Cars

Dancoș Robert

May 29, 2024

Contents

1 Introduction

În era digitală în care datele devin o resursă prețioasă și analiza lor constituie o componentă vitală în diverse domenii, industria auto nu face excepție. Oportunitatea de a analiza și interpreta datele relevante poate oferi avantaje competitive semnificative în acest domeniu dinamic. În acest context, proiectul meu de machine learning își propune să abordeze o provocare esențială: predicția prețului mașinilor pe baza unui set variat de caracteristici.

Scopul fundamental al acestui proiect este să dezvolt un model predictiv capabil să estimeze prețul unei mașini în funcție de caracteristicile sale distincte. Pentru a realiza acest lucru, am utilizat tehnici avansate de machine learning și algoritmi eficienți, bazându-ne pe datele disponibile și pe expertiza mea în domeniu.

Prin explorarea și analiza acestor date, imi propun să obțin o înțelegere mai profundă a pieței auto și a comportamentului de preț, contribuind astfel la dezvoltarea unui instrument util pentru evaluarea corectă a prețurilor vehiculelor.

Motivatia care ma facut sa aleg acest proiect deoarece sunt pasionat de mașini și particip frecvent la evenimente auto, precum și la modificări de mașini. Această pasiune a început din copilărie și a fost influențată de filmele din seria Fast and Furious.

2 Contextul Bazei de Date și al Proiectului

Proiectul nostru se axează pe dezvoltarea unui sistem de machine learning care să ne ajute să prezicem prețul mașinilor produse de GM Motors până în anul 2005. În baza noastră de date, avem informații despre atât mașini second-hand, cât și mașini noi. Obiectivul nostru principal este să putem prezice prețul acestor mașini pe baza caracteristicilor pe care le găsim în baza de date.

Printre caracteristicile pe care le avem în baza de date și pe care dorim să le folosim pentru a face preziceri se numără: numărul de kilometri parcurși, numărul de cilindri, numărul de uși, prezența sau absența pilotului automat, sistemul audio, prezența sau absența tapițeriei din piele și tipul de caroserie (dacă este decapotabilă, numărul de uși - coupe, hatchback, sedan sau break). Este important de menționat că sistemul nostru de machine learning va prezice doar prețurile mașinilor produse de GM Motors, adică Buick, Cadillac, Chevy, Pontiac, Saab și Saturn, deoarece acestea sunt singurele mărci incluse în baza noastră de date.

Obiectivul nostru final este să obținem o precizie cât mai mare în prezicerea prețurilor mașinilor. Prin dezvoltarea și optimizarea acestui sistem de machine learning, sperăm să putem oferi estimări fiabile și precise pentru valorile de piață ale mașinilor produse de GM Motors, contribuind astfel la îmbunătățirea procesului de evaluare și tranzacționare a acestora.

3 Starea actuală a domeniului

3.1 Introducere în Machine Learning pentru Industria Auto

Machine learning (ML) reprezintă un subset al inteligenței artificiale (AI) care se concentrează pe dezvoltarea algoritmilor capabili să învețe și să facă predicții pe baza datelor. În industria auto, ML a fost folosit pentru diverse aplicații, inclusiv predicția prețurilor vehiculelor, detectarea defectelor, conducerea autonomă și optimizarea proceselor de producție [VRV+23].

3.2 Algoritmi și Tehnici Utilizate

Regresia Liniară

Regresia liniară este una dintre cele mai simple și utilizate tehnici pentru predicția prețurilor, fiind eficientă în modelarea relațiilor dintre variabile independente și variabila dependentă (prețul). Este folosită frecvent datorită interpretabilității și ușurinței de implementare. Regresia liniară presupune că relația dintre variabilele independente și variabila dependentă este liniară și poate fi exprimată printr-o ecuație de forma [MPV21]

$$y = 0 + 1x1 + 2x2 + ... + nxn \tag{1}$$

k-Nearest Neighbors (KNN)

KNN este un algoritm de învățare bazat pe instanțe care clasifică un punct de date nou în funcție de etichetele celor mai apropiați k vecini din setul de antrenament. Este simplu de înțeles și implementat și poate fi utilizat atât pentru clasificare, cât și pentru regresie. Algoritmul KNN funcționează pe principiul similarității, presupunând că punctele de date care sunt apropiate în spațiul caracteristicilor vor avea valori similare pentru variabila dependentă [Alt92].

