Temă predicție

STUDENT:

BURSUC ALEX-GEORGE

1. Introducere

Proiectul constă în implementarea unei soluții folosind tehnici de machine learning capabile să prezică prețul pentru o listă de mașini second-hand pe baza mai multor informații precum: marca, modelul, anul fabricației, numărul de kilometri, transmisia, dotări, etc.

2. Caracteristici ale setului de date

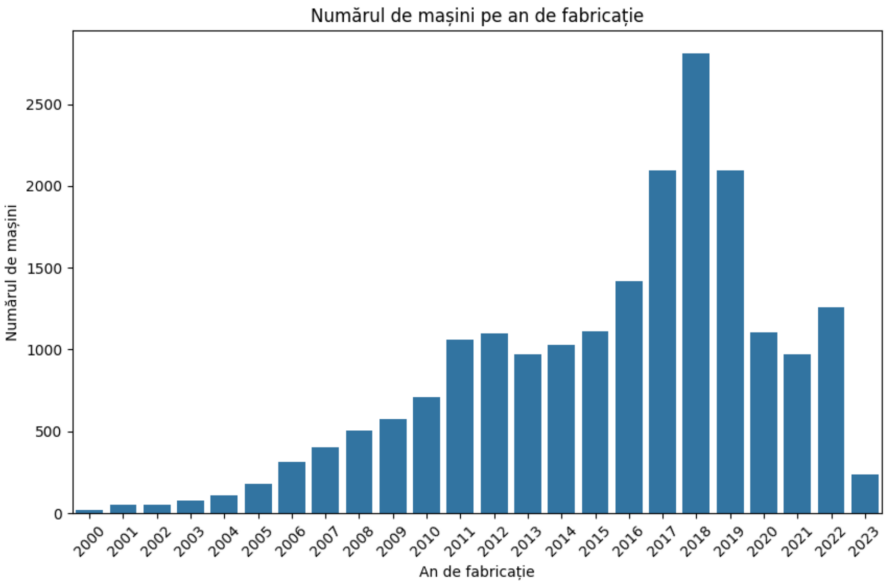
Înainte de a începe procesul de implementare al soluției s-au realizat mai multe ploturi (atăt pe setul de antrenare cât și pe setul de testare) cu scopul de a extrage caracteristicile ce țin de seturile de date:

a) Relația dintre numărul de kilometri și preț în funcție de marcă – se poate observa faptul că majoritatea mașinilor cu un număr mic de kilometri au un preț mai ridicat. Acest lucru ne poate ajuta la validarea modelui (spre exemplu dacă modelul prezice un pret scăzut pentru o mașină cu cu puțini kilometri putem trage concluzia că modelul nu ține cont de aceste caracterisici)

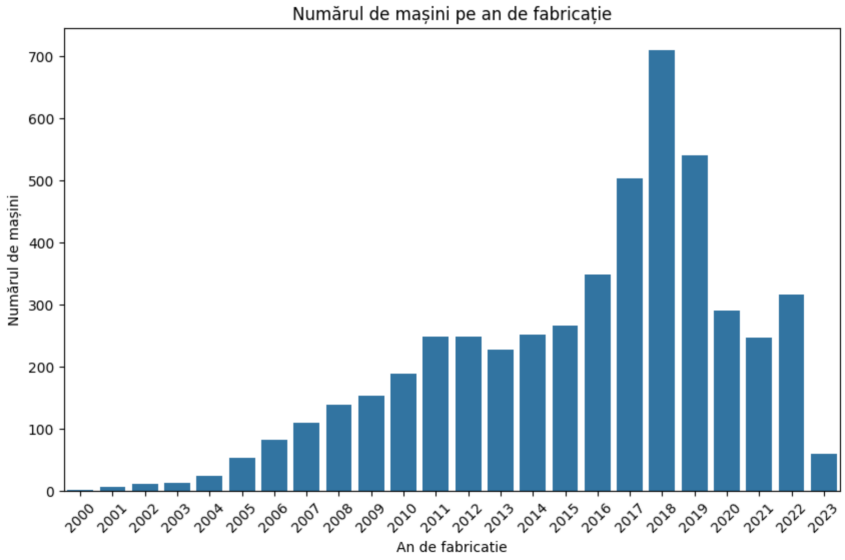
(train.json)

b) Numărul de mașini pe an de fabricație – Acest lucru ne indică numărul de mașini per an de fabricație. Putem identifica extremele, iar acest lucru poate implica un proces suplimentar in ceea ce priveste implementarea modelului. Spre exemplu,

masinile vechi sau masinile noi (extremele) pot influenta major procesul de antrenare al modelului.



