M.I. gyakorlat – Beadandó

Készítette: Markó Balázs (J7AJ3X)

A feladat egy ingatlanárakat becslő regressziós modell létrehozása volt. A beadandó leírásában található Boston Housing Dataset-et használtam: [Boston House Prices (kaggle.com)](https://www.kaggle.com/datasets/vikrishnan/boston-house-prices). Az alkalmazás egészét Jupyter Notebook-ban készítettem.

## Megoldáshoz használt technológiák

* Python 3.12.7
* sklearn
* numpy
* pandas
* joblib
* tkinter
* matplotlib
* seaborn

## A felhasznált adatbázis szerkezete

* CRIM: bűnözési ráta
* ZN: nagyobb lakótelkek százaléka
* INDUS: nem lakóterületek aránya
* CHAS: Charles-folyó melletti ingatlan (0 vagy 1)
* NOX: légszennyezési koncentráció
* RM: szobák átlagos száma
* AGE: régebbi épületek aránya
* DIS: távolság a foglalkoztatási központoktól
* RAD: hozzáférési index autópályákhoz
* TAX: ingatlanadó
* PTRATIO: tanuló/tanár arány
* B: Black (faji jelző érték)
* LSTAT: alacsony státuszú népesség aránya
* MEDV: ház árának mediánja

## A kód

Először elő kellett készíteni az adatokat. Sajnos az adatbázis formátuma nem éppen optimális (minden adat be van zsúfolva egy oszlopba), így először azt kellett átformálni. Ez az alábbi kódrészletben látható:

A képen szöveg, képernyőkép, Betűtípus látható

Automatikusan generált leírás

Adatvizualizáció korrelációs mátrix segítségével:

A képen szöveg, képernyőkép, tér, szoftver látható

Automatikusan generált leírás

Adatok normalizálása, ahogy az értékek azonos skálán mozogjanak, illetve tanító- és teszthalmaz szétosztása:

A képen szöveg, szoftver, Betűtípus, Multimédiás szoftver látható

Automatikusan generált leírás

Ezt követően jött a modelltanítás. Négy modellel dolgoztam, egy lineáris regresszióval, egy ridge regresszióval, egy döntési fával és egy random forest-tel. Így a különböző modelleket össze lehet hasonlítani és kvázi „megversenyeztetni”.

Példa a lineáris regresszióra:

A képen szöveg, képernyőkép, Betűtípus látható

Automatikusan generált leírás

MAE (Mean Absolute Error): modell pontatlansága (minél kisebb, annál jobb)

R^2: adatok varianciájának magyarázata (0-1, minél nagyobb, annál jobb)

A többi modellt ugyanerre a mintára írtam meg.

Ezután összehasonlítottam a modellek teljesítményét:

A képen szöveg, képernyőkép, képernyő, szoftver látható

Automatikusan generált leírás

Vizuálisan ábrázolva:

A képen szöveg, képernyőkép, Színesség, Téglalap látható

Automatikusan generált leírás

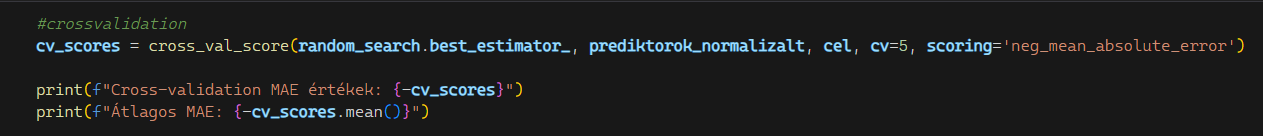
Hyperparaméter-keresés: optimalizálja a modell teljesítményét a legjobb paraméter megtalálásával.

A képen szöveg, képernyőkép, szoftver, Multimédiás szoftver látható

Automatikusan generált leírásA képen szöveg, képernyőkép, Betűtípus, szoftver látható

Automatikusan generált leírás

Cross-validation alkalmazása:



Tényleges és előre jelzett árak összehasonlítása vizuálisan:

A képen szöveg, képernyőkép, sor, szoftver látható

Automatikusan generált leírás

Ezután már csak egy függvényt írtam, ami árat becsül, illetve megjelenít a felhasználónak egy vizuális felületet, ahol adatokat vihet fel.

A képen szöveg, képernyőkép, szám látható

Automatikusan generált leírás