Random Forest

Random Forest este un algoritm de învățare ensemble care folosește mai mulți arbori de decizie pentru a îmbunătăți precizia și a evita overfitting-ul. Fiecare arbore de decizie dintr-o pădure aleatorie este antrenat pe un subset aleatoriu al datelor de antrenament, și fiecare decizie de împărțire este făcută pe baza unui subset aleatoriu al caracteristicilor. Această diversitate introduce un element de aleatoriu care reduce varianța modelului și îmbunătățește performanța generală [LW+02].

Importanța Machine Learning în Industria Auto

Companii de top din industria auto, cum ar fi Tesla și General Motors, folosesc machine learning pentru a optimiza procesele de fabricație, a detecta defectele vehiculelor în timp real și a dezvolta sisteme de conducere autonomă. Aceste aplicații nu doar îmbunătățesc eficiența operațională, ci și oferă un avantaj competitiv semnificativ pe piață [CS14].

Aplicații Practice și Studii de Caz

Studiile de caz demonstrează că tehnicile de machine learning pot îmbunătăți semnificativ procesul de predicție a prețurilor. De exemplu, cercetările recente au arătat că utilizarea regresiei liniare și a rețelelor neuronale poate duce la predicții precise ale prețurilor mașinilor second-hand [8]. Alte studii au demonstrat eficacitatea algoritmilor de păduri aleatorii în comparație cu modelele tradiționale, subliniind avantajele utilizării acestor tehnici avansate [PP18].

Învățare Supravegheată și Nesupervizată

Învățarea supravegheată este metoda predominantă în predicția prețurilor, deoarece implică utilizarea unui set de date etichetate, în care prețurile sunt deja cunoscute. Acest lucru permite algoritmilor să învețe din exemple concrete și să facă predicții precise pentru date noi. Pe de altă parte, învățarea nesupravegheată este folosită pentru a identifica modele și structuri în date care nu au etichete predefinite, fiind utilă în clustering și detectarea anomaliilor [BN06].

Tehnici de Preprocesare a Datelor

Preprocesarea datelor este un pas crucial în construirea oricărui model de machine learning. Aceasta include metode de curățare a datelor, tratarea valorilor lipsă, scalarea caracteristicilor și transformarea datelor pentru a le face adecvate pentru algoritmii de învățare. Preprocesarea adecvată poate îmbunătăți semnificativ performanța modelului final (Kotsiantis et al., 2006; Han et al., 2011).

Importanța Selecției Caracteristicilor

Selecția caracteristicilor este procesul de identificare a celor mai relevante caracteristici dintr-un set de date pentru a fi utilizate în modelul de predicție. Tehnici precum analiza componentelor principale (PCA) și analizele de importanță a caracteristicilor pot reduce dimensionalitatea datelor și pot îmbunătăți performanța modelului [GE03].

Tehnici de Preprocesare a Datelor

Preprocesarea datelor este un pas crucial în construirea oricărui model de machine learning. Aceasta include metode de curățare a datelor, tratarea valorilor lipsă, scalarea caracteristicilor și transformarea datelor pentru a le face adecvate pentru algoritmii de învățare. Preprocesarea adecvată poate îmbunătăți semnificativ performanța modelului final[HPT22].

