Proiecții Predictive pentru Prețurile Mașinilor Mercedes Uzate

Andranovici Dumitru  
December 2023

**ABSTRACT**

Această analiză se concentrează pe explorarea și predicția prețurilor vehiculelor Mercedes dintr-un set de date privind mașinile uzate. Datele, colectate și puse la dispoziție pe platforma Kaggle, furnizează informații valoroase despre caracteristicile automobilelor, cum ar fi transmisia, kilometrajul și tipul de combustibil. Scopul principal al studiului este să dezvolte modele predictibile pentru prețurile vehiculelor pe baza acestor caracteristici.

Analiza începe prin explorarea setului de date, includând factorizarea variabilelor categorice și o analiză descriptivă detaliată. Relațiile dintre variabilele cheie, precum transmisia, kilometrajul și prețurile, sunt examinate în profunzime. De asemenea, se utilizează tehnici de vizualizare grafică pentru a evidenția tendințele și modelele.

Modelarea lineară reprezintă un element central al acestui studiu, implicând dezvoltarea și evaluarea mai multor modele. Variabilele precum transmisia, kilometrajul și tipul de combustibil sunt luate în considerare în diverse combinații pentru a identifica modele semnificative care influențează variația prețurilor.

Rezultatele obținute demonstrează că anumite caracteristici, cum ar fi transmisia și kilometrajul, au o asociere semnificativă cu prețurile vehiculelor Mercedes. Modelele lineare dezvoltate oferă o înțelegere mai profundă a interacțiunilor dintre aceste caracteristici și susțin procesul de predicție.

În concluzie, această analiză detaliată a setului de date privind mașinile uzate Mercedes aduce contribuții valoroase în înțelegerea factorilor cheie care influențează prețurile vehiculelor. Rezultatele obținute pot oferi perspective importante pentru strategiile de marketing și analiza pieței în domeniul vehiculelor de lux.

**INTRODUCERE**

Industria auto a parcurs o evoluție continuă în ultimele decenii, adaptându-se la schimbările tehnologice, cerințele de mediu și preferințele consumatorilor. În acest peisaj dinamic, piața mașinilor uzate reprezintă un segment esențial și în continuă transformare. Acest context complex generează un interes crescut din partea proprietarilor și a potențialilor cumpărători de mașini Mercedes uzate, care doresc să înțeleagă direcția în care se îndreaptă prețurile acestor vehicule pe piața second-hand. Analiza proiecțiilor predictive devine astfel esențială pentru a anticipa schimbările viitoare și pentru a lua decizii informate în privința vânzării sau achiziției unei mașini Mercedes uzate. Acest articol propune o investigație detaliată a acestor proiecții, luând în considerare factori variabili și dinamici care influențează piața automobilelor second-hand. O atenție deosebită va fi acordată tendințelor generale ale pieței auto, având în vedere inovațiile tehnologice care au impact asupra valorii mașinilor uzate. De asemenea, vom explora evoluția specifică a mărcii Mercedes în industrie și modul în care aceasta influențează proiecțiile prețurilor. În cadrul analizei noastre, vom examina implicațiile aduse de schimbările economice și sociale asupra pieței mașinilor uzate, evidențiind cum aceste influențe se reflectă în particularitățile vehiculelor Mercedes. Factori precum cererea în creștere pentru mașini ecologice, tendințele de mobilitate urbană și schimbările în preferințele consumatorilor vor fi analizați în contextul pieței mașinilor Mercedes uzate. Astfel, acest articol își propune să ofere o perspectivă comprehensivă asupra proiecțiilor predictive ale prețurilor mașinilor Mercedes uzate, într-un context mai amplu al industriei auto și al schimbărilor socioeconomice. Această abordare cuprinzătoare va oferi cititorilor o bază solidă pentru înțelegerea evoluțiilor viitoare și pentru luarea deciziilor informate într-un mediu de piață în continuă schimbare [1].

# MATERIALE ȘI METODE

Metodologia acestui studiu a implicat mai multe etape, începând cu o prelucrare detaliată a setului de date privind autovehiculele Mercedes uzate. Prima fază a constat în gestionarea valorilor lipsă, eliminarea duplicatelor și transformarea datelor pentru a le aduce într-un format potrivit pentru analiză.

