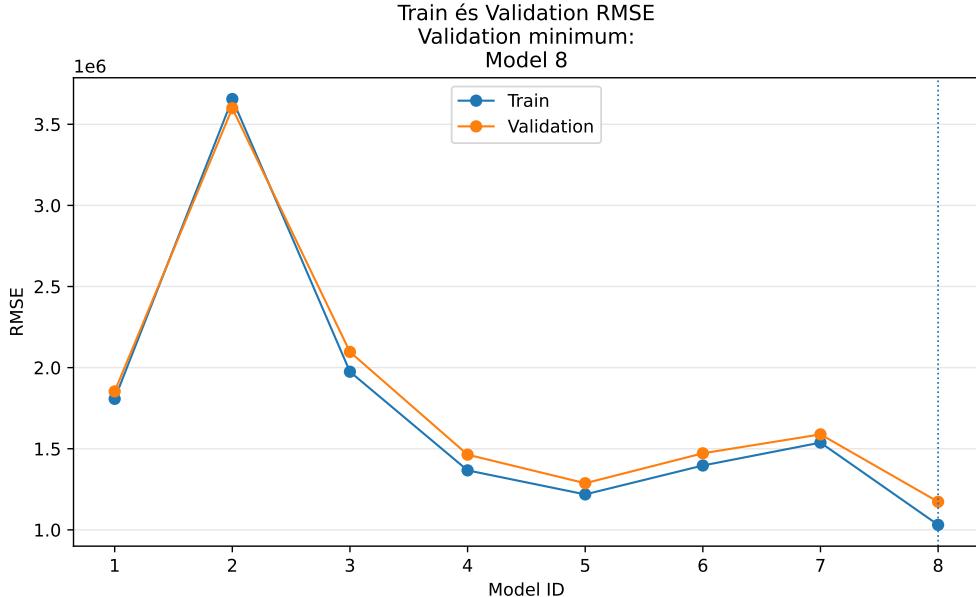


Figure 1: Direkt modell RMSE értékei

3.2 Indirekt módszer

Az indirekt módszer esetén a Train halmazt nem bontjuk tovább két részre, hanem az Akaike (AIC) és Bayesian (BIC) információs kritérium alapján rangsoroljuk őket. Az információs kritériumok büntetik modell komplexitását, viszont meglepő módon ebben az esetben is a legkomplexebb modell került ki győztesnek.

Place	Model ID	IC	Size	—	Place	Model ID	IC	Size
1	8	AIC	13		1	8	BIC	13
2	5	AIC	4		2	5	BIC	4
3	7	AIC	5		3	7	BIC	5
4	4	AIC	3		4	4	BIC	3
5	6	AIC	4		5	6	BIC	4
6	1	AIC	1		6	1	BIC	1
7	3	AIC	2		7	3	BIC	2
8	2	AIC	1		8	2	BIC	1

Table 2: AIC és BIC rangsorok

Ezek az eredmények már érdekesebb lehetnek számunkra. Mivel ebben az esetben is a legkomplexebb modellt kaptuk meg, mint legjobb előrejelző modell, elköpzelhető, hogy a modelldefinícióinkkal. Látszik, a direkt módszer alapján is, hogy akár egy 4 magyarázó változójú modell is hasonló RMSE-t tud produkálni a Validation halazon, szóval elköpzelhető az is, hogy rossz modelleket definiáltunk.

3.3 Lasso regularizáció

A Lasso regularizáció eljárás alapfeltétele az, hogy a magyarázó változókat "leszűkítse", akár 0-ra redukálja a modellben. Az eljárást a legbővebb modellre futtatjuk. A kiválasztott alpha 0-körüli (mintha rendes OLS becslés lenne), amiből arra tudunk következtetni, hogy a legjobb előrejelző modellhez sok magyarázó változóra van szükségünk.

4 Szimulációk

4.1 Stabilitás

Az adatok véletlenszerű felosztását most 100-szor fogjuk elvégezni, hogy egy "stabil" legjobb modellt tudjuk kiválasztani. A direkt módszert és a Lasso regularizációt továbbá keresztpárlatival fogjuk számolni.

