Tema 2

Analiză de regresie pentru oferte de vânzări auto
Mineritul datelor și analiza datelor (MDAD)

2024 - v1.0
Deadline: 17.05.2024 (23:59)

Descriere generală

Pentru această temă va trebui să antrenați și evaluați modele de regresie pentru un set de date cu oferte de vânzări auto. Rezolvarea corectă și completă a datelor presupune următoarele:

1. Citirea și încărcarea datelor din fișierul la dispoziție
2. Transformarea datelor
    a. Din format JSON în formatul necesar pentru restul pipeline-ului
    b. Descoperirea și corectarea erorilor care au apărut din procedura de colectare
    c. Adăugarea sau eliminarea de coloane (acolo unde este cazul, de exemplu prin transformarea celor existente)
3. Stocarea datelor într-un sistem/format ales de voi
4. Antrenarea unor modele de regresie folosind datele de antrenare puse la dispoziție
5. Evaluarea modelelor pe datele de test furnizate
6. Încărcarea fișierului de test în *leaderboard* pentru analiza de performanță
7. Prezentarea soluției implementate într-un raport tehnic

Setul de date

Pentru această temă veți avea la dispoziție un set de date cu oferte de vânzări auto pentru care veți aplica algoritmi de regresie pentru a estima prețul unei oferte. Setul de date îl găsiți în fișierul *auto_train.json*. Veți împărți datele din fișier în 2 subseturi, antrenare și validare. Va fi nevoie însă să prelucrați datele înainte de antrenare pentru a avea rezultate acceptabile. Exemple de pre-procesări ar fi:

- Eliminarea unităților de măsură
- Verificarea separatorului pentru zecimale / mii
- Convertirea tuturor valorilor din anumite coloane în valori de același tip (de exemplu, dacă toate valorile în afară de una singură sunt numere)
- Extragerea de informații din valori compuse (anumite coloane au valori care conțin multiple informații)
- Imputarea valorilor lipsă

Leaderboard

Pentru a verifica și compara performanțele modelului vostru veți putea încărca rezultatele obținute de modelele voastre pe platforma ***ML Leaderboard*** pe care o găsiți la adresa: http://141.85.224.171/

Performanțele se testează folosind Root Mean Squared Error (RMSE), adică eroarea pătratică medie.

Pentru a vă crea un cont veți introduce adresa voastră de email **@upb.ro** și un nume de utilizator care va apărea public în tabelul de pe platformă. După introducerea detaliilor veți primi un email ce conține un link pentru confirmarea contului.
Folosiți butonul **UPLOAD** din tabul ***My Submissions*** pentru a încărca o soluție nouă. Soluția va fi un fișier *.csv* cu coloanele: **id, value**. Valoarea **id** o găsiți în fișierul de test pentru fiecare exemplu în parte, iar **value** este valoarea prezisă de modelul vostru pentru exemplul respectiv.

Puteți încărca o soluție nouă ori de câte ori doriți, rezultatele voastre de pe parcurs vor rămâne în pagina ***My Submissions***, iar cel mai bun rezultat va fi afișat public în pagina ***All Submissions*** alături de rezultatele colegilor.

Algoritmi de regresie

Pentru realizarea regresiei veți antrena diverși algoritmi. Puteți utiliza orice algoritm de regresie doriți însă ne așteptăm să regăsim printre aceștia cel puțin algoritmii discutați la curs: Regresie Liniară (cu și fără regularizări), Arbori de Regresie, Random Forest, ExtraTrees, AdaBoost, Gradient Boosting. Este de așteptat ca pentru acești algoritmi să căutați cei mai buni hiperparametri (prin mai multe rulări sau prin tehnici de căutare a hiperparametrilor). Evident, pentru această căutare veți avea nevoie să apelați la metode de tipul cross-validation. Modul în care ați realizat căutarea și rezultatele obținute pentru fiecare set de valori de hiperparametri testați se vor regăsi în raportul tehnic pe care îl veți preda odată cu tema.

Livrabile

Pe lângă rezultatele evaluării va trebui să puneți la dispoziție atât *codul sursă* utilizat pentru încărcarea și analiza datelor, antrenarea și evaluarea modelelor, precum și un *raport tehnic* în care veți descrie modul în care ați realizat EDA-ul, transformarea datelor, antrenarea modelelor și evaluarea rezultatelor. Alternativ, dacă rezolvați tema folosind notebook-uri (jupyter) puteți adăuga aceste descrieri în notebook. Veți încărca pe Moodle o arhivă *.zip* cu cu numele vostru și prefixul

*Assignment_2_* (e.g. Assignment_2_Ionel_Popescu.zip) în care veți include codul sursă și raportul tehnic.

Bonus

Top 5 teme din *leaderboard* vor primi un bonus de 0.25p adăugate direct la nota finală (a cursului) cu condiția rezolvării corecte a temei (nu se acceptă soluții aleatoare norocoase sau „date în bob”) și a respectării codului de etică al facultății.

Succes!