În ceea ce privește modificările aduse setului de date, acestea au inclus adăugarea de coloane semnificative și transformări care au extins și adâncit analiza. S-au introdus categorii pentru modele și perioade de producție, iar variabilele au fost transformate pentru a facilita analizele ulterioare. Principalele schimbări au fost descrise, precum și motivul și impactul lor asupra studiului.

Analiza exploratorie a datelor a constat într-o examinare atentă a diferitelor caracteristici, precum modelul, anul fabricației, prețul, tipul de transmisie, kilometrajul, tipul de combustibil, taxa, eficiența de combustibil și cilindreea motorului. S-au efectuat vizualizări grafice pentru a evidenția relațiile dintre aceste caracteristici și prețurile mașinilor Mercedes uzate.

Ulterior, s-a trecut la dezvoltarea modelelor predictive folosind tehnici de învățare automată, cum ar fi regresia liniară, random forest și gradient boosting. S-au ajustat parametrii modelelor pentru a obține performanțe optime. Rezultatele modelelor, reprezentate prin diagrame și coeficienți, au fost prezentate, evidențiind varietatea de abordări aplicate.

Validarea și evaluarea modelelor s-au realizat prin împărțirea setului de date într-un set de antrenare și un set de testare. S-au utilizat metrici precum eroarea medie pătratică și coeficientul de determinare pentru a evalua performanța modelelor. Interpretarea rezultatelor a inclus analiza influențelor semnificative ale diferitelor caracteristici asupra prețurilor mașinilor Mercedes uzate.

Prin această metodologie complexă, studiul a adus o contribuție valoroasă la înțelegerea relațiilor complexe dintre caracteristicile autovehiculelor Mercedes uzate și prețurile acestora, oferind o bază solidă pentru analize viitoare și dezvoltarea de modele predictive în domeniul automobilelor second-hand.

Prelucrarea Datelor:

În prima etapă, am efectuat o prelucrare detaliată a setului de date. Acest proces a inclus gestionarea valorilor lipsă, eliminarea duplicatelor și transformarea datelor în formatul potrivit pentru analiză.