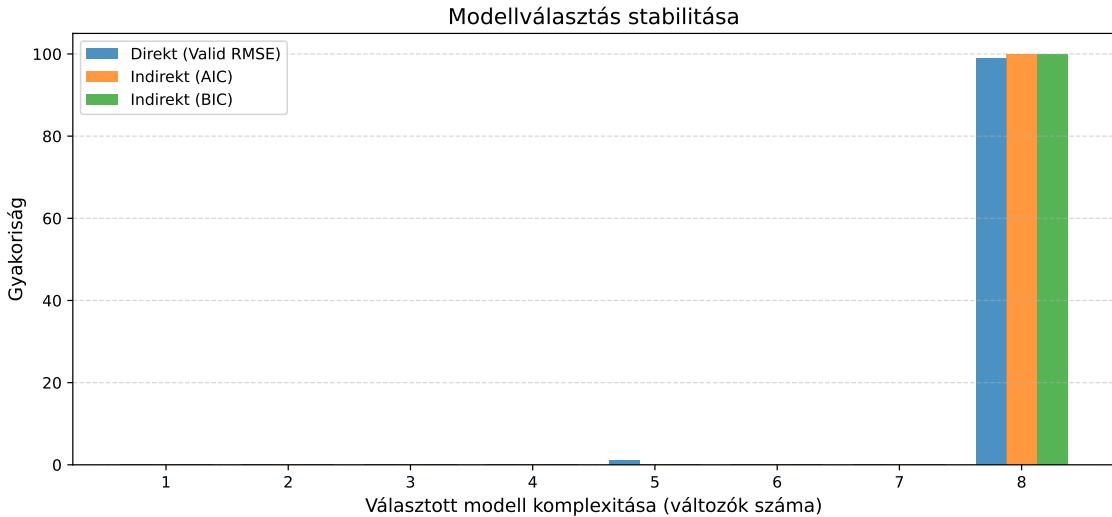


Figure 2: Modellválasztás stabilitása

100-szoros véletlen felosztás után is látható, hogy szinte mindenkor a 8-as, legkomplexebb modellt választjuk minden a direkt és indirekt módszer alapján.

4.2 Holdout teljesítmény

Végső teljesírtékelésként a szimuláció során a Test halazra előrejelzéseket végzünk a kiválasztott modellekkel, és ábrázoljuk boxploton az RMSE értékeket. Látható, hogy az átlagok és a kvartilisek is hasonlóan alakulnak egymáshoz.

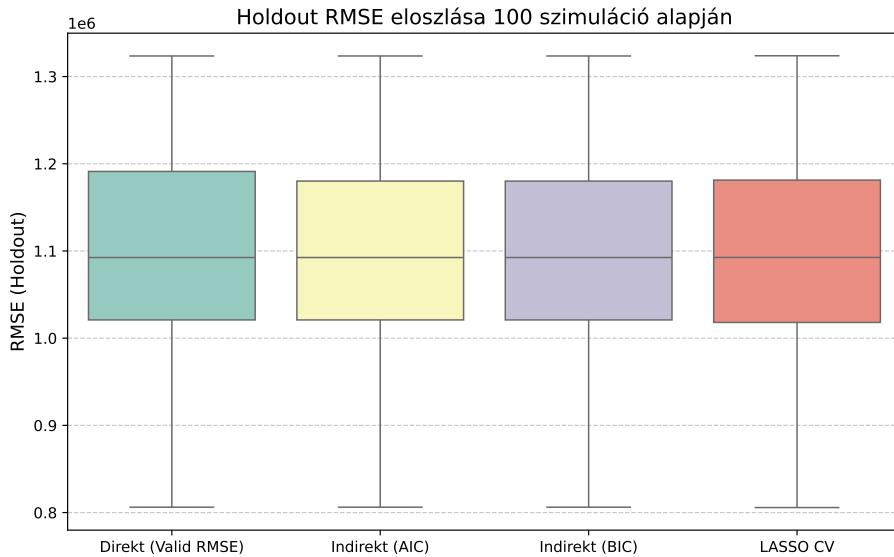


Figure 3: Holdout teljesítmény

5 Értékelés

A szimulációk során szinte minden esetben a legkomplexebb modellünk lett a legjobb előrejelző. Ez azt sugallhatja, hogy egy lakás beárazásához az adott magyarázó változók szinte mindegyike szükséges, azaz mindegyikük hordoz magában új információkat (esetleg legondolkozhatunk a multikollinearitás vizsgálatán is).

Viszont néhány esetben a szimuláció során a 5-ös modell is kiválasztára került (kizárolag a direkt módszer alapján), ami azt sugallhatja, hogy lehetséges egy olyan modellt építeni, ami a legjobb előrejelző képességgel rendelkezik, de mégsem a legkomplexebb. Az 5-ös modell például csak az alapterületet, a fürdők számát, az emeletek számát és a preferált fekvést tartalmazza, mint magyarázó változót.

Egy lehetséges továbbvitel az lehet, hogy megépítjük az összes több, mint 10 magyarázó változót tartalmazó komplex modellt, és ezeket vetjük össze egymáshoz képest.

Melléklet

	Direkt RMSE	Indirekt AIC és BIC RMSE	Lasso CV RMSE	Lasso λ
Átlag	1097324.97	1096453.78	1096704.21	0.001521
Szórás	112776.17	111817.84	112608.01	0.001705

Table 3: Holdout átlag RMSE 100 szimulációra

	Model ID	Set	RMSE	Size
0	1	Train	1807551.37	1
1	2	Train	3655572.72	1
2	3	Train	1974737.18	2
3	4	Train	1366788.10	3
4	5	Train	1218600.66	4
5	6	Train	1396699.33	4
6	7	Train	1537834.73	5
7	8	Train	1030802.25	13
8	1	Validation	1853442.64	1
9	2	Validation	3599062.57	1
10	3	Validation	2096772.02	2
11	4	Validation	1463575.88	3
12	5	Validation	1287483.27	4
13	6	Validation	1471958.40	4
14	7	Validation	1588596.61	5
15	8	Validation	1173102.82	13

Table 4: Direkt módszer Train-Validation RMSE

Konstans -4.57e+05	'area' 2.17e+02	'bedrooms' 1.90e+05	'bathrooms' 9.23e+05	'stories' 4.40e+05	'mainroad' 5.94e+05	'guestroom' 1.78e+05
basement' 3.93e+05	'hotwaterheating' 6.41e+05	'airconditioning' 9.52e+05	'parking' 2.77e+05	'prefarea' 5.89e+05	'dfurnished' 3.88e+05	'dsemifurnished' 2.40e+05

Table 5: Direkt modell béta értékeinek átlaga (100-szoros szimuláció)