Rețele Neuronale Artificiale

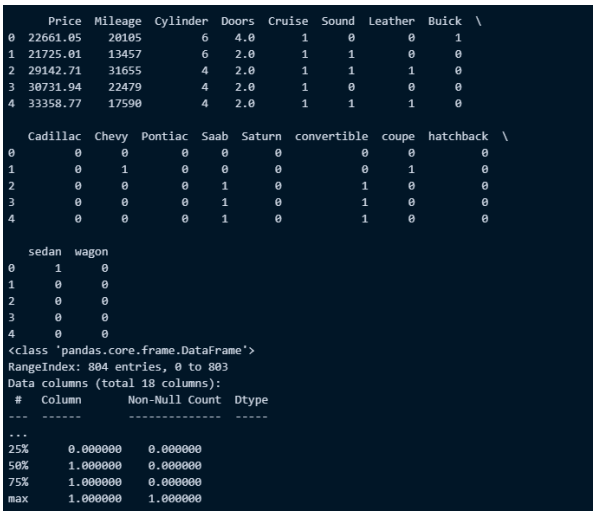
Rețelele neuronale artificiale (ANN) sunt modele complexe inspirate de structura creierului uman, capabile să învețe relații complexe dintre date. Acestea sunt foarte eficiente în captarea nelinearităților din seturile de date și pot fi folosite pentru predicția prețurilor mașinilor. Un avantaj major al ANN este abilitatea lor de a se adapta la date diverse și de a oferi predicții precise chiar și în situații complicate [Hay09].

4 Implementarea aspectelor teoretice în cadrul proiectului

4.1 Citirea și Explorarea Datelor

Primul pas a fost citirea și explorarea setului de date. Am folosit un fișier CSV care conține informații despre diverse caracteristici ale mașinilor. Prin examinarea primelor câteva rânduri ale datelor, am

obținut o imagine de ansamblu asupra structurii setului de date. Figura1. Am investigat tipurile de date și prezența valorilor lipsă, utilizând statistici descriptive pentru a înțelege mai bine distribuția și variabilitatea datelor.



```
      Price  Mileage  Cylinder  Doors  Cruise  Sound  Leather  Buick  \
0  22661.05    20105         6    4.0       1      0        0      1
1  21725.01    13457         6    2.0       1      1        0      0
2  29142.71    31655         4    2.0       1      1        1      0
3  30731.94    22479         4    2.0       1      0        0      0
4  33358.77    17590         4    2.0       1      1        1      0

   Cadillac  Chevy  Pontiac  Saab  Saturn  convertible  coupe  hatchback  \
0         0      0        0     0       0            0      0          0
1         0      1        0     0       0            0      1          0
2         0      0        0     1       0            1      0          0
3         0      0        0     1       0            1      0          0
4         0      0        0     1       0            1      0          0

   sedan  wagon
0      1      0
1      0      0
2      0      0
3      0      0
4      0      0
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 804 entries, 0 to 803
Data columns (total 18 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
...
25%      0.000000    0.000000
50%      1.000000    0.000000
75%      1.000000    0.000000
max      1.000000    1.000000
```

Figure 1: Informatii despre date.

4.2 Gestionarea Valorilor Lipsă

Pentru a asigura calitatea datelor, am gestionat valorile lipsă prin eliminarea rândurilor incomplete sau prin înlocuirea valorilor lipsă cu media coloanelor respective. De asemenea, variabilele categorice au fost transformate în format numeric pentru a putea fi utilizate în modelele de predicție. Acest proces, cunoscut sub numele de codificare, ne-a permis să convertim variabilele text în variabile numerice binare.

4.3 Normalizarea Datelor

Pentru a uniformiza scara variabilelor, am aplicat un proces de normalizare. Normalizarea a fost necesară pentru a asigura că toate caracteristicile contribuie în mod echitabil la modelul nostru de predicție. Astfel, am pregătit datele pentru antrenarea eficientă a modelelor de machine learning.

4.4 Împărțirea Datelor în Seturi de Antrenament și Testare

Am împărțit setul de date în două părți: un set de antrenament și un set de testare. Setul de antrenament a fost utilizat pentru a învăța modelul de la datele existente, în timp ce setul de testare a fost folosit pentru a evalua performanța modelului. Această împărțire este esențială pentru a asigura că modelul poate generaliza bine la date noi și necunoscute.