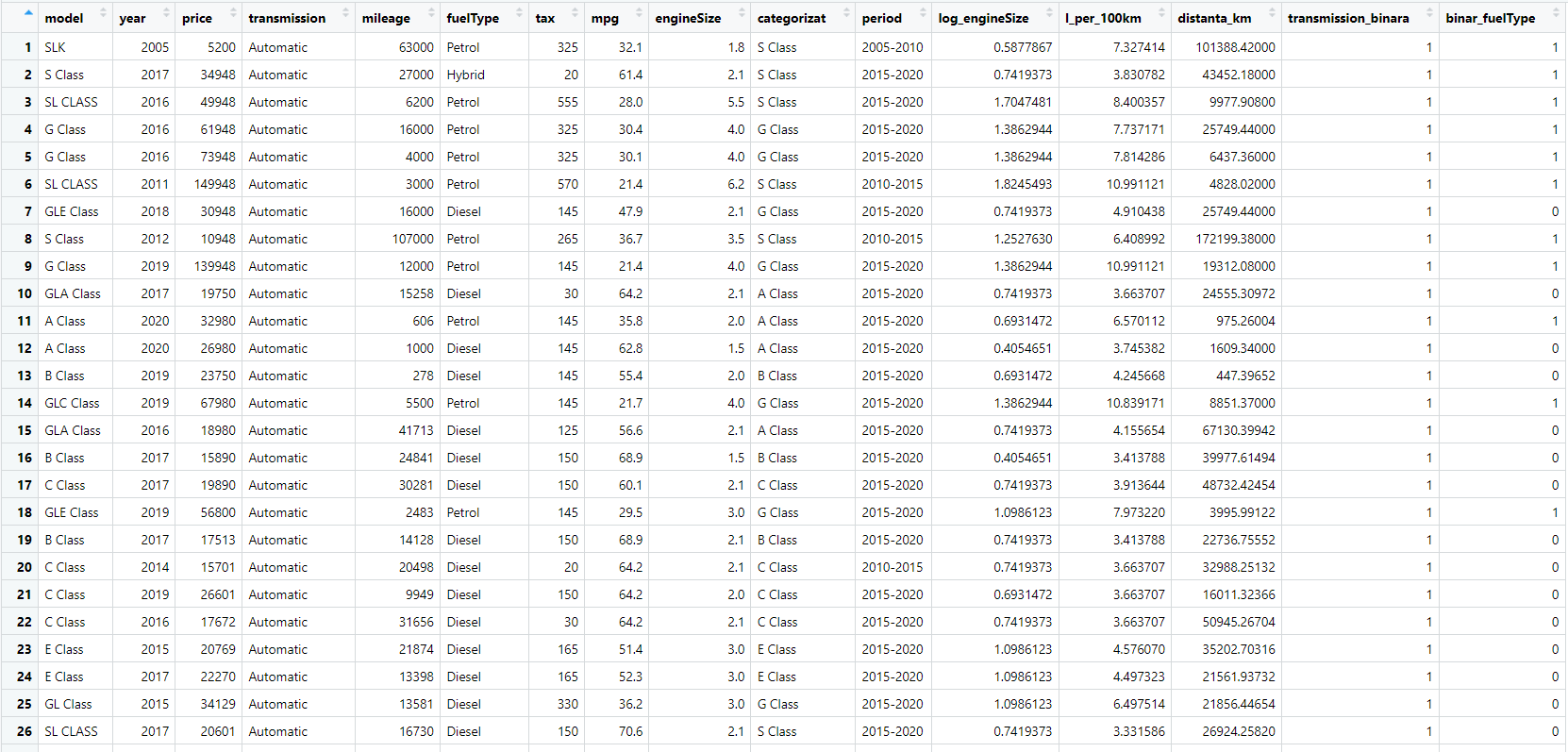


Figura 1. Setul de date a autovehiculelor Mercedes uzate

În urma schimbărilor aduse setului de date, s-au efectuat câteva modificări semnificative. Iată o comparație între setul de date înainte și după implementarea acestor schimbări:

Setul de date înainte de schimbări:

* Număr de rânduri: 13,119
* Număr de coloane: 9
* Coloane principale: model, year, price, transmission, mileage, fuelType, tax, mpg, engineSize

Setul de date după schimbări:

* Număr de rânduri: 13,105
* Număr de coloane: 16
* Coloane adiționale: categorizat, period, log\_engineSize, l\_per\_100km, distanta\_km, transmission\_binara, binar\_fuelType, categorizat\_numeric

Principalele schimbări aduse:

A fost adaugata categorizarea dupa modelul masinii, o nouă coloană "categorizat" a fost introdusă pentru a grupa modelele în categorii specifice, cum ar fi "S Class" sau "G Class", simplificând astfel analiza datelor în funcție de tipul de model.

A fost introdusa categorizarea dupa periada de productie fiecare interval are cite 5 ani, a fost adăugată coloana "period", indicând perioada de fabricație a mașinii pentru a evidenția anii relevanți în procesul de analiză.

A fost applicate transformări logaritmice, au fost adăugata o coloana "log\_engineSize", reflectând logaritmul natural al cilindreei motorului pentru a facilita interpretarea datelor.

Au fost transformate datele din tip char in variabile binare. Două coloane binare, "transmission\_binara" și "binar\_fuelType", au fost create pentru a transforma variabilele categorice precum tipul de transmisie și tipul de combustibil în format binar, facilitând astfel utilizarea acestora în modelele predictive. Au mai fost creat categorizarea dupa timpul modelului de tip numeric.

Aceste schimbări aduc o mai mare complexitate și adâncime analizei noastre, permițând identificarea unor pattern-uri și corelații mai precise între caracteristicile mașinilor Mercedes uzate și prețurile acestora.

