Feladat

A feladat során a www.hasznaltauto.hu oldal személyautó, kishaszongépjármű és haszongépjármű hirdetéseiből készítettem egy adathalmazt, amelyből a kiugró értékek eltávolítása és a hiányzó elemek pótlása után készítettem egy Tableau vizualizációt. A dataset elkészítése sok időbe telt, így az árbecslő modellre csak néhány lehetőséget próbáltam ki bármilyen paraméter beállítása nélkül. Az így készült modell csak olyan autók árát tudja megbecsülni, amelyről minden információnk megvan (márka, típus, üzemanyagtípus, évjárat, lóerő, futott kilométerek).

Dataset (*hasznaltautohu_db.csv*)

Az adatbázis alapját a **datagen.py** kód szedi össze. Ez a script az oldalak HTML kódjából vonja ki az adatot illetve a következő oldal URL-jét, amivel aztán tudja folytatni az adatgyűjtést. A script opcionálisan három szálon fut (a három kategória miatt), ezzel kb 5 óra alatt el is készült a kezdeti adatbázisunk, ami a képen aláhúzott információkat fogja tartalmazni illetve az autó kategóriáját (személy, kishaszon, haszon).

A script a túl hiányos hirdetéseket (forintos ár nélküli, üzemanyag típus nélküli, stb.) átugorja. Megengedjük viszont, hogy a hengerűrtartalom illetve a futott kilométerek hiányozzanak, hiszen ezeket elég jól tudjuk pótolni átlagokkal. Elektromos autó esetében előbbi fixen nulla, más hajtású járműnél meglepő módon nem hiányzott a hengerűrtartalom egyszer sem. A futott kilométerek esetében a 2021-es autóknál a hiányzó adatot 0-val pótoltam.

2021-es évjáratútól különböző autók hiányzó futott kilométer jellemzőjét a **na_handler.py** pótolja az adott év autóinak futott kilométer átlagával.

Ezt követően a Tableau-ban kerestem gyanús értékeket vagy elírt/rosszul megadott autómárkákat stb.. Ezeket sajátkezűleg törölgettem ki vagy írtam át a .csv-ben:

- 4500 lóerős Daewoo, 500M Ft-os Fiat 500, 123456789 km-et futott maruti és hasonlók törlése
- REPLIKA/EGYÉB/EGYEDI automárkák eltávolítása
- két IVECO márka összevonása
- Hibrid (Benzin), Hibrid (Dízel) összevonása Hibriddé.
- LAND -> LAND ROVER átírás

A régi autók esetében nagyon sok volt a túlárazott illetve hiteltelen kilométeróradatokkal rendelkező hirdetés. A **rm_susp.py** törli az 1982 előtti járművek hirdetéseit, a 2019 előtti olyan autókat, amelyekben 1000-nél kevesebb kilométer van illetve a járműveket amelyek 1.500.000 km-nél többet jártak; legtöbb valószínűleg csak tessék-lássék módon megadott adat volt a hirdetésfeltöltő részéről és nem valódi értékek.

A dataset összeállítása és a végül elkészített vizualizáció (ha nem számolom az időt amíg a kódok futottak) ~6.5 órát vett igénybe.

Viz: https://public.tableau.com/profile/m.t.kis#!/vizhome/hasznaltauto_dashboard/Dashboard1?publish=yes

Model árbecslésre (*model.ipynb*)

Erre a feladatrészre kevés időm maradt, amit itt végül csináltam az az, hogy a kategorikus oszlopokat helyettesítettem számokkal, megkevertem és feldaraboltam a datasetet egy train és egy validációs setre. Három féle modelt próbáltam ki mindenfajta paraméterállítás nélkül: Egyik egy lineáris regresszor volt, amolyan baseline-ként, hogy megtudjuk mi az aminél biztos, hogy jobban kell teljesítsünk (az ár egyáltalán nem lineáris függvénye pl az évjáratnak vagy a futott kilométernek). A második egy decision tree volt; ez nem követel skálázást illetve jól elbánik a nemlinearitásokkal. A harmadik kipróbált modeltípus pedig a random forest, ami végeredményben a legjobb modellnek bizonyult. A notebook végén néhány más használtautós site-ról vett példára kipróbáltam a modell jóslatait. A pár példán én azt vettem észre, hogy a (kis)haszongépjárműveknél a tévedés viszonylag nagy.