Universitatea Babeş-Bolyai

Facultatea de Ştiinţe Economice şi Gestiunea Afacaerilor

PROIECT

PREZICEREA PREŢURILOR ÎN FUNCȚIE DE CARACTERISTICILE TEHNICE ALE MAȘINILOR

Student: Meseșan Daria

Introducere

În contextul unei industrii auto în continuă evoluție, analiza datelor privind mașinile second-hand din perioada 1970-2024 oferă o perspectivă valoroasă asupra tendințelor de piață și a comportamentului consumatorilor. Prin analiza unui set de date complex ce cuprinde peste 90.000 de mașini, cercetarea actuală își propune să dezvolte un model predictiv capabil să estimeze prețurile vehiculelor second-hand bazându-se pe variabile cheie precum: transmisia, producătorul, tipul de combustibil, modelul sau anul de fabricație. Aceste caracteristici au fost alese deoarece prezintă o influență majoră asupra prețului prin valorile rezultate în urma calculelor realizate în acest studiu.

Obiectivul principal al acestui studiu este de a construi un model de regresie robust care să permită prezicerea prețurilor autovehiculelor luând îm considerare diferite caracteristici tehnice prezente în baza de date. Întrebările de cercetare vizează determinarea influenței specifice a fiecărei variabile asupra prețului de vânzare și evaluarea preciziei și eficacității diferitelor modele statistice în predicția prețului.

În trecut, diverse studii au abordat estimarea prețurilor auto, dar adesea s-au limitat la date mai restrânse sau la perioade mai scurte. Prin urmare, aplicarea unui model predictiv pe un set de date extins și diversificat, cum este cel disponibil pentru perioada 1970-2024 va aduce o contribuție importantă în înțelegerea dinamicii pieței auto second-hand.

Studiul nostru se adresează atât mediului academic de cercetare a piețelor economice și a comportamentului consumatorului, cât și profesioniștilor din domeniul auto care caută să optimizeze strategiile de prețuri și să înțeleagă mai bine factorii care influențează valorile de piață ale vehiculelor.

Întrebările de cercetare la care ne propunem să răspundem sunt următoarele:

1. În ce măsură influențează tipul de transmisie, mărimea motorului, anul de fabricație și modelul prețul final al unei mașini second-hand, în contextul unui set de date care acoperă peste 50 ani în dezvoltarea industriei auto?
2. Cum variază capacitatea de prezicere a prețurilor în funcție de producător și modelul specific al mașinii?
3. Care sunt caracteristicile mașinilor care sunt relevante pentru prețul autovehiculelor?

Analiza detaliată și modelul de prezicere dezvoltat în cadrul acestei lucrări urmăresc să ofere o bază solidă pentru decizii informate și eficiente în industria auto. Aceasta introducere setează un cadru inițial pentru explorarea modelelor de regresie și a impactului lor asupra predicției prețurilor auto, subliniind relevanța și importanța întrebărilor de cercetare alese.

Setul de date

Sursa setului de date utilizat în cadrul acestui proiect este disponibilă pe platforma online Kaggle. https://www.kaggle.com/datasets/meruvulikith/90000-cars-data-from-1970-to-2024/data

Detalii specifice despre originea datelor din industria auto nu au fost furnizate de autor. Datele sunt structurate într-un tabel ce include următoarele coloane:

- **Model**: Modelul mașinii, indicând specificația și versiunea.
- **Anul**: Anul de fabricație al mașinii, care este un indicator al vârstei vehiculului.
- **Preț**: Prețul actual al mașinii pe piața de vehicule second-hand.
- **Transmisie**: Tipul de transmisie al mașinii, cum ar fi manuală sau automată.
- **Kilometraj**: Numărul de kilometri parcurși de mașină, reflectând uzura generală.
- **FuelType:** Tipul de combustibil utilizat de mașină, de exemplu benzină, diesel, electric sau hibrid.
- **Taxă:** Cota de impozitare aplicabilă mașinii, care poate varia în funcție de diverse criterii, inclusiv emisiile de CO2.
- **MPG:** Eficiența consumului de combustibil a mașinii, măsurată în mile pe galon, indicând economia de combustibil.
- **EngineSize:** Dimensiunea motorului mașinii, exprimată în litri, care este un factor determinant al performanței și al consumului de combustibil.
- **Producător:** Compania care a fabricat vehiculul, indicând brandul și, implicit, segmentul de piață al mașinii.

Variabilele din punctul de vedere al tipului de date:

- **Integer (int):** year, price, mileage, tax
- **String (chr):** model, transmission, fuelType, Manufacturer
- **Numerical (num)**: mpg, engineSize

Analizând aceste date și realizând grafice pentru fiecare variabilă în mod individual, putem face observații preliminare care vor informa analiza ulterioară și modelarea predictivă pentru prețul mașinilor bazată pe caracteristicile enumerate.

Analiza preliminară a datelor

O primă observație realizată asupra setului de date este faptul că unele variabile se comportă ca niște clase, chiar dacă ele au o structură numerică. Un exemplu ar fi variabila **year** sau variabila **engineSize** care, deși au formă continuă, ele reprezintă niște categorii în care se încadrează prețurile autovehiculelor. ***Figura 1*** prezintă o descriere vizuală a acestui fenomen, în care prețurile sunt aranjate pe coloane verticale în funcție de anul producției și nu au o structură liniară continuă. Din acest motiv, variabilele din baza date sunt factorizate pentru a putea îmbunătăți analiza prețurilor. ***Figura 2*** este o vizualizare a prețurilor în funcție de anul producției sub formă de candele, în urma factorizării variabilei **year.**