Analiza Exploratorie a Datelor:

Setul de date furnizează informații valoroase despre diverse modele de mașini Mercedes uzate. Datele sunt:

**model (modelul mașinii)**: Coloana de tip text (chr) indică modelul specific al mașinii Mercedes, cum ar fi "SLK", "S Class", "SL CLASS", "G Class", etc.

**year (anul fabricației)**: Coloana numerică (dbl) indică anul fabricației fiecărei mașini din set. Datele acoperă o gamă largă de ani, de la modele mai vechi precum 2005 la modele recente din 2020 și 2021.

**price (prețul)**: Coloana numerică (dbl) furnizează informații despre prețul fiecărui vehicul în set. Prețurile sunt exprimate în valori numerice, reflectând variația lor.

**transmission (tipul de transmisie)**: Coloana de tip text (chr) indică tipul de transmisie al fiecărui vehicul, cum ar fi "Automatic" sau "Manual".

**mileage (kilometraj)**: Coloana numerică (dbl) oferă date despre kilometrajul fiecărui vehicul, furnizând astfel informații despre cât de mult a fost utilizat.

**fuelType (tipul de combustibil)**: Coloana de tip text (chr) indică tipul de combustibil utilizat de mașină, precum "Petrol", "Diesel" sau "Hybrid".

**tax (taxa de înmatriculare)**: Coloana numerică (dbl) oferă informații despre taxa de înmatriculare asociată fiecărui vehicul.

**mpg (miles per gallon)**: Coloana numerică (dbl) reflectă eficiența de combustibil a mașinii, măsurată în mile per galon.

**engineSize (cilindreea motorului)**: Coloana numerică (dbl) furnizează informații despre cilindreea motorului al fiecărui vehicul, un indicator important al performanței acestuia.

Aceste detalii variate din setul de date oferă o imagine comprehensivă asupra mașinilor Mercedes uzate și vor servi drept bază solidă pentru analiza și dezvoltarea de modele predictive pentru prețurile acestora, având în vedere caracteristicile precum transmisia, kilometrajul, tipul de combustibil și altele.

# REZULTATE

Am efectuat o analiză exploratorie a datelor pentru a înțelege distribuția variabilelor cheie, relațiile între caracteristici și posibile tendințe. Această etapă ne-a ajutat să identificăm corelații semnificative și să selectăm caracteristicile relevante pentru modelele predictive.

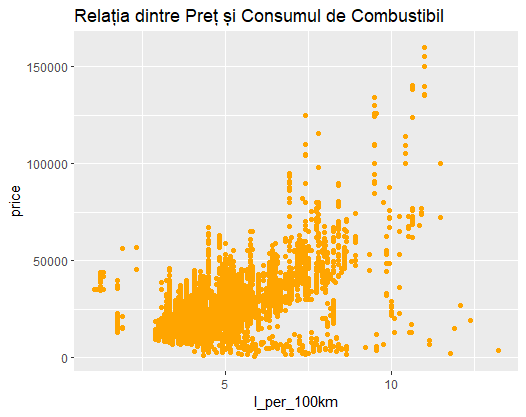


Figura 2. Relația dintre Preț și Consumul de Combustibil

Graficul ilustrează relația între prețul mașinilor Mercedes și consumul de combustibil pe distanța de 100 de kilometri. Se observă o corelație, sugerând că mașinile cu consum mai mic au adesea prețuri mai ridicate și viceversa.

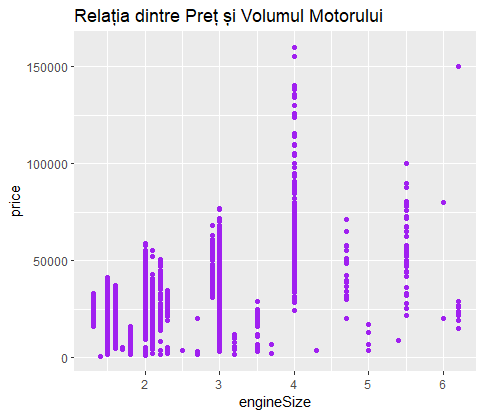


Figura 3. Relația dintre Preț și Capacitatea Motorului

Acest grafic evidențiază o relație strânsă între capacitatea motorului și prețurile mașinilor Mercedes. Punctele indică o corelație evidentă, sugerând că mașinile cu motoare de capacitate mai mare au tendința de a avea prețuri mai ridicate, subliniind impactul capacității motorului asupra valorii acestora.