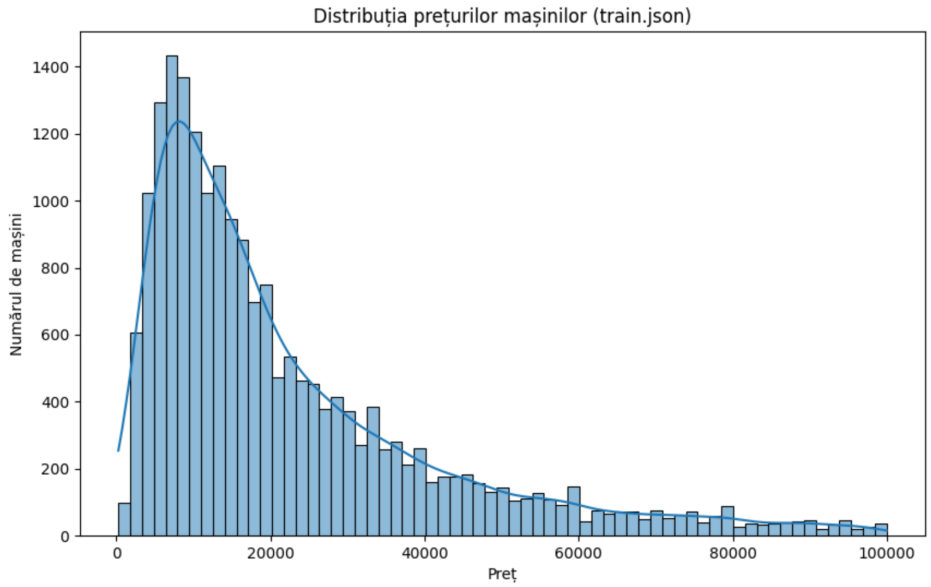
(train.json)



(test.josn)

Putem trage concluzia că seturile de date sunt echilibrate. Diferă doar numărul de mașini in funcție de an.

c) Distribuția prețurilor în raport cu numărul total de mașini – Acest lucru ne indică faptul că unele mărci au prețuri concentrate într-un anumit interval, iar altele au o variație mare. Acest lucru poate ajuta la implementarea modelului prin tratarea valorilor extreme.



3. Implementare

În urma analizei setului de date, am identificat caracteristici diverse ale vehiculelor, cum ar fi marca, modelul, anul de fabricație, kilometrii parcurși, puterea motorului, tipul cutiei de viteze, combustibilul, capacitatea cilindrică, transmisia, caroseria, culoarea și multe altele. Valorile acestor caracteristici se impart în două categorii: numerice și categoriale.

Un prim pas a constat în procesarea caracteristicilor de dotări (addons): Am creat caracteristici binare pentru cele mai frecvente dotări din lista de addons. Acest lucru permite luarea în considerare de către model a celor mai frecvente dotări.

Al doilea pas a constat verificarea valorilor lipsă din setul de date și completatea acestora. Pentru acest lucru am folosit SimpleImputer.

Al treilea pas a constat în identificarea caracteristiclor numerice si categoriale. Pentru caracteristicile categoriale am aplicat tehnica One-Hot-Encoding care are rolul de a le transforma in caracteristici numerice pentru a fi folosite de model.

Următorul pas a constat în alegerea modelului capabil să prezică prețurile folosind datele menționate anterior. Modelul a fost ales în urma unor comparții dintre XGBoost și LiniarRegression, DecisionTreeRegression, unde XGBoost a oferit rezultate mai bune decât celelalte modele. Pentru a optimiza procesul de antrenare și datorită variantei datelor numerice am considerat ca scalarea logaritmică a etichetelor (prețurilor) poate aduce beneficii. Alegerea hiperparametrilor a fost facută aleatoriu, încercând mai multe combinații.

Pentru evaluarea modelului am folosit un subset din setul de antrenare, care a

reprezentat aproximativ 10% din întregul set de date. Am ales să folosesc Kfold pentru a împărți setul de date in subseturi pentru a face o validare încrucișată, iar parametrii folosiți sunt (n_splits=5, shuffle=True, random_state=42), unde n_splits reprezintă numărul de subseturi, shuffle reprezintă posibilitatea de a amesteca datele înainte de a fii împărțite, iar random_state reprezintă seed ul pentru generatorul de numere aleatoare.

