Big Data elemzési módszerek házi feladat
Eredmények összefoglalása

Adatbetöltés, adatelőkészítés és feature engineering

Legelőször betöltöttük az adatokat .parquet formátumban, és ellenőriztük az oszlopok adattípusát és a sorok számát. 2017-ben több mint 110 millió sor volt, ezért ezt mintavétellel szűkítettük le (250 000 adatsoronként hónapokra lebontva).

Ezt követően elvégeztük a szükséges adattisztítási feladatokat. A trip_amount esetében nincs jelentős számú kiugró érték. Voltak azonban olyan oszlopok, amelyekben jelentős számú hiányzó értékek voltak. Ezen túlmenően, ahol szükséges volt, a meglévő jellemzőket megfelelő formátumba konvertáltuk, és új változókat hoztunk létre.

A modellezés előkészítéséhez még néhány adatelőkészítési feladatot el kellett végezni. Minden nem numerikus új változót numerikusra alakítottunk át. A modellépítéshez nem használható változókat kizártuk. Ezenkívül a variancia csökkentése érdekében a dask_ml programmal standard scaler-rel normalizáltuk az adatkészletet (targeten kívül).

Felderítő adatelemzés

A felderítő elemzés során mindkét évi adatsorra leíró statisztikát. Az utazások gyakoriságát az év első hetére nézve hisztogramon ábrázoltuk, amin szépen kivehető a csúcsiszakok kiugró utazásszáma. Hisztogramon ábrázoltuk továbbá az utazások hosszát, amelyből látható, hogy a rövidebb utazások dominálnak. Scatter ploton ábrázoltuk a tip_amount és a trip_distance kapcsolatát, amiből látható, hogy a legtöbb utazás a 0-50$ tip_amount és a 0-50 mérföld trip_distance által meghatározott négyszögbe esik. Mindezeken kívül ábrázoltuk a tip_percent-et a hét napjaira vonatkozóan, az utasszám, valamint az egyes zónák alapján.

Csináltunk egy-egy korreláció mátrixot is a két adatsorra, amiből látszik, hogy 2017-ben a tip-amount-tal leginkább a weekday, a trip_distance és a tolls_amount korrelált. Ezzel szemben 2021-ben a változók sokkal kevésbé korrelálnak már egymással. A tip_amount-tal leginkább az mta_tax korrelál, de az is csak közepesen erősen (0,35.).

Big data adatvizualizáció

A vizualizáció célja, hogy egy interaktív hőtérképet kapjunk, amely segítségével információt nyerhetünk az adatok helyalapú eloszlásából. Ebben a projektben a borravaló mértékét vizsgáljuk, ezért készítettünk egy olyan hőtérképet is, amely az átlagos borravalók értékét mutatja térképes nézeten.

Mivel a feladatkiírásban szereplő linken ehhez nem találtunk adatot, a következő linken elérhető forrás adatokat használtuk a plotoláshoz:

https://s3.amazonaws.com/datashader-data/nyc_taxi_wide.parq

Modellépítés a feature importance meghatározásához

Végül mindkét évre vonatkozóan két modellt futtattunk: lineáris regressziót és random forest modellt. Kipróbáltuk az ensemble learnert és a dask_ml néhány modelljét is, de semmi értelmezhetőt nem tudtunk kihozni belőlük.

Kipróbáltuk a korrelációs mátrix eredményei alapján leszűrni a változókat, de nagyon rossz pontosságú modelleket kaptunk. Ezért hagytuk az eredeti változókészletet. A modellben sok olyan változó szerepelt, ami között triviális volt az összefüggés, ezért viszonylag magas pontosságot értünk el. Ezeket a változókat leszámítva 2017-ben a lineáris regresszió a trip_distance-et, a RateCodeID-t és a tolls_amount-ot emelte ki a koefficiensek alapján, míg a random forest-nél a trip_distance, a duration és a PULocationID / DOLocationID. 2021-ben a random forest esetében meglepő módon nem történt változás, míg a lineáris regresszió esetében az mta_tax és az extra költségtényezők értéke felértékelődött.