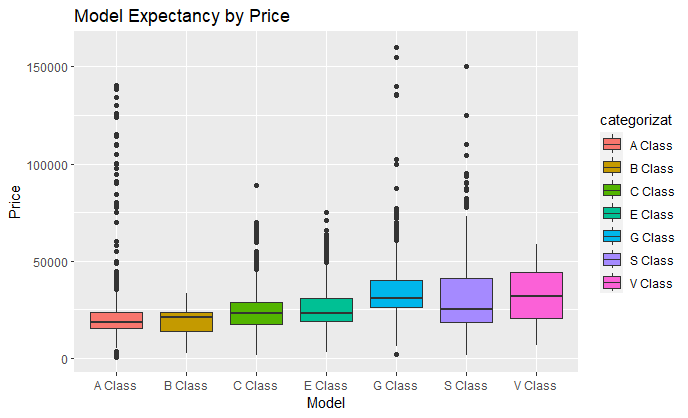


Figura 4. Relația dintre Preț și Model

Acest grafic arată că cele mai mari prețuri în gama de mașini Mercedes sunt asociate modelelor "A Class", "G Class" și "S Class". Aceasta este evidențiată de poziția superioară a cutiilor (boxplot) corespunzătoare acestor modele în raport cu celelalte modele, indicând o tendință către prețuri mai ridicate în aceste categorii.

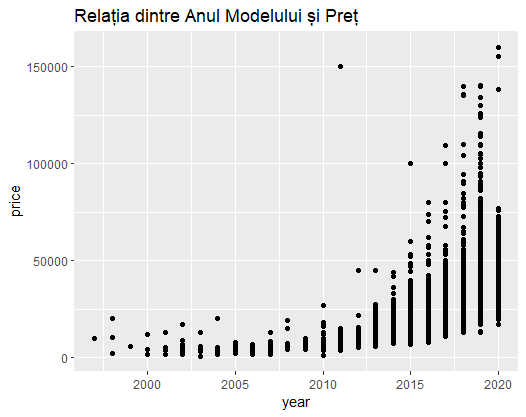


Figura 5. Relația dintre Preț și Anul Modelului

Acest grafic explorează relația strânsă dintre anul modelului și prețurile mașinilor Mercedes. Punctele indică o corelație evidentă, sugerând că mașinile mai recente au prețuri mai ridicate, evidențiind importanța anului modelului în determinarea valorii acestora.

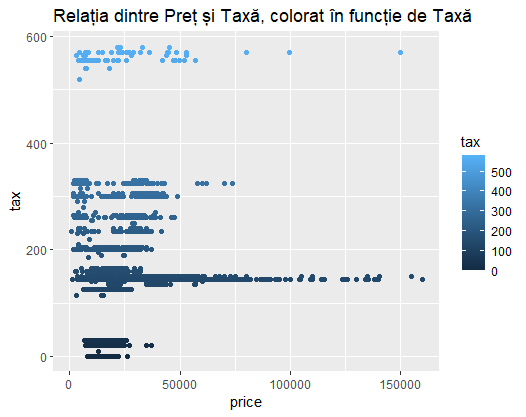


Figura 6. Relația dintre Preț și Taxă, colorat în funcție de Taxă

Acest grafic explorează conexiunea dintre prețul și taxa mașinilor Mercedes, ilustrând distribuția lor. Diferitele culori reprezintă nivelurile de taxă, evidențiind variații semnificative în funcție de preț și taxă.