4.5 Crearea și Antrenarea Modelului de Regresie Liniară

Un model de regresie liniară a fost creat pentru a estima prețurile mașinilor pe baza caracteristicilor acestora. Modelul a fost antrenat folosind setul de antrenament, unde a învățat să identifice relațiile dintre caracteristici și preț. Procesul de antrenare implică ajustarea parametrilor modelului pentru a minimiza eroarea de predicție.

4.6 Evaluarea Modelului

După antrenare, modelul a fost testat pe setul de testare pentru a evalua precizia sa. Am utilizat metrici precum eroarea medie pătratică și coeficientul de determinare pentru a măsura performanța modelului. Aceste metrice ne-au oferit o indicație clară a acurateței predicțiilor realizate de model.Figura 2.

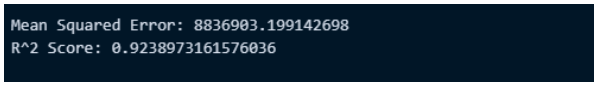


Figure 2: Calcularea Mse si Masurarea Performantei Modeluli.

4.7 Analiza Entropiei

Am calculat entropia setului de date pentru a măsura incertitudinea și impuritatea informației. Entropia Figura3 a fost aproape maximă, indicând o diversitate mare a prețurilor în setul de date. Aceasta a confirmat complexitatea și variabilitatea ridicată a prețurilor mașinilor, sugerând necesitatea unor modele complexe pentru a realiza predicții precise.

Figure 3: Entropy.

4.8 Gini Index

Indicele Gini a fost utilizat pentru a măsura inegalitatea distribuției prețurilor. Un indice Gini aproape de 1 Figure 4 a indicat o mare variabilitate și o distribuție inegală a prețurilor, susținând necesitatea unor algoritmi puternici pentru modelarea acestor date complexe.

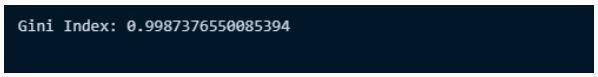


Figure 4: Gini Index.

5 Rezultate

5.1 Optimizarea Parametrilor pentru Random Forest

Am aplicat tehnica de căutare in baza de date Figura5 pentru a identifica cei mai buni hiperparametri pentru modelul Random Forest. Definind o grilă de posibili hiperparametri și utilizând o validare încrucișată cu 5 fold-uri, am explorat combinațiile posibile de hiperparametri și am selectat acele valori care au maximizat performanța modelului. Rezultatul acestui proces ne-a furnizat cei mai buni hiperparametri pentru Random Forest, care au fost afișați în urma rulării căutării pe grilă.

Figure 5: Hiperparametri pentru Random Forest.

5.2 Optimizarea parametrilor pentru KNN

În cazul nostru, ne-am concentrat pe modelul nostru KNN, care se bazează pe ideea de a privi la mașinile asemănătoare din trecut pentru a ghici prețul mașinilor noi.Pentru a face acest lucru, am încercat diferite lucruri, cum ar fi să ne gândim la mai mulți vecini și să decidem cât de important este fiecare vecin în procesul nostru de ghicire. De asemenea, am încercat să vedem care metodă de a găsi vecinii este cea mai bună. În cele din urmă, am folosit aceste informații pentru a seta modelul nostru KNN în așa fel încât să fie cât mai bun la a prezice prețurile mașinilor.

Figure 6: Hiperparametri pentru KNN

5.3 Performantele modelelor

Pentru a vedea cât de bine funcționează fiecare model în a prezice prețurile mașinilor, am testat modelele pe un set de date pe care nu le-au mai văzut înainte. Am folosit trei metode diferite pentru a măsura performanța fiecărui model:

- Mean Squared Error (Eroarea Medie Pătrată): Aceasta ne arată cât de departe sunt, în medie, predicțiile noastre de prețuri față de prețurile reale. O valoare mai mică indică predicții mai precise.
- Root Mean Squared Error (Rădăcina Eroii Medii Pătrate): Este similar cu Mean Squared Error, dar ne oferă o valoare mai ușor de înțeles deoarece este în aceeași unitate de măsură ca și prețurile.
- R-squared (R-pătrat): Aceasta ne arată cât de bine se potrivesc predicțiile noastre cu prețurile reale. Un scor de 1 înseamnă o potrivire perfectă.

