CARSTOIU Daniel-Petru 313CA

Proiect PCLP3

Construirea si explorarea unui dataset tabelar

Problema rezolvata de mine este de regresie. Aceasta consta in prezicerea rezultatelor unor student bazandu-ne pe o serie de factori.

Am ales generarea sintetica a unui dataset. Consideram 8 caracteristici mai mult sau mai putin importante despre studenti:
1. Genul (Masculin / Feminin) = este un sir de caractere cu o sansa configurata de 51% ca studentul sa fie barbat si 49% sa fie femeie

2. Facultatea (5 optiuni) = valoare categoriala cu probabilitate configurata de 25% ca studentul sa fie de la facultatea de “Computer Science”, 25% la “Mathematics”, 20% la “Psychology”, 20% la “Electrical Engineering” si 10% la “Nuclear Sciences”

3. Age = numar intreg random intre 18 si 25

4. Bursa (true / false) = 30% sanse sa aiba bursa, 70% sa nu

5. Ore de studiu / saptamana = numar real, in medie 10 cu o deviatie de 5, rotunjit la o zecimala

6. Prezenta = numar real, in medie 70% cu o deviatie de 10%, rotunjit la o zecimala

7. Nota anterioara = numar real, in medie 7.5 cu o deviatie de 1.1, rotunjit la doua zecimale

8. Nota finala (VARIABILA TINTA) = numar real obtinut dintr-o formula gasita adecvata de catre mine. Formula este urmatoarea: nota anterioara * 0,7 + nr de ore * 0,15 + prezenta * 0,02 + 0,3 daca are bursa (+ 0 daca nu are)

Dupa generarea datelor am adaugat cateva valori lipsa, 5% mai exact pe cele 3 coloane numerice importante (ore studiate, prezenta, nota din trecut).

Apoi, am salvat aceste date intr-un csv denumit sugestiv “unprepared_data.csv”.

In urmatorul fișier sursa, data_prepare.py, incarc datele din csv-ul anterior si tratez mai intai valorile lipsa prin imputare, strategia fiind de a completa datele lipsa cu mediana. Impart datele intr-o proportie de 28.5% din totalul de sample-uri (2000) pentru datele de test, iar restul celor de antrenare. Considerand ca am ales 2000 de teste (desi fisierul sursa date_generate poate genera oricate din moment ce primeste ca parametru numarul de sample-uri), 28,5% inseamna 1430 de teste de antrenare si 570 de testare ceea ce este destul de asemanator cu proportia 500-200 din cerinta. Dupa impartirea datelor, acestea sunt salvate in csv-uri separate “train.csv” si “test.csv”.

In al treilea si ultimul fișier sursa, data_analysis.py, am facut o analiza exploratorie a datelor. Am folosit describe() pentru a genera statistici descriptive despre toate campurile (indiferent ca unele statistici erau mai mult sau mai putin relevante).

Fac analiza distributiei variabilelor prin a genera o sumedenie de histograme pentru variabilele numerice si de countplot-uri pentru variabilele categorice.

Pentru detectarea outlierelor am folosit regula intercuartilica (IQR). Extrag primul cuartil (25%) si al treilea (75%) si calculez intervalul IQR = primul - al treilea. Definim limitele printr-o formula si detectam outlierele prin a ne da seama daca fac parte din intervalul marginit de limita inferioara si cea superioara.

Am analizat corelatiile pt variabilele numerice cu ajutorul heatmapurilor. De asemenea, am generat si ploturi pt a evidentia relatia dintre campul tinta (scorul din examen) si restul campurilor. Am generat atat scatterploturi, cat si violinploturi.

Pentru antrenarea si evaluarea modelului am ales regresia liniara, iar dupa separarea datelor (campul tinta si restul), am antrenat si prezis notele. La o evaluare atenta, rezultatele sunt imbucuratoare:

MAE: 0.15 (asadar, in medie, predictiile sunt la doar 0.15 puncte fata de scorul real)

RMSE (root mean squared error): 0.3 confirma iar o eroare mica

R^2: 0.92 (foarte bun, indica ca modelul explica 92% din variatia notelor)
Dupa aceea am generat si niste ploturi de eroare (predictions_vs_actual)

Am folosit git pt a putea controla versionarea acestui proiect, toate fisierele sursa, csv-urile si ploturile sunt postate pe repo-ul meu public.

Link github: https://github.com/Dani340/Proiect-pclp3