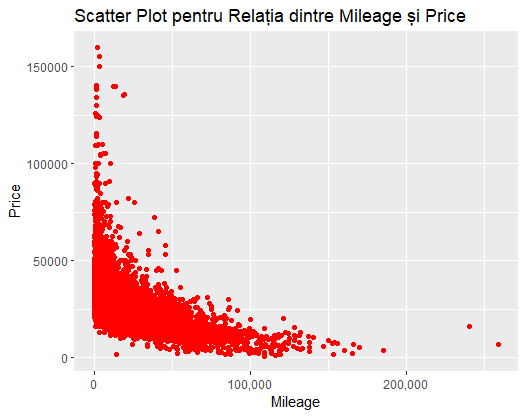


Figura 7. Scatter Plot pentru Relația dintre Mileage și Price

Acest scatter plot evidențiază relația între kilometrajul mașinilor Mercedes și prețurile lor. Punctele roșii indică această asociere, sugerând o relație strânsă, unde mașinile cu kilometraj mai mic tind să aibă prețuri mai ridicate și viceversa.

Interpretarea Rezultatelor:

Am folosit tehnici de învățare automată, cum ar fi regresia liniară, random forest, gradient boosting, pentru a dezvolta modele predictive pentru prețurile mașinilor Mercedes uzate. Am ajustat și optimizat parametrii modelelor pentru a obține performanțe optime. Am împărțit setul de date într-un set de antrenare și un set de testare pentru a valida modelele. Am utilizat metrici precum eroarea medie pătratică (MSE) și coeficientul de determinare (R-squared) pentru a evalua performanța modelelor noastre. Am interpretat rezultatele obținute și am identificat influențele semnificative ale diferitelor caracteristici asupra prețurilor mașinilor uzate Mercedes. Această etapă a fost esențială pentru a furniza înțelegere și context în spatele proiecțiilor predictive.

Rezultatele la regresia liniara:

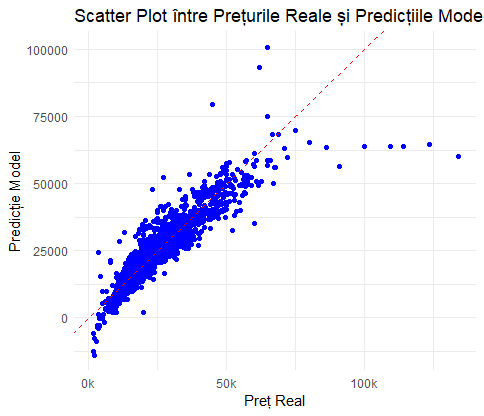
****

Figura 8. Scatter Plot între Prețurile Reale și Predicțiile Modelului Regresie Liniara

Rezultatele analizei modelului de regresie liniară arată că acesta explică aproximativ 76.36% din variația prețurilor automobilelor în setul de date de testare. Acest procent, denumit R-squared, indică cât de bine variabilitatea în prețuri poate fi capturată de caracteristicile incluse în model. În cazul regresiei liniare tale, R-squared este 0.7636, ceea ce înseamnă că aproximativ 76.36% din variabilitatea prețurilor este explicată de caracteristicile precum mileage, tax, mpg, engineSize, year, categorizat\_numeric, și l\_per\_100km. Residual standard error, o măsură a dispersiei reziduurilor (diferența dintre valorile prezise și cele observate), este în acest caz 5819. Aceasta reprezintă deviația standard a reziduurilor și indică cât de aproape sunt valorile prezise de prețurile reale.

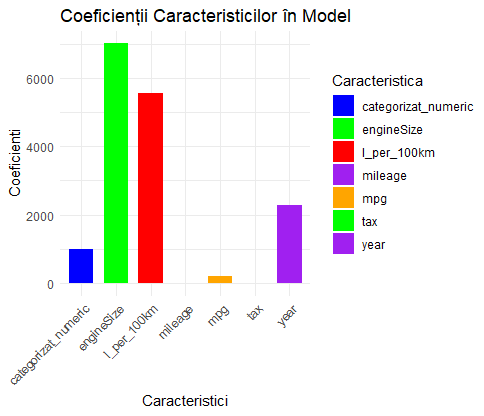


Figura 9. Coeficienții Caracteristicilor în Model pentru Regresie Liniara

Modelul de regresie liniar arată că variabilele precum kilometrajul, impozitul, capacitatea motorului și anul de fabricație au o influență semnificativă asupra prețului mașinii.