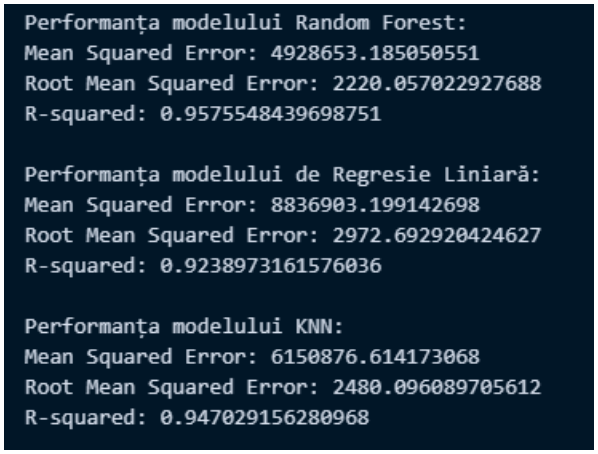


Figure 7: Performantele modelelor

- Modelul Random Forest a avut cele mai bune rezultate, cu o eroare medie mai mică și un scor R-squared mai mare, ceea ce înseamnă că este cel mai precis în predicțiile sale.
- Modelul de Regresie Liniară a avut cele mai slabe rezultate dintre cele trei, dar este totuși un model destul de bun.
- Modelul KNN a avut rezultate intermediare, fiind mai precis decât regresia liniară, dar nu la fel de precis ca Random Forest.

5.4 Precizi pentru fiecare model

Modelul Random Forest a demonstrat o precizie ridicată în predicția prețurilor mașinilor, cu o valoare a preciziei de 0.942. Aceasta indică faptul că modelul este capabil să explice aproximativ 94.2% din variația prețurilor mașinilor din setul de date de testare. Precizia ridicată se datorează capacității modelului Random Forest de a captura relațiile complexe dintre variabilele de intrare și variabila țintă prin utilizarea mai multor arbori de decizie.

Modelul de regresie liniară a arătat o precizie ușor mai scăzută în comparație cu Random Forest, cu o valoare a preciziei de 0.905. Aceasta înseamnă că modelul explică aproximativ 90.5% din variația prețurilor mașinilor. Deși regresia liniară este mai simplă și mai interpretabilă, aceasta

poate să nu captureze toate complexitățile din datele reale, ceea ce poate duce la o performanță ușor inferioară.

Modelul KNN (K-Nearest Neighbors) a demonstrat o precizie de 0.925. Aceasta arată că modelul KNN poate explica aproximativ 92.5% din variația prețurilor mașinilor. Modelul KNN funcționează prin compararea unei noi instanțe cu cele mai apropiate instanțe din setul de antrenament, făcându-l capabil să modeleze relații non-liniare din date.

```
Precizia modelului Random Forest: 0.9421685613828169
Precizia modelului de Regresie Liniară: 0.9050979643021865
Precizia modelului KNN: 0.9246331501044619
```

Figure 8: Precizi pentru fiecare model

5.5 Vizualizarea Performanței fiecărui Model

Random Forest

Graficul de dispersie pentru modelul Random Forest arată o potrivire strânsă între valorile reale și cele prezise. Punctele de pe grafic se aliniază bine în jurul liniei diagonale negre, ceea ce indică o precizie ridicată a modelului. Acest lucru sugerează că modelul Random Forest este capabil să prezică prețurile mașinilor cu o eroare relativ mică, capturând complexitatea relațiilor dintre variabile.