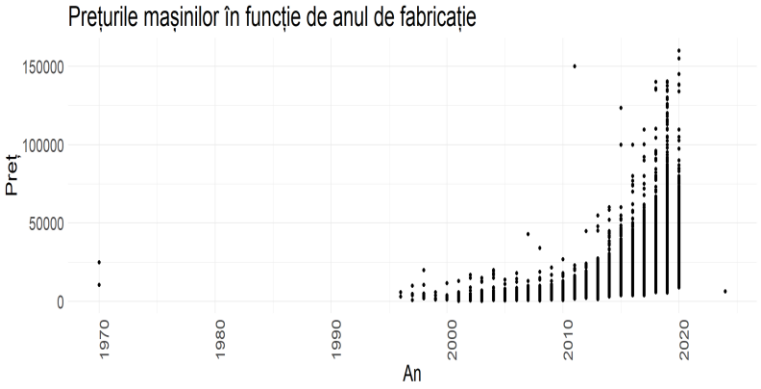

Fig1. Reprezentarea pe puncte a prețurilor în funcție de anul de fabricație

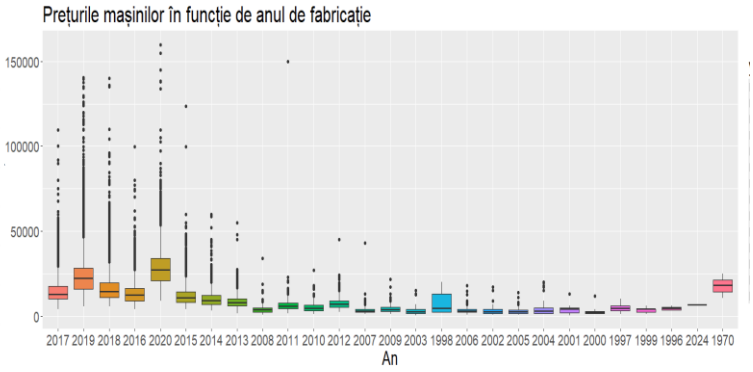

Fig2. Reprezentarea folosind candele

Astfel, înainte de a începe procesul de regresie liniară, variabilele engineSize, year, transmission, Manufacturer, model, fuelType și tax au fost factorizate. Următorul capitol va aprofunda influența caracteristicilor mașinii asupra prețului lor folosind regresia liniară.

Regresia liniară – rezultate și discuții

Studiul curent propune realizarea unei predicții numerice, întrucât variabila dependentă, prețul, este de tip numeric. Se va analiza detaliat influența variabilelor existente în baza de date în vederea alegerii unor caracteristici relevante.

Un prim pas este crearea modelelor liniare folosind funcția **lm** pentru a ajusta diferite variabile în funcție de preț. Se va exprima variabila dependentă luând combinații de variabile independente de la general la specific. Tabelul de mai jos cuprinde rezultatele aceste modelări, împreună cu descrierea detaliată a valorilor extrase.

Analiza comparativă a rezultatelor modelării pentru fiecare variabilă din baza de date

Tabelul 1. Rezultatele obținute în urma apelării metodei summary pentru regresiile liniare ale fiecărei variabile și evidențierea primelor 5 cele mai semnificative variabile independente

Model	R-squared (%)	Adjusted R-squared	F-statistic	p-value
mod_price_all	90.24	0.9021	2868	< 2e-16
mod_price_model	61.79	0.6172	812.9	< 2e-16
mod_price_year	31.52	0.315	1729	< 2e-16
mod_price_transmission	30.61	0.3061	14370	< 2e-16
mod_price_mileage	17.47	0.1747	20690	< 2e-16
mod_price_tax	17.42	0.1738	438.4	< 2e-16
mod_price_fuelType	5.29	0.05283	1363	< 2e-16
mod_price_mpg	8.73	0.08731	9349	< 2e-16
mod_price_engineSize	49.1	0.4908	2416	< 2e-16
mod_price_Manufacturer	28.49	0.2848	4865	< 2e-16

Prin analizarea rezultatelor fiecărui model de regresie, observăm că modelul complet (mod_price_all), care include toate variabilele disponibile, oferă cea mai înaltă valoare a coeficientului R-squared, indicând că poate explica aproximativ 90.24% din varianța prețurilor auto. Acest lucru sugerează o potrivire excelentă a modelului la date și relevanța tuturor variabilelor selectate pentru predicția prețului. F-statistica este mare pentru toate variabilele cu un p-value < 2.2e-16, ceea ce indică că toate modelele sunt, în ansamblu, semnificative. Cu toate acestea, modelele care utilizează o singură caracteristică, cum ar fi tipul de combustibil (mod_price_fuelType) sau eficiența consumului de combustibil (mod_price_mpg), au R-squared mult mai scăzut, sugerând că aceste caracteristici izolate sunt predictori slabi pentru prețul auto.

Când comparam eficiența diferitelor modele, observăm că mod_price_engineSize, cu un R-squared de aproape 49.1%, este mai eficient comparativ cu alte modele, indicând o corelație semnificativă între dimensiunea motorului și prețul auto. Acesta este un indiciu că specificațiile tehnice ale unei mașini, cum ar fi capacitatea motorului, sunt foarte valoroase pentru determinarea prețului.

Pe de altă parte, mod_price_Manufacturer și mod_price_model demonstrează că marca și modelul unei mașini sunt, de asemenea, indicatori importanți ai prețului, cu R-squared-uri de 28.49% și 61.79%, respectiv. Aceasta subliniază influența brandului și a modelului specific în percepția valorii de piață. În același mod se poate observa influența anului de fabricație, cu un R-squared de 31.52%.

Analiza comparativă a valorilor rezultate ne indică importanța crescută a modelului, a anului de fabricație, a tipului de transmisie, a dimensiunii motorului și a producătorului. Studiul va continua prin aprofundarea analizei pentru trei dintre cele mai semnificative variabile și se va explica fiecare rezultat în detaliu. Mai apoi, variabilele vor fi combinate pentru a genera modele mai complexe.