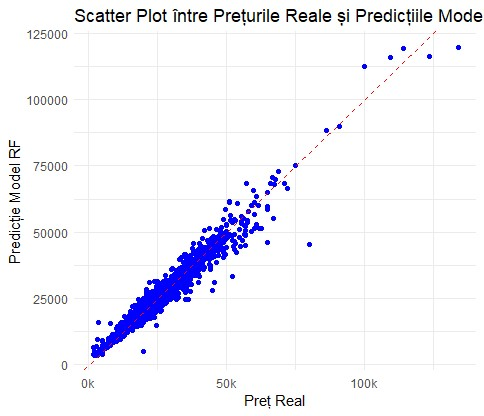


Figura 10. Scatter Plot între Prețurile Reale și Predicțiile Modelului Random Forest

Modelul Random Forest a fost antrenat și evaluat pe setul de date Mercedes pentru a realiza predicții privind prețurile automobilelor. Acesta a fost construit folosind caracteristici precum kilometrajul, anul de fabricație, impozitul, dimensiunea motorului și altele. Rezultatele arată că modelul are un Mean Squared Error (MSE) de 6139636, care reprezintă media valorilor pătratice ale diferențelor dintre prețurile reale și predicțiile modelului. De asemenea, R-squared (coeficientul de determinare) este de 0.9516009, indicând că aproximativ 95.16% din variația prețurilor este explicată de model.

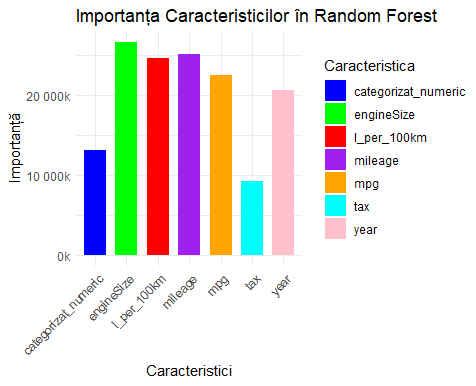


Figura 11. Coeficienții Caracteristicilor în Model pentru Random Forest

Kilometrajul și anul de fabricație sunt cele mai influente în modelul Random Forest pentru prezicerea prețului mașinilor. Aceste variabile au o contribuție semnificativă, indicând importanța lor în procesul de luare a deciziilor al algoritmului.

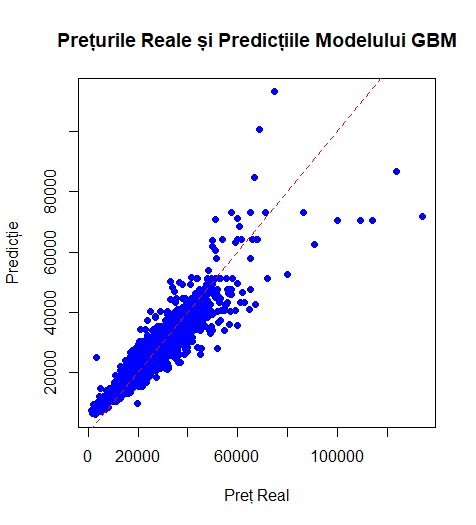


Figura 12. Scatter Plot între Prețurile Reale și Predicțiile Modelului Gradient Boosting

Rezultatele analizei modelului Gradient Boosting Machine (GBM) indică că engineSize, mileage și year sunt cele mai influente caracteristici în predicția prețurilor automobilelor, contribuind semnificativ la variația acestora. Acest lucru sugerează că specificațiile tehnice ale vehiculelor, precum dimensiunea motorului, kilometrajul și vechimea, au un impact semnificativ asupra evaluării prețului. Cu un R-squared de 0.8189, se constată că modelul GBM explică aproximativ 81.89% din variația totală a prețurilor în setul de date de testare.

