Webshop árelőrejelzés

Papp Júlia

Technológia

Python

Jupyter Notebook

Adatgyűjtés módja

Néhány opció:

- létező adathalmaz (pl. Kaggle)
- API
- **webscraping**

Megfontolások:

- jogilag lehet-e scrapelni, pl Mediamarkt webshopot nem lehet ÁSZF alapján
- nem szabad túlterhelni az oldalt

Választott oldal:

- ebay.com

Termékcsoport kiválasztása

Szempontok:

- egy olyan csoport, ahol az attribútumok összehasonlíthatóak
  - pl ruhákon belül csak női felsőket lenne érdemes
- megfelelő mennyiségű attribútum
  - pl könyv esetében ez elég kevés lenne

Választás:

- Laptopok

Az adatról

Sorok száma:

- Gyűjtött adat: 145 sor
- Szűrések után: 121 sor (duplikátum, túl sok hiányzó érték, kilógó ár adat)

Változók száma:

- Összes lehetséges feature száma: 128
- Szűrés után: 11 (túl sok hiányzó érték, nehezen értelmezhető, túl egyedi értékek)
  - Condition_categ
  - Processor_categ
  - SSD Capacity_ordinal
  - RAM Size_ordinal
  - Processor Speed_ordinal
  - Screen Size_ordinal
  - Brand_categ
  - Storage Type SSD
  - Maximum Resolution Full HD or better
  - GPU_categ
  - Operating System_categ

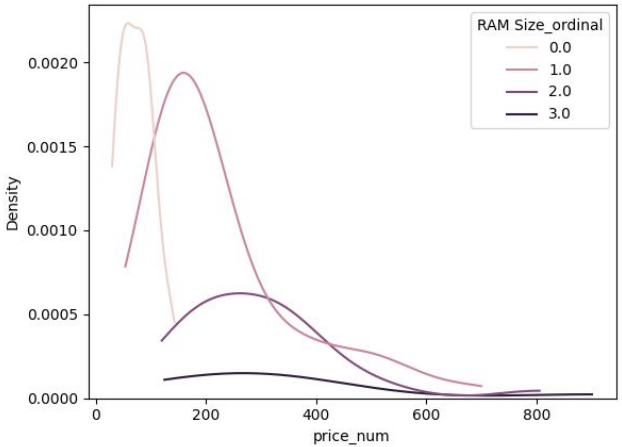
Adatelőkészítés

Minden változó szöveges volt

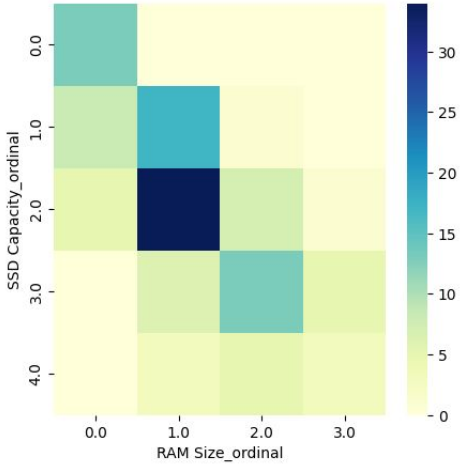
- tisztítás
- ordinális változó készítése, ha van sorrendiség - kategóriánként legalább 5 elem
- kategorikus változó készítése (one-hot-encoding) - kategóriánként legalább 5 elem
- boolean készítése
- hiányzó érték kitöltés

EDA

Változók kapcsolata az árral:
Legtöbb helyen látszott valamilyen minta



Változók kapcsolata egymással:
SSD Capacity és RAM Size között volt a legerősebb kapcsolat



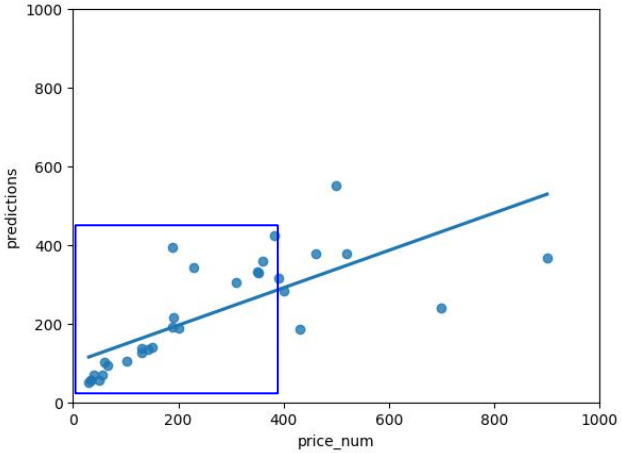
Modellezés

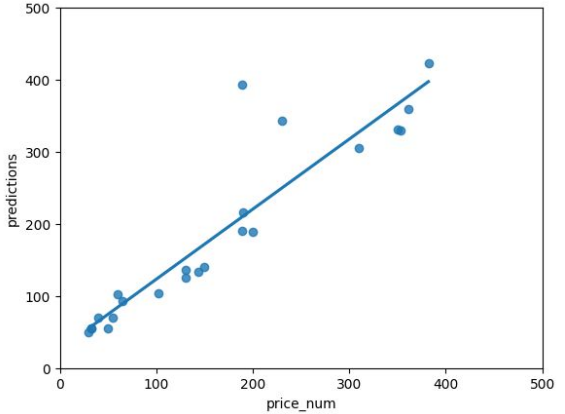
- Train-test split
- Standardizálás
- RandomForestRegressor
- Gridsearch
- MAE (Dummy regresszorhoz képest)
- Feature importance (ez alapján változók további szűrése)
- Predikció összehasonlítása a tényleges árral vizuálisan

Eredmény

- MAE a dummy regresszorhoz képest jobb
- Változók szűkítése segít
- Gridsearch eredmény kicsit rosszabb lett

Az ábrákból látszik, hogy az alacsonyabb áraknál nagyon jól teljesít, a magas árak esetén problémásabb.





Lehetséges következő lépések és ötletek

Adatelőkészítés:

- Végignézni, hogy van-e esetleg olyan feature, ami valójában megegyezik egy másikkal, csak kicsit más a neve
- Screensize mint kategorikus változó
- A "Seller Notes" változóra rá lehetne esetleg menni valamilyen szövegelemzéssel, akár kulcsszavakat vagy negatív kifejezéseket kigyűjteni, hátha van hozzáadott értéke.

Modellezés:

- Más model kipróbálása
- Más paraméterekre gridsearch
- Másfajta search (pl randomsearch)
- Cross validation
- Ránézni, hogy miért a magasabb áraknál teljesít rosszabbul, min lehetne javítani