1. Regresia simplă **Preț (price)** în raport cu **Modelul (model)**

Tabelul 2. Rezultatele pentru Residuals si Coefficients (în cazul a 5 modele selectate manual) obținute în urma sumarizării modelului mod_price_model

Residuals				
Min	1Q	Median	3Q	Max
-63702	-2967	-324	2508	120641
Coefficients				
Variabilă	Coeficient Estimat	Eroare Standard	Valoare t	p-valoare
(Intercept)	15810.911	137.852	114.695	< 2e-16 ***
model 2 Series	3634.434	224.471	16.191	< 2e-16 ***
model S3	4568.534	1445.804	3.160	0.001579 **
model A2	-13320.911	6107.636	-2.181	0.029184 *
model Tourneo Connect	-1948.505	1088.180	-1.791	0.073359 .
model Kadjar	-1379.244	3528.041	-0.391	0.695844

Explicarea rezultatelor la nivelul Coefficients:

Exemplificarea s-a realizat doar asupra unor modele selectate pentru a captura o diversitate mai mare a coeficienților, fără a aglomera spațiul de studiu. Modelul de regresie arată că prețul de bază mediu este de aproximativ 15810.911 unități monetare. Comparativ cu acest preț de bază, modelul "2 Series" are un preț mediu cu 3634.434 unități mai mare, fiind extrem de semnificativ din punct de vedere statistic. Modelul "S3" are un preț mediu cu 4568.534 unități mai mare decât interceptul, cu

o semnificație statistică foarte mare. În schimb, modelul "A2" scade prețul mediu cu 13320.911 unități, ceea ce îl face mai ieftin decât prețul de bază și cu o semnificație statistică mai scăzută. Modelul "Tourneo Connect" scade prețul mediu cu 1948.505 unități și este doar marginal semnificativ statistic. Modelul "Kadjar" scade prețul mediu cu 1379.244 unități și nu este semnificativ statistic.

Explicarea rezultatelor pentru celelalte caracteristici ale modelului

- **Multiple R-squared:** 0.6179 indică faptul că modelul explică aproximativ 61.79% din variația totală a prețurilor
- **Adjusted R-squared:** 0.6172, ajustat după numărul de predictori în model
- **F-statistic:** Valoarea F-statisticii este 812.9, arătând că modelul, în ansamblu, este statistic semnificativ la nivelul datelor colectate.
- **p-value (F-statistic):** p-value < 2.2e-16 confirmă semnificația generală a modelului la nivelul datelor existente.

Semnificația coeficienților (simbolizată prin steluțe):

- ***: $p < 0.001$, extrem de semnificativ statistic.
- **: $p < 0.01$, foarte semnificativ statistic.
- *: $p < 0.05$, semnificativ statistic.
- .: $p < 0.1$, marginal semnificativ statistic.
- Fără steluțe: $p \geq 0.1$, nu este semnificativ statistic.

2. Regresia simplă **Preț (price)** în raport cu **dimensiunea motorului (engineSize)**

Tabelul 3. Rezultatele pentru Residuals si Coefficients (în cazul a 5 motoare selectate manual) obținute în urma sumarizării modelului mod_price_engineSize

Residuals				
Min	1Q	Median	3Q	Max
-98045	-4094	-789	3644	116932

Coefficients				
Variabilă	Coeficient Estimat	Eroare Standard	Valoare t	p-valoare
(Intercept)	11111.78	53.88	206.240	< 2e-16 ***
engineSize2	9863.32	68.90	143.154	< 2e-16 ***
engineSize1.2	-1323.09	101.41	-13.047	< 2e-16 ***
engineSize1.8	4596.87	178.38	25.771	< 2e-16 ***
engineSize1.5	5406.24	85.51	63.224	< 2e-16 ***
engineSize5.2	95683.43	1469.33	65.120	< 2e-16 ***

Explicarea rezultatelor la nivelul Coefficients:

Această analiză detaliată a impactului dimensiunii motorului asupra prețului arată că modelele cu motoare mai mari tind să fie semnificativ mai scumpe, reflectând costurile mai mari de producție și percepția de performanță și lux asociate cu acestea. Coeficienții pozitivi mari pentru dimensiunile mari ale motorului și coeficienții negativi pentru cele mai mici subliniază importanța acestei caracteristici în formarea prețurilor vehiculelor.

Explicarea rezultatelor pentru celelalte caracteristici ale modelului

- **Residual standard error:** 7042, arată variația medie a erorilor modelului de la valorile reale.
- **Degrees of Freedom:** 97672, indicând numărul mare de observații ajustate pentru numărul de predicții.
- **Multiple R-squared:** 0.491, sugerează că aproximativ 49.1% din varianța prețului este explicată de varianța dimensiunii motorului.
- **Adjusted R-squared:** 0.4908, ajustarea R-squared pentru numărul de variabile predictor.
- **F-statistic:** 2416, indică semnificația modelului în ansamblu.
- **p-value**: < 2.2e-16, confirmă semnificația statistică a modelului.

3. Regresia simplă **Preț (price)** în raport cu **tipul de transmisie (transmission)**

Tabelul 4. Rezultatele pentru Residuals si Coefficients obținute în urma sumarizării modelului mod_price_transmission

Residuals				
Min	1Q	Median	3Q	Max
-22735	-4577	-1255	3413	135764
Coefficients				
Variabilă	Coeficient Estimat	Eroare Standard	Valoare t	p-valoare
(Intercept)	12076.62	34.89	346.090	< 2e-16 ***
transmissionSemi-Auto	12158.60	65.18	186.533	< 2e-16 ***
transmissionAutomatic	9435.59	67.92	138.928	< 2e-16 ***
transmissionOther	4142.50	2740.46	1.512	0.131

Explicarea rezultatelor la nivelul Coefficients:

Această analiză arată cum diferitele tipuri de transmisie afectează prețul vehiculelor, cu transmisiile semi-automate și automate aducând un surplus de valoare comparativ cu cele manuale sau alte tipuri de transmisie.