Validarea și Evaluarea Modelului:

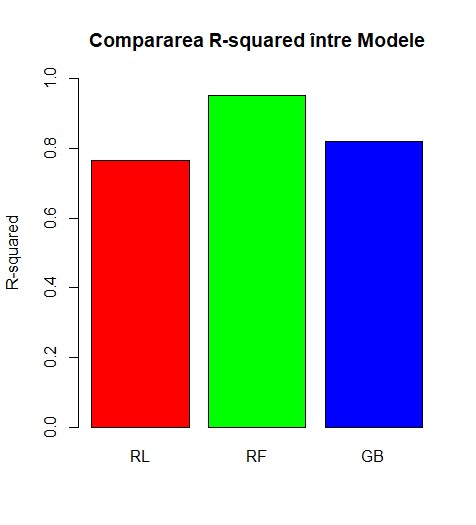


Figura 13. Compararea R-squared între Modele

În procesul de construire a modelelor pentru prețurile mașinilor, am evaluat trei tehnici de regresie distincte: regresia liniară, Random Forest și Gradient Boosting. Fiecare dintre acestea aduce propriile sale avantaje și considerații. Modelul de regresie liniară oferă o abordare simplă și interpretabilă, presupunând relații liniare între caracteristici și preț. Analiza rezultatelor indică faptul că variabile precum 'engineSize', 'year', și 'mileage' au o influență semnificativă asupra prețului mașinii, iar modelul are un R-squared de 0.76, semnificativ pentru un model liniar. Random Forest, un model bazat pe ansamblu de arbori de decizie, surprinde relații complexe în setul de date și obține un R-squared remarcabil de 0.95. Totodată, importanța caracteristicilor indică că 'engineSize', 'mileage' și 'l\_per\_100km' au avut o contribuție semnificativă. Gradient Boosting, prin antrenarea succesivă a modelelor slabe, furnizează o precizie și mai mare, cu un R-squared de 0.82. Aici, 'engineSize', 'mileage' și 'year' sunt caracteristici cheie în determinarea prețului.

În evaluarea performanței, Random Forest pare să ofere un echilibru optim între interpretare și precizie, cu un MSE de 6,139,636. În alegerea celui mai bun model, trebuie să luăm în considerare contextul specific al proiectului și să ponderăm importanța interpretabilității în raport cu precizia predicțiilor. Aceste modele oferă un instrumentar versatil, fiecare potrivindu-se anumitor scenarii și obiective.

Modelul de regresie liniară pentru prețurile mașinilor Mercedes uzate evidențiază câteva caracteristici cheie. Mărimea motorului se dovedește a fi cel mai influent factor, urmat de kilometraj și anul de fabricație. Eficiența energetică și nivelul de taxe completează lista factorilor relevanți. Aceste caracteristici oferă potențialilor cumpărători ghiduri valoroase în alegerea unei mașini uzate, evidențiind importanța performanței, eficienței și costurilor asociate. Este esențială înțelegerea acestor factori pentru a face achiziții informate și pentru a reflecta schimbările generale din domeniul mașinilor uzate Mercedes.

**CONCLUZIE**

În concluzie privind proiecțiile predictive pentru prețurile mașinilor Mercedes uzate, am efectuat o analiză exhaustivă a trei modele diferite de regresie: regresie liniară, random forest și gradient boosting. Fiecare dintre aceste modele a fost antrenat și evaluat pe seturi de date distincte pentru a estima prețurile vehiculelor. Rezultatele au evidențiat importanța anumitor caracteristici, cum ar fi motorul, kilometrajul și anul de fabricație, în determinarea prețurilor. Aceste constatări au implicații semnificative pentru cumpărătorii de mașini uzate, oferindu-le o înțelegere mai profundă a factorilor cheie care influențează prețurile. În concluzie, proiectul nostru aduce o contribuție valoroasă în domeniul proiecțiilor predictive pentru piața mașinilor uzate și oferă informații utile pentru consumatori în procesul decizional de achiziție.

**BIBLIOGRAFIE**

1. 30.11.2023 cele mai dorite modele de mercedes benz second hand [https://ziarulclujean.ro/care-sunt-cele-mai-dorite-modele-de-mercedes-benz-second-hand-in-romania/].
2. 14.10.2023 Mercedes Used Car Listing [https://www.kaggle.com/datasets/mysarahmadbhat/mercedes-used-car-listing ].