Pénzügyi adatelemzés 3. Házi Feladat

Beadási határidő: 2025. 11. 23. 23:59, a Moodle-ben kijelölt helyre feltöltve, 1 darab pdf fájlban, ahol a fájl nevében legyen benne a készítő vezetékneve a következő formátumban: „KOVACS.pdf".
A házi feladatot egyénileg készítsétek el. Nevet, Neptun kódot és email címet a házi feladat címlapján is tüntessétek fel!
A dolgozatot vezetői beszámoló (*executive summary*[1]) stílusban, röviden, tömören, gyorsan áttekinthető módon (max 4 oldal), a készítsétek el. Törekedjetek a világos, lényegre törő megfogalmazásra, az üzenet átadását a dolgozat vizuális megjelenítése is segítse. Használjatok ábrákat, táblázatokat, színeket.
Másolás gyanúja esetén minden érintett hallgató nulla pontot kap!
Ezzel a házi feladattal legfeljebb **15 pont** szerezhető.

Cél: Predikciós modellszelekció demonstrálása lakáspiaci adatokon.
Célváltozó: Price
Készíts egy elemzést a következő kérdésekre válaszolva a következő lakásadatbázis segítségével:
https://www.kaggle.com/datasets/yasserh/housing-prices-dataset?resource=download

Az elemzés során véletlenszerűen 80-20% arányban oszd fel a mintádat work (train - 64% + validation -16%) és holdout (teszt) részre. Építs 8 különböző komplexitású modellt. Figyelj arra modellválasztásnál, hogy ez egy predikciós feladat, és a cél a különböző modell szelekciós megoldások bemutatása, ami akkor tud könnyen prezentálható lenni, hogyha lényeges különbség van az egyes modellek között.

1) **Hasonlítsd össze a következő 3 modellszelekciós eljárást** (melyik modellt választja ki, hogyan alakul a train és validation rész RMSE értéke az egyes modellek esetén). Próbáld meg az eredményeket, táblázat, és ábra segítségével tömören bemutatni.
A) *Direkt:* modellek fit a **train**-en; kiválasztás **validation RMSE** alapján (CV nélkül).
B) *Indirekt:* modellek fit **train+validation**-on; kiválasztás **BIC** (és később AIC) alapján.
C) *Regularizáció:* LASSO fit **train+validation**-on; **λ** kiválasztása **validation RMSE** (később 5-fold CV) alapján.
Mindhárom eljárás **kiválasztott** modelljével készíts előrejelzést a **holdout/test** halmazra, és hasonlítsd össze a holdout RMSE-ket (elsődleges), plusz MAE-t (másodlagos).
2) **Keresztvalidációs kiterjesztés** Ismételd meg az A) és C) eljárást **5-fold CV**-val a kiválasztási lépésben (preprocess *fold-on belül*!); hasonlítsd össze, mennyiben változik a holdout teljesítmény.
3) **AIC vs. BIC** A B) eljárást futtasd AIC-val is; vitasd meg az eltérést (modellméret, illeszkedés, holdout teljesítmény).
4) **Ismétléses értékelés (stabilitás)** Véletlenszerű szétválasztást **100×** ismételd; jelentsd a kiválasztott modellek arányát és a holdout metrikák **átlagát ± szórását** táblázatban és ábrán.

A probléma megoldása során próbálj meg minél több szempontot figyelembe véve válaszolni a kérdésre. Használd az órán tanult koncepciókat!

Mit tartalmazzon a dokumentum:
1–2 ábra: kiválasztási görbék (pl. RMSE vs. modellindex/λ), és a 100× ismétlés eloszlásai/boxplotjai.
2–3 táblázat: train/val RMSE per modell; AIC/BIC rangsorok; holdout eredmények (átlag±sd).
Rövid módszertani indoklás: miért ez a modellkészlet; előfeldolgozás lépések;
Melléklet: részletes modell- és hiperparaméter-leírás

[1] http://en.wikipedia.org/wiki/Executive_summary