Explicarea rezultatelor pentru celelalte caracteristici ale modelului

- **Residual standard error**: 8221 pe 97708 grade de libertate indică o dispersie considerabilă a erorilor față de valorile prezise de model.
- **Multiple R-squared**: 0.3061, arată că aproximativ 30.61% din varianța prețurilor este explicată de tipul de transmisie.
- **Adjusted R-squared**: 0.3061, este ajustat pentru numărul de predictori în model.
- **F-statistic**: 14370 pe 3 și 97708 grade de libertate, indică o semnificație statistică a modelului.
- **p-value**: < 2.2e-16, confirmă că modelul este statistic semnificativ la nivelul datelor.
-

Combinarea variabilelor pentru crearea unui model complex – **engineSize+year+transmission+model**

Modelul complex de regresie generat combină patru variabile importante: anul fabricației (**year**), tipul de transmisie (**transmission**), dimensiunea motorului (**engineSize**), și modelul mașinii (**model**). Alegerea acestor variabile este fundamentată pe premisa că acestea sunt printre principalii factori care influențează prețul unui vehicul, idee demonstrată în studiul anterior al variabilelor.

Modelul complex își propune să capteze în mod eficient variațiile de preț din piața auto prin evaluarea impactului combinat al acestor caracteristici, oferind o înțelegere mai profundă a dinamicii prețurilor. În cod, acest model poate fi indentificat prin numele **” mod_price_complex”.**

Analiza rezultatelor modelului în urma apelării metodei summary indică o corelație puternică între cele mai semnificative patru variabile dependente față de prețul autovehiculelor. **„mod_price_complex”** are un R-squared de 0.8779, indicând că aproximativ 87.79% din variația prețurilor este explicată de variabilele incluse în model. Acest lucru demonstrează o ajustare foarte bună a modelului la date. De asemenea, significanța statistică este confirmată de p-value-uri extrem de mici pentru majoritatea variabilelor, indicând că rezultatele sunt robuste și puțin probabil să fie datorate șansei.

Combinarea variabilelor pentru evaluarea legăturii **producător-model asupra prețului**

Pentru a înțelege mai bine impactul combinat al producătorului **(Manufacturer)** și modelului (**model**) asupra prețurilor mașinilor, am construit un model de regresie care include aceste două variabile. Combinația acestora vizează să capteze cum percepțiile de brand și caracteristicile specifice ale fiecărui model influențează valoarea de piață a vehiculelor. Prin integrarea acestor două variabile într-un singur model, putem identifica nu doar valoarea adăugată de brand-ul auto în sine, ci și cum diferitele modele ale aceluiași producător se compară între ele în termeni de preț.

”**mod_price_model_Manufacturer”** a înregistrat un **R-squared** de 0.6179, ceea ce indică faptul că aproximativ 61.79% din variația prețurilor este explicată de variabilele incluse în model, și demonstrează o ajustare bună a modelului la date. Este un indicator solid că variabilele alese au o influență semnificativă asupra prețurilor vehiculelor. P-value-urile extrem de mici pentru majoritatea coeficienților sugerează că relațiile detectate între preț și variabilele incluse sunt statistic semnificative, cu o probabilitate foarte scăzută ca aceste efecte să fie rezultatul fluctuațiilor aleatorii. Coeficienții modelului reflectă contribuția specifică a fiecărui producător și model la prețul vehiculelor. De exemplu, modelul 8 Series de la BMW are un coeficient semnificativ mare, ceea ce indică o asociere cu prețuri substanțial mai ridicate comparativ cu prețul de bază, reflectând performanțele superioare ale modelului.

Prin combinarea producătorului și modelului în analiza regresiei, am putut obține o imagine complexă și nuanțată a modului în care identitatea de brand și particularitățile modelului individual influențează prețurile vehiculelor. Astfel, se poate observa că prețul mașinilor sunt influențate într-o mare măsura de atât de producător, cât și de modelul caracteristic fiecărei mărci.

Antrenarea modelelor – analiză și rezultate

Împărțirea setului de date în date de antrenament și date de test

În studiul prezent, am antrenat mai multe modele de regresie liniară pentru a prezice prețul autovehiculelor. Inițial, setul de date complet Cars este împărțit în două subseturi: unul pentru antrenament (70%) și altul pentru test (30%). Utilizarea unei seed aleator (set.seed(123)) asigură reproductibilitatea împărțirii, o componentă cheie în cercetarea aplicată.

Rezolvarea erorilor generate de absența unor nivele din setul de test

Așa cum am menționat anterior, deși variabilele independente sunt numerice, ele se pot comporta ca niște clase. Acesta este cazul și pentru variabila engineSize, motiv pentru care am factorizat-o. Cu toate acestea, în timpul predicției s-a generat o eroare datorată absenței unor nivele în subsetul de testare. Pentru a remedia eroarea, se crează rânduri dummy în setul de antrenament pentru fiecare nivel lipsă identificat în setul de testare. Acest procedeu previne apariția de erori când modelul se confruntă cu categorii neîntâlnite anterior.

Antrenarea modelelor

Se creează modele de regresie liniară pe cele mai importante variabile independente asupra setului de antrenament. Modelele generate și antrenate în cadrul acestui studiu sunt:

- **Modelul Complex**: Se antrenează un model de regresie liniară care include ca predictori **anul fabricației, tipul de transmisie, dimensiunea motorului și producătorul**. Acest model este cel mai complet și încearcă să capteze interacțiunile dintre toți predictorii importanți în estimarea prețului.
- **Modelul pe Dimensiunea Motorului**: Se evaluează influența dimensiunii motorului asupra prețului, prin antrenarea unui model ce utilizează doar această variabilă.
- **Modelul pe Tipul de Transmisie**: Similar, se analizează efectul tipului de transmisie asupra prețului.
- **Modelul pe Anul de Fabricație**: Se investighează cum variază prețurile autovehiculelor în funcție de anul în care au fost fabricate.
- **Modelul pe Producător**: Acest model explorează diferențele de preț între diferiți producători de autovehicule.

Aceste modele sunt salvate în variabile și sunt folosite mai departe pentru a face predicții asupra setului de test. Predicțiile sunt din nou salvate în variabile și sunt incluse în setul de test pentru a se putea face plotarea. Pentru a evalua performanța modelelor, fiecare predicție impelemntată va avea o reprezentare vizuală explicată detaliat în secțiunea următoare.