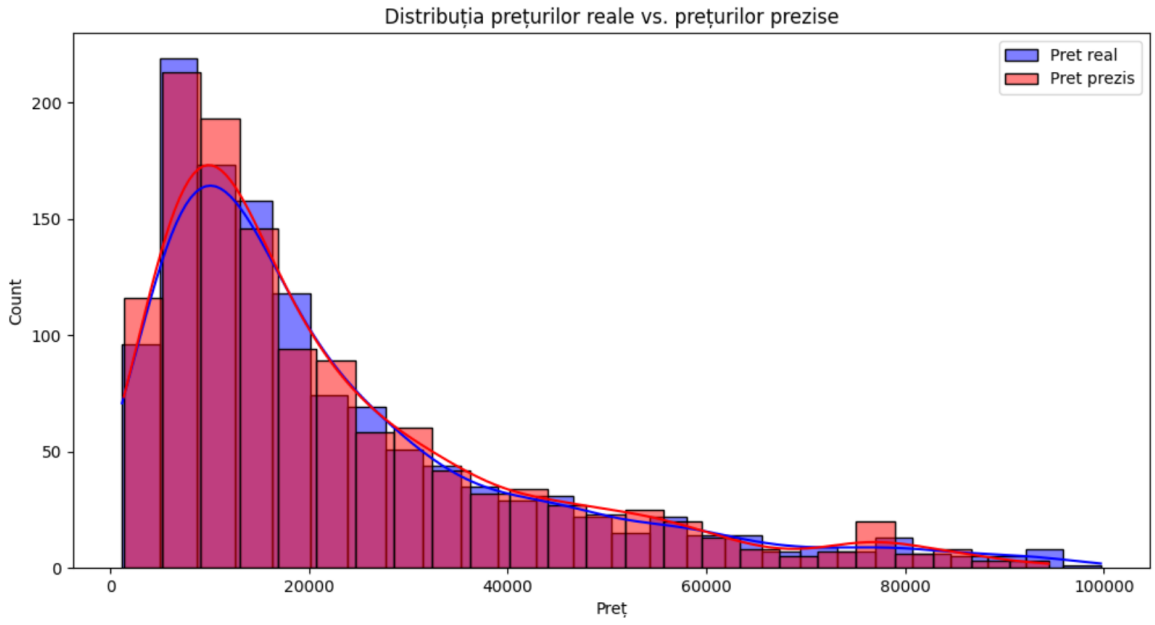
În imaginea de mai jos de găsesc scorurile asociate fiecărui subset:

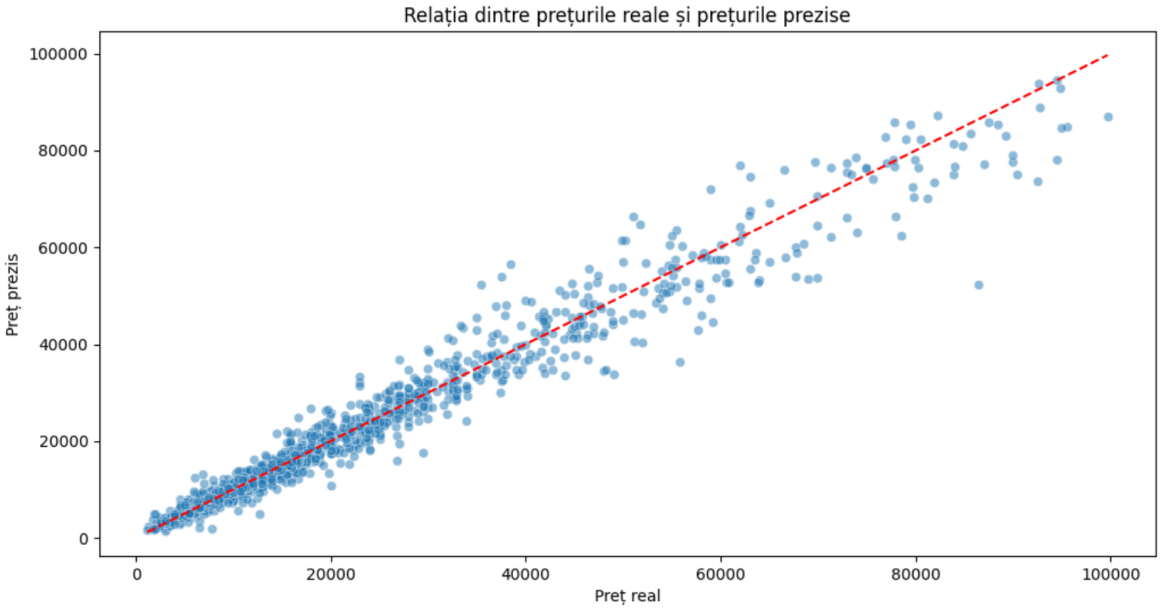
```
CV R2 scores: [0.96031453 0.95479926 0.95536704 0.94997343 0.95552292]
```