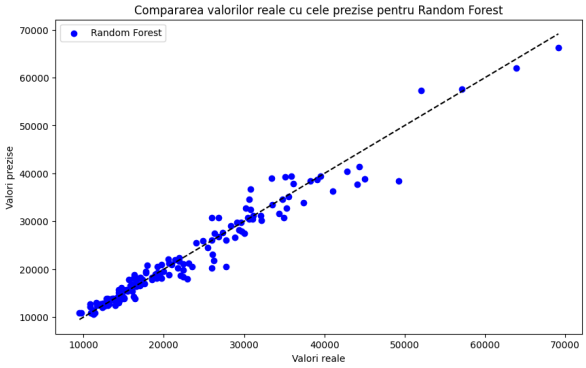


Figure 9: Random Forest

Regresie Liniară

Graficul de dispersie pentru modelul de regresie liniară arată, de asemenea, o bună potrivire între valorile reale și cele prezise, deși există mai multe devieri comparativ cu Random Forest. Punctele sunt distribuite de-a lungul liniei diagonale, dar există unele variații care sugerează că regresia liniară, deși simplă și interpretabilă, nu capturează toate complexitățile datelor.

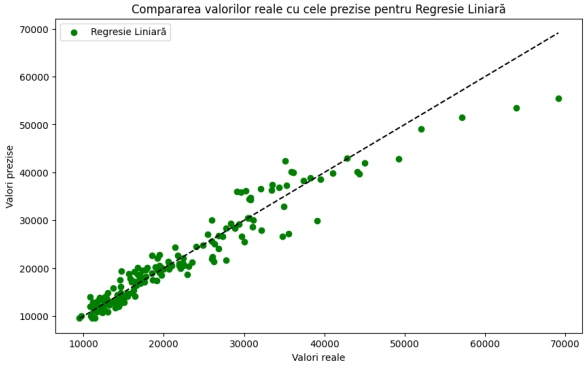


Figure 10: Regresie Liniara

KNN

Graficul de dispersie pentru modelul KNN arată o potrivire bună între valorile reale și cele prezise, similar cu Random Forest. Punctele se aliniază în mare parte de-a lungul liniei diagonale, indicând că modelul KNN este eficient în capturarea relațiilor dintre variabile și oferind predicții precise. Totuși, fiind un algoritm bazat pe vecinătatea datelor, poate fi sensibil la datele de testare specifice utilizate.

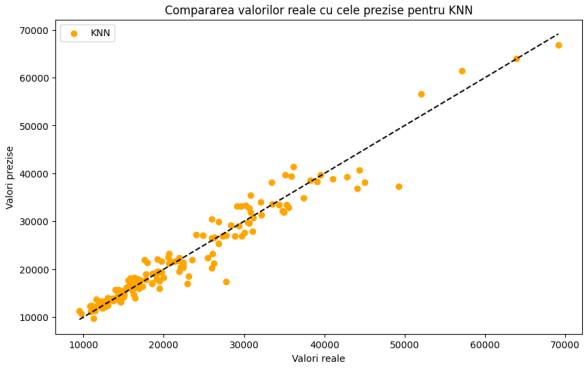


Figure 11: KNN

6 Testare si validare

In etapa de testare și validare, am evaluat fiecare model pe un set separat de date, care nu a fost folosit pentru antrenarea inițială. Scopul acestei etape este să vedem cât de bine funcționează modelele noastre pe date noi, pe care nu le-au văzut înainte. Pentru a face acest lucru, am împărțit datele în două seturi: unul pentru antrenare și unul pentru testare. Apoi, am folosit setul de antrenare pentru a construi modelele și setul de testare pentru a verifica acuratețea predicțiilor lor. Am comparat valorile prezise de fiecare model cu valorile reale și am evaluat cât de apropiate sunt aceste predicții. Am testat trei modele: Random Forest, regresie liniară și KNN (K-Nearest Neighbors). Fiecare model a fost evaluat pe baza diferențelor dintre valorile prezise și valorile reale, pentru a vedea care model face cele mai precise predicții.