Reprezentarea vizuală a predicțiilor generate

1. Modelul Complex – year + engineSize + transmission + Manufacturer

Figura 3 reprezintă proiecția prețurilor reale în comparație cu cele prezise de modelul complex. Punctele reprezintă vehiculele individuale din setul de test, iar axa x arată prețul real al fiecărui vehicul, în timp ce axa y arată prețul prezis de model. O linie de regresie este trasată pentru a indica direcția generală a predicțiilor - dacă punctele sunt aproape de această linie, acest lucru sugerează că modelul oferă predicții precise. Putem observa o acuratețe mai înaltă a predicției în cazul prețurilor mai joase și o eroare mai mare pentru prețurile mai ridicate.

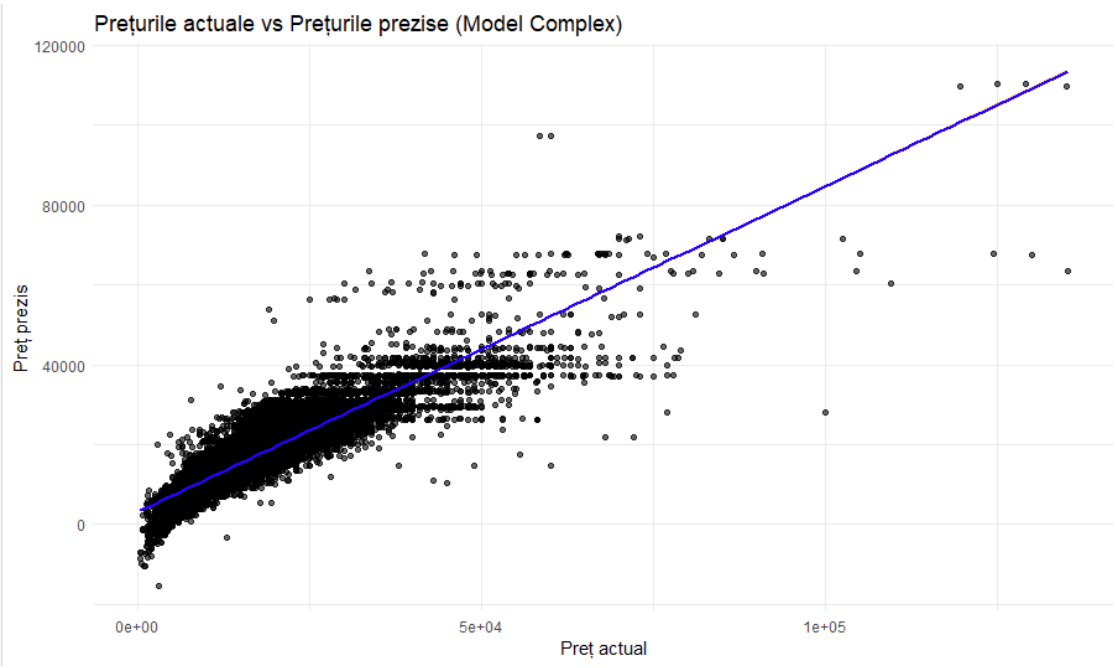

Fig. 3: Prețurile reale în comparație cu prețurile prezise de modelul complex în setul de test

2. predicted_price_engineSize – prețul prezis în funcție de dimensiunea motorului

Figura 4 evidențiază relația între dimensiunea motorului (engineSize) și prețul vehiculelor (price). În general, se poate observa că mașinile cu motoare mai mari (de exemplu, cele marcate cu 3, 4, 5.5) par să aibă prețuri mai mari. Pentru dimensiunile extreme ale motorului (cele foarte mici sau foarte mari), valorile anticipate au variații mai mari față de medianele reale, sugerând că modelul poate avea dificultăți în a prezice cu exactitate impactul dimensiunilor extreme asupra prețului.

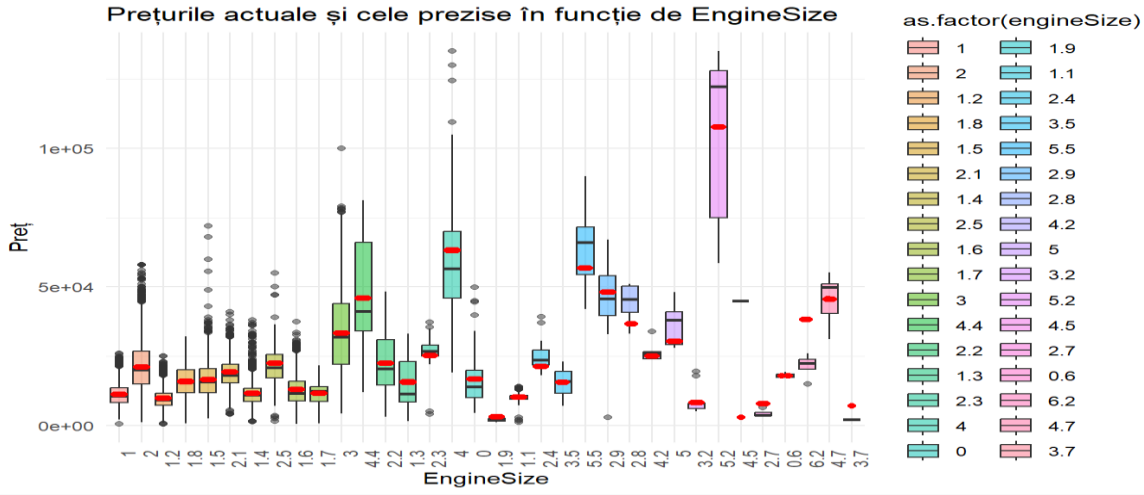

Fig. 4: Prețurile reale în comparație cu prețurile prezise în setul de test în funcție de dimensiunea motorului (engineSize)

3. predicted_price_Manufacturer – prețul prezis în funcție de producător

Figura 5 prezintă comparația între prețurile reale și cele prezise pe baza producătorului vehiculului. Se poate observa că există o variație considerabilă a prețurilor între diferiți producători. De exemplu, Audi și BMW par să aibă prețuri mai ridicate comparativ cu Hyundai și Vauxhall. Predicțiile modelului pentru Ford, Toyota sau Vauxhall sunt aproape de mediana boxplot-urilor, sugerând că modelul prezice bine prețurile pentru aceste mărci. O concluzie ar fi că modelul prezice mai bine prețul autovehiculelor pentru brandurile care au modele cu prețuri mai mici și care nu se îndepărtează prea mult de primele trei quartile. În cazul producătorilor cu o varietate mai mare a prețurilor, care nu se încadrează în 75% din dimensiunea datelor, putem osbserva o supraestimare a prețurilor.