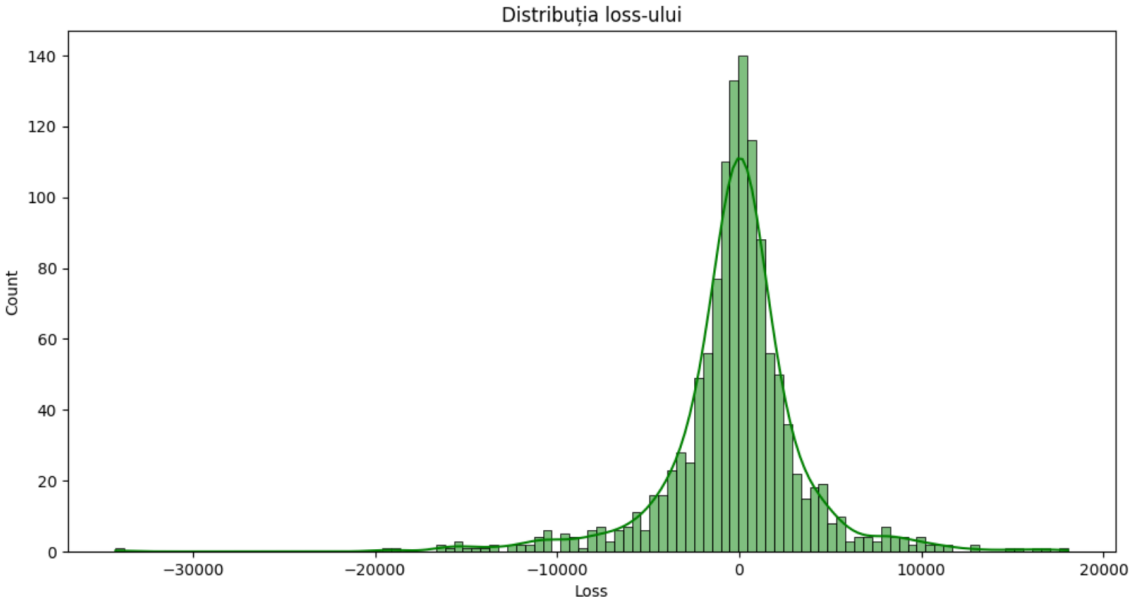
Valoarea scorului R2 pe intreg setul de test se poate observa în imaginea de mai jos:

```
Scorul R2 pe setul de test este: 0.9609163581472802
```

Pentru a evalua performanțele modelului s-au realizat mai multe plot-uri:




Relația dintre prețurile reale și prețurile prezise
Preț prezis
Preț real
0
20000
40000
60000
80000
100000


Distribuția loss-ului
Count
Loss
0
20
40
60
80
100
120
140
−30000
−20000
−10000
0
10000
20000

	ID	Pret prezis	Pret real
19000	19000	12498.0	11490
19001	19001	7583.0	5250
19002	19002	50421.0	44990
19003	19003	18370.0	22200
19004	19004	92926.0	94800
...	...	...	...
20240	20240	15721.0	14756
20241	20241	9579.0	10750
20242	20242	10298.0	9485
20243	20243	66668.0	62951
20244	20244	1862.0	1990

4. Concluzii

Se poate trage concluzia ca modelul a obținut un scor bun, iar valorile prezise sunt destul de apropiate de cele reale, cu mențiunea că pentru unele instanțe pot exista diferențe mai mari legat de valoarea prețului. Consider că extremele din cadrul setului de date (masinile vechi si masinile noi) au avut un impact asupra acelor valori mai puțin apropiate de cele reale.

Link GitHub : https://github.com/alexbursuc/SecondHandCarPricePredictor