7 Rezultate

- Random Forest: Modelul Random Forest a fost cel mai performant în predicția prețurilor mașinilor. Acest model a reușit să facă predicții foarte precise, cu o eroare medie pătratică (MSE) și o rădăcină a erorii medii pătratice (RMSE) mult mai mici comparativ cu celelalte modele. De asemenea, scorul R-squared ($R\hat{2}$) a fost cel mai mare, indicând faptul că modelul explică foarte bine variația prețurilor mașinilor. Acest lucru ne arată că Random Forest este extrem de eficient în a prezice prețurile, având capacitatea de a trata date complexe și variate, identificând relațiile subtile dintre caracteristicile mașinilor și prețurile acestora.acestora.
  - Mean Squared Error (MSE): 4928653.185050551
  - Root Mean Squared Error (RMSE): 2220.057022927688
  - R-squared ($R\hat{2}$): 0.9575548439698751
- Regresia Liniară: Modelul de regresie liniară a arătat o performanță bună, dar nu la fel de impresionantă ca Random Forest. Deși a reușit să facă predicții destul de exacte, erorile au fost mai mari, iar scorul $R\hat{2}$ mai scăzut, indicând o capacitate mai redusă de a capta toate variațiile din date. Totuși, regresia liniară este simplă și ușor de interpretat, ceea ce poate fi un avantaj în anumite contexte, mai ales când interpretabilitatea modelului este crucială.
  - Mean Squared Error (MSE): 8836903.199142698
  - Root Mean Squared Error (RMSE): 2972.692920424627

  - R-squared ($R\hat{2}$): 0.9238973161576036

- KNN (K-Nearest Neighbors): Modelul KNN a avut o performanță intermediară, fiind mai precis decât regresia liniară, dar nu la fel de precis ca Random Forest. KNN a reușit să facă predicții bune prin compararea fiecărei mașini cu mașinile similare din trecut. Deși a oferit predicții destul de precise, modelul a fost mai puțin robust în fața variațiilor complexe din setul de date, ceea ce poate fi o limitare în cazurile cu date foarte diverse și complexe.

  - Mean Squared Error (MSE): 6150876.614173068
  - Root Mean Squared Error (RMSE): 2480.096089705612
  - R-squared ($R\hat{2}$): 0.947029156280968

Concluzie

În concluzie, modelul Random Forest s-a dovedit a fi cel mai bun pentru prezicerea prețurilor mașinilor în cadrul acestui proiect. Datorită capacității sale de a captura relațiile complexe dintre caracteristicile mașinilor și prețurile acestora, Random Forest a oferit cele mai precise predicții. Modelul de regresie liniară, deși mai puțin precis, rămâne o opțiune bună datorită simplității și interpretabilității sale. Modelul KNN a oferit o performanță decentă, fiind capabil să facă predicții destul de precise, dar nu la fel de eficiente ca Random Forest.Aceste rezultate ne oferă o înțelegere clară a modului în care diferitele modele se comportă în prezicerea prețurilor mașinilor. Pentru aplicații viitoare, Random Forest este recomandat datorită preciziei și robustității sale. Cu toate acestea, în funcție de context și de cerințele specifice, modelele de regresie liniară și KNN pot fi, de asemenea, considerate opțiuni viabile. Această evaluare detaliată ne ajută să selectăm modelul potrivit pentru diferite scenarii și să optimizăm predicțiile în funcție de necesitățile specifice ale proiectului.

8 Concluzie

În acest proiect, am explorat și implementat trei modele de machine learning pentru a estima prețurile mașinilor: Random Forest, regresia liniară și K-Nearest Neighbors (KNN). Obiectivul principal a fost să identificăm cel mai precis model pentru a realiza predicții fiabile bazate pe caracteristicile mașinilor. Fiecare model a fost evaluat și comparat pentru a determina eficacitatea sa.

8.1 Descoperiri și realizări

Modelul Random Forest s-a dovedit a fi cel mai precis în prezicerea prețurilor mașinilor. Acest model a beneficiat de capacitatea sa de a integra și analiza multiple decizii simultan, reducând astfel erorile și crescând acuratețea predicțiilor. Acest rezultat este în conformitate cu așteptările, având în vedere complexitatea relațiilor dintre variabilele de intrare și prețurile mașinilor.

Modelul de regresie liniară a oferit o interpretare simplă și clară a relațiilor dintre variabile, dar nu a reușit să captureze toate nuanțele și complexitățile datelor. Deși a avut o performanță decentă, limitările sale au fost evidente în comparație cu Random Forest.