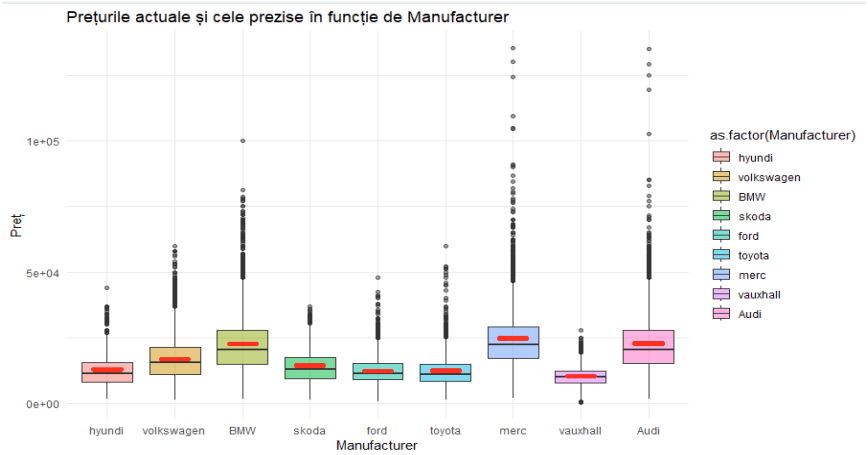

Fig. 5: Prețurile reale în comparație cu prețurile prezise în setul de test în funcție de producători (Manufacturer)

4. predicted_price_transmission – prețul prezis în funcție de tipul transmisiei

În **Figura 6** prețurile sunt comparate în funcție de tipul de transmisie. La prima vedere, tipurile de transmisie Manual și Automatic par să aibă o distribuție mai largă a prețurilor reale comparativ cu Semi-Auto și Other, care sunt mai concentrate. Rezultatele sugereaza ca modelul este mai exact în predicția prețurilor pentru transmisii Manual și Automatic și mai puțin precis pentru celelalte tipuri. De asemenea, prezicerea foarte slabă pentru transmisia Other poate fi datorită lipsei de semnificație statistică a acestei clase, demonstrată în secțiunile anterioare.

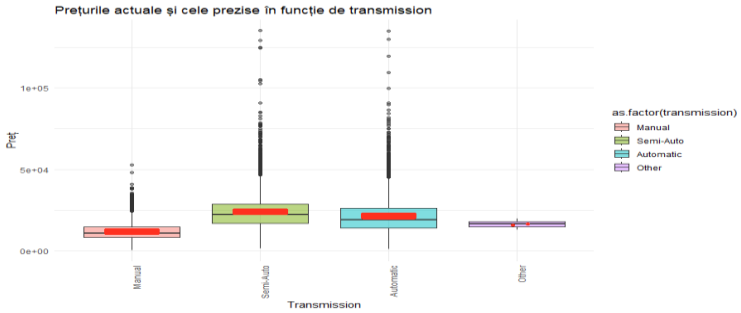

Fig. 6: Prețurile reale în comparație cu prețurile prezise în setul de test în funcție de tipul transmisiei (transmission)

5. predicted_price_year – prețul prezis în funcție de anul producției

Figura 7 reprezintă prețurile reale ale mașinilor sortate după anul de fabricație. Se poate observa că prețurile reale ale mașinilor (boxplot-urile) tind să fie mai mari pentru mașinile mai noi (spre partea stângă a graficului) și mai mici pentru mașinile mai vechi (dreapta). Predicțiile modelului (punctele roșii) urmează această tendință generală, cu unele predicții aliniate aproape de medianele boxplot-urilor, ceea ce sugerează că modelul este mai precis pentru anii mai recenți. Pentru mașinile foarte vechi (în special cele de dinainte de anul 2000), predicțiile sunt sub mediana prețurilor reale, ceea ce ar putea indica o exagerare a devalorizării atovehiculului de către model.

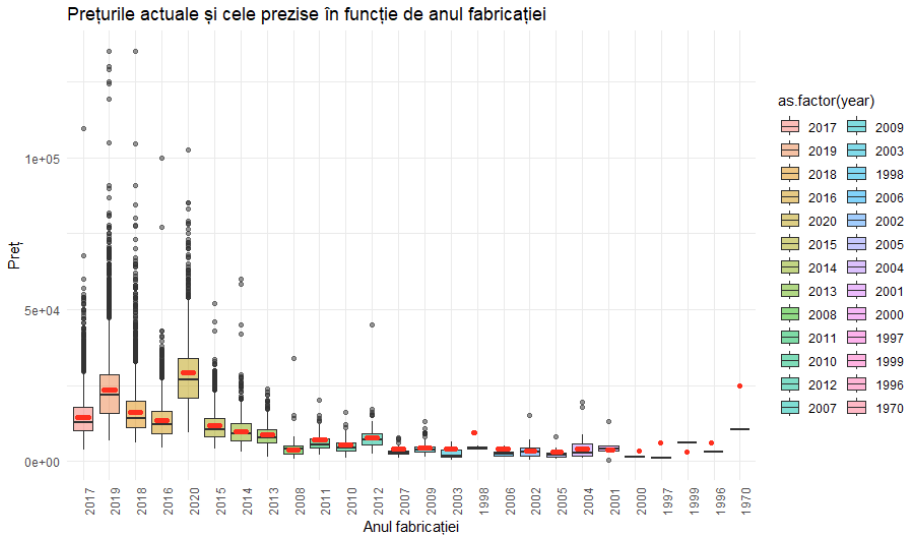

Fig. 7: Prețurile reale în comparație cu prețurile prezise în setul de test în funcție de anul producției (year)

Arborele de decizie

În abordarea noastră folosind R-Studio, am demarat procesul prin divizarea setului de date în două porțiuni distincte: una pentru antrenare, reprezentând 70% din total, și cealalta destinată testării, care a inclus restul de 30%. După această împărțire inițială, am procedat la construirea efectivă a modelului de arbore de decizie, aplicând algoritmul corespunzător pentru această sarcină.

Acest arbore de decizie modelează **prețul** mașinilor bazat pe variabilele independente. Arborele de decizie este un instrument de modelare predictivă care se divide în noduri, ramuri și frunze, reprezentând decizii și rezultatele lor.

Fiecare nivel este influentat de diferinte valori ale variabilelor independente, ceea ce sugereaza faptul ca prețul autovehiculelor este determinat de o multitudine de factori puternic legați între ei. **Fig 10** reprezintă arborele de decizie rezultat în funcție de toate variabilele independente relevante.

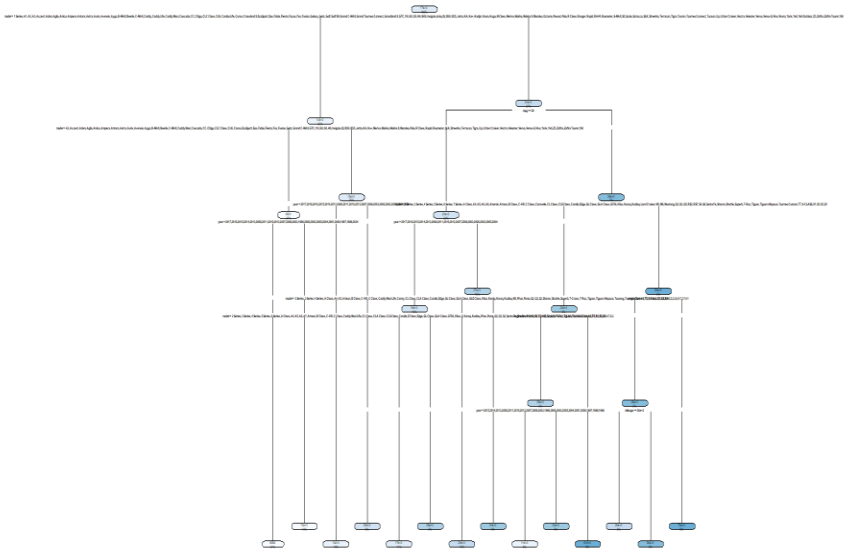

Fig10. Arborele de decizie

1. Nodul Rădăcină: Toate mașinile sunt inițial divizate în două grupuri majore pe baza modelului mașinii.

2. Noduri Intermediare:

Nod 2: Include modele mai accesibile de mașini, cum ar fi 1 Series, A1, A2, etc. Acest grup este ulterior subdivizat pe baza anului de fabricație:

- **Nod 4:** Acest nod conține mașini fabricate în anii mai vechi, precum 2017, 2016, 2015, și alții specificați.
- **Nod 5**: Acest nod include mașini mai noi, fabricate în anii 2019 și 2020.

Nod 3: Include modele mai scumpe de mașini, cum ar fi 2 Series, 3 Series, 4 Series, etc. Diviziunea ulterioară este, de asemenea, bazată pe anul de fabricație:

- Nod 6: Subdivizare pentru modele fabricate în anii specificați anterior.
- Nod 7: Subdivizare pentru modele fabricate în anii mai recenți, 2019 și 2020.

3. Mai multe Divizări: După an, subdiviziunile continuă pe baza mărimii motorului (engineSize), diferite grupuri indicând importanța acestei variabile în predicția prețului.

4. Frunze (Noduri Terminale): Fiecare traseu în arbore se încheie cu o frunză care reprezintă predicția medie a prețului pentru mașinile din grupa respectivă.

5. Procente și Valori din Frunze: Numerele din nodurile terminale (de exemplu, 8686.249, 12198.510) reprezintă predicția medie a prețului mașinilor. Acestea reflectă probabilitatea relativă a traseului în comparație cu întregul set de date.

6. Importanța Variabilelor: Modelul mașinii este variabila utilizată la nodul rădăcină, sugerând o influență majoră asupra prețului. Anul de fabricație și mărimea motorului urmează în importanță, evidențiindu-se și în subdiviziuni.

7. Predicții și Setul de Date: Numărul de observații în fiecare frunză, precum și devianța, indică cât de multe mașini din setul de date se încadrează în fiecare predicție finală și cât de variabil este prețul în cadrul fiecărui grup.

Concluzii Ajustate la Arborele de Decizie pentru Predicția Prețurilor Mașinilor

Această structură a arborelui de decizie subliniază că prețul unei mașini este influențat puternic de modelul acesteia, anul de fabricație și dimensiunea motorului, cu mașinile mai noi și cele cu motoare mai mari având tendința de a fi mai scumpe. Predicțiile finale pentru prețuri sunt rezultatul analizei complexe a acestor variabile, și fiecare nod terminal oferă o estimare a prețului mediu pentru grupul respectiv de mașini.

Modelul Mașinii: Este primul factor folosit pentru divizarea setului de date, ceea ce subliniază influența semnificativă a modelului asupra prețului. Modelele sunt clasificate în categorii care reflectă diferite segmente de preț, de la cele mai accesibile la cele premium.

Anul de Fabricație: Anul fabricației este un predictor important al prețului, cu mașinile mai noi, fabricate în 2019 sau 2020, având prețuri mai ridicate comparativ cu mașinile mai vechi.

Dimensiunea Motorului: Dimensiunea motorului este analizată în nodurile mai profunde ale arborelui, indicând o relație complexă cu prețul. Motoarele de diferite dimensiuni conduc la prețuri variate, posibil datorită performanțelor diferite sau eficienței combustibilului.

Performanța și Optimizarea Modelului

Se constată că algoritmul aplicat creează diverși arbori folosind cross-validation, selecționând un arbore și un parametru de complexitate al costurilor, notat cu α, cu scopul de a optimiza suma dintre eroarea pătratică medie (SSE) și produsul dintre α și numărul de noduri terminale ale arborelui, simbolizat prin |T|.