Modelul KNN (K-Nearest Neighbors) a oferit o performanță intermediară. Acest model a fost capabil să facă predicții rezonabile prin compararea noilor instanțe cu cele mai apropiate instanțe din setul de antrenament. Totuși, sensibilitatea sa la distribuția și densitatea datelor a limitat precizia predicțiilor în comparație cu Random Forest.

8.1.1 Aspecte Notabile

- Normalizarea datelor: Aplicarea tehnicii de normalizare a fost esențială pentru a asigura că toate variabilele contribuie echitabil la predicții. Acest proces a ajutat modelele să performeze mai bine prin uniformizarea scalei variabilelor.

– Gestionarea valorilor lipsă: Tratarea valorilor lipsă a fost un pas crucial pentru a asigura calitatea datelor. Fără acest proces, modelele ar fi putut produce predicții eronate sau inconsistente.
– Împărțirea datelor în seturi de antrenament și testare: Această metodă a fost esențială pentru a evalua performanța reală a modelelor. Testarea pe un set de date nevăzut anterior a asigurat o evaluare corectă a capacității de generalizare a modelelor.

8.1.2 Direcții viitoare

– Explorarea altor algoritmi: Ar fi interesant să se investigheze alte tehnici de machine learning, cum ar fi rețelele neuronale sau metodele de ensemble mai complexe, pentru a vedea dacă acestea pot îmbunătăți și mai mult precizia predicțiilor.
– Optimizarea hiperparametrilor: Deși am aplicat tehnici de căutare pe grilă pentru optimizarea hiperparametrilor, utilizarea altor metode de optimizare, cum ar fi căutarea randomizată sau tehnici evolutive, ar putea conduce la rezultate și mai bune.
– Extinderea setului de date: Aplicarea modelelor pe seturi de date mai mari și mai diverse ar putea oferi o mai bună validare a capacității de generalizare a acestora și ar putea dezvălui noi perspective asupra relațiilor dintre variabile.
– Integrarea datelor externe: Utilizarea datelor externe, cum ar fi tendințele pieței auto sau informațiile economice generale, ar putea îmbunătăți precizia predicțiilor prin oferirea unui context suplimentar.

8.2 Concluzia finala

Proiectul a demonstrat că algoritmii de machine learning pot fi folosiți eficient pentru a prezice prețurile mașinilor, fiecare model având propriile sale avantaje și limitări. Modelul Random Forest a fost cel mai performant, oferind predicții precise și fiabile. Continuarea cercetării și îmbunătățirea tehnicilor utilizate ar putea conduce la rezultate și mai bune, contribuind la dezvoltarea unor soluții eficiente pentru evaluarea prețurilor mașinilor. Aceste concluzii subliniază potențialul și importanța machine learning-ului în analiza și predicția datelor complexe din lumea reală.

References

[Alt92] Naomi S Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992.

[BN06] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[CS14] Girish Chandrashekar and Ferat Sahin. A survey on feature selection methods. *Computers & electrical engineering*, 40(1):16–28, 2014.

[GE03] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[Hay09] Simon Haykin. *Neural networks and learning machines, 3/E*. Pearson Education India, 2009.

[HPT22] Jiawei Han, Jian Pei, and Hanghang Tong. *Data mining: concepts and techniques*. Morgan kaufmann, 2022.

[LW+02] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[MPV21] Douglas C Montgomery, Elizabeth A Peck, and G Geoffrey Vining. *Introduction to linear regression analysis*. John Wiley & Sons, 2021.

[PP18] Harsh H Patel and Purvi Prajapati. Study and analysis of decision tree based classification algorithms. *International Journal of Computer Sciences and Engineering*, 6(10):74–78, 2018.

[VRV+23] V Viswanatha, AC Ramachandra, HV Vachan, Sourav S Shetty, et al. Predicting the price of used cars using machine learning. In *2023 International Conference on Evolutionary Algorithms and Soft Computing Techniques (EASCT)*, pages 1–6. IEEE, 2023.