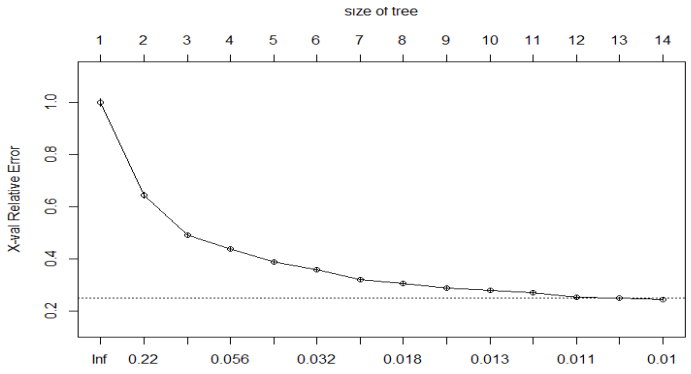

Fig11. Graficul pentru optimizarea costului - complexitate

Conform graficului analizat, cel mai eficient nivel al SSE se obține pentru o valoare de α stabilă la 0.01, ceea ce corespunde cu un arbore de dimensiune 14.

Analizând diferite modele pentru a identifica arborele cel mai eficient, am ajustat **minsplit** (numărul minim de observații necesare pentru divizarea ulterioară a unui nod) și **maxdepth** (adâncimea maximă permisă a arborelui, măsurată prin numărul de noduri interne de la rădăcină până la frunze).

	minsplit	maxdepth	cp	error
1	7	9	0.01	0.2406270
2	15	12	0.01	0.2413602
3	12	14	0.01	0.2417812
4	8	11	0.01	0.2421207
5	9	13	0.01	0.2422453

Tabelul 5. Rezultatele pentru minsplit si maxdepth

În urma acestei analize, am extras primele cinci variante superioare, rezultând că versiunea ideală a arborelui necesită cel puțin 7 observații într-un grup pentru a împărți mai departe și nu mai mult de 9 noduri interne de la rădăcină la frunză. Acest model echilibrează complexitatea arborelui cu capacitatea de a capta esențialul din date fără a suprancărca modelul (overfitting), având cel mai mic scor de eroare (0.2406270).

RMSE (Root Mean Square Error) este o măsură a erorilor între valorile prezise de un model și valorile observate. Cu cât RMSE este mai mic, cu atât modelul este considerat mai precis și mai eficient în predicții. În tabelul de ma jos este prezentată comparația performanței între modelele de regresie liniară și modelele de arbore de decizie, folosind metrica RMSE pentru a evalua erorile dintre valorile prezise și cele observate.

Predicții	**RMSE**
Modelului complex(engineSize+year+transmission+model)	4316.27880200683
Modelului pe EngineSize	7001.21502327627
Modelului pe Transmission	8147.50754439958
Modelului pe Year	8067.99740161827
Modelului pe Manufacturer	8271.82372619649
Arbore de decizie	4751.96

Tabelul 6. Rezultatele pentru RMSE in funcție de modelul folosit

Modelul complex, care include combinarea mai multor variabile (EngineSize, Year, Transmission, și Model), a prezentat cel mai mic RMSE (4316.28), sugerând o capacitate superioară de a prezice relații subtile și interacțiuni între variabilele multiple. Aceast rezultat indică eficiența utilizării unui model complex în determinarea prețurilor autovehiculelor. Pe de altă parte, modelele care au utilizat

o singură caracteristică, cum ar fi EngineSize, Transmission, Year, sau Manufacturer, au arătat RMSE-uri mult mai mari, reflectând limitările lor în capturarea variației prețurilor. Rezultatele subliniază importanța combinării variabilelor pentru a îmbunătăți acuratețea în modelarea predictivă.

Concluzii

Studiul curent analizează în profunzime variabilitatea prețurilor mașinilor second-hand, folosind o bază de date extensivă ce acoperă perioada 1970-2024. Scopul principal este de a construi și evalua modele predictive care să estimeze prețurile vehiculelor pe baza unor caracteristici tehnice esențiale.

Prin utilizarea regresiei liniare și a arborilor de decizie, s-a demonstrat că prețurile sunt influențate semnificativ de anul de fabricație, mărimea motorului, tipul de transmisie și modelul mașinii. De asemenea, a fost evidentă superioritatea modelului complex de regresie liniară în predicția prețurilor, oferind cea mai mică eroare de predicție (RMSE) comparativ cu modelele ce au folosit variabile izolate. Acest lucru indică faptul că un model care combină mai multe caracteristici tehnice oferă o perspectivă mai precisă și mai profundă asupra factorilor care influențează prețurile.

Pe de altă parte, arborii de decizie au facilitat vizualizarea și înțelegerea modului în care diferitele segmente de caracteristici tehnice conduc la formarea prețurilor. Deși nu au fost la fel de precișși ca regresia liniară în predicția numerică exactă a prețurilor, arborii de decizie au oferit o perspectivă clară asupra ierarhiei și importanței caracteristicilor. De exemplu, divizarea inițială a datelor pe baza modelului mașinii a indicat o influență predominantă a acestui factor asupra prețului, urmată apoi de anul de fabricație și mărimea motorului, reflectând diferitele niveluri de variații în calitatea mașinilor.

În concluzie, acest studiu contribuie la literatura existentă prin extinderea înțelegerii factorilor care influențează prețurile mașinilor second-hand, demonstrând valoarea combinării diferitelor metode statistice pentru a îmbunătăți precizia analizelor predictive în domeniul auto. Pentru date cu relații liniare clare și directe, regresia liniară poate fi suficientă și mai ușor de interpretat, în timp ce pentru seturi de date cu mai multe variabile interconectate și relații complexe, arborii de decizie pot oferi informații analitice suplimentare și o înțelegere mai nuanțată. Astfel, cercetarea prezentă include aplicarea unor metode mixte și analiza detaliată a diferitelor tehnici statistice pentru a perfecționa acuratețea și eficacitatea predicțiilor în sectorul auto