UNIVERSITATEA BABEȘ-BOLYAI

FACULTATEA DE ȘTIINȚE ECONOMICE ȘI GESTIUNEA AFACERILOR

PROIECT – FUNDAMENTE DE BIG DATA

Analiza și predicția preturilor

pentru mașinile Mercedes

STUDENT: DANCI ANCA DENISA

2022

1. **INTRODUCERE**

Datorită numărului fără precedent de mașini achiziționate și vândute, predicția prețului mașinilor second hand este un subiect de interes ridicat. Accesibilitatea mașinilor folosite în țările în curs de dezvoltare a determinat tendința oamenilor de a cumpăra mai mult mașini la mâna a doua. De cele mai multe ori vânzătorii de mașini profită de acest scenariu enumerând prețuri nerealiste din cauza cererii crescute. Prin urmare, apare necesitatea unui model care sa poată atribui un preț corect pentru un autovehicul luând în calcul prețurile altor masini, acesta fiind dependent de o multitudine de factori precum anul fabricației, rulajul, s.a.m.d .

Acest proiect își propune să ofere modele de predicție a prețurilor, pentru a ajuta la ghidarea persoanelor care caută sa cumpere sau să vândă mașini și pentru a le oferi o perspectivă mai bună asupra sectorului auto, cu precădere marca Mercedes. Cumpărarea unei mașini second hand de la un dealer poate fi o experiență frustrantă și neplacută, deoarece unii dealeri sunt cunoscuți că recurg la tactici înșelatoare pentru a încheia o vânzare. Prin urmare, pentru a ajuta consumatorii să evite să cadă victimele unor astfel de tactici, acest studiu îi va îndruma să facă alegerea optimă când vine vorba de achiziții, având la dispoziție date reale pe baza cărora se va face o predicție de preț. Un alt obiectiv al proiectului este de a explora noi metode de evaluare a prețurilor mașinilor folosite și de a le compara acuratețea.

Întrebările de cercetare propuse sunt următoarele:

1. Există o legatură între prețul unei masini și anul fabricației, rulaj, taxe , marimea motorului și consum?
2. Acești factori influențează în mod egal prețul unei masini?
3. Este posibilă realizarea unei estimari de pret în situația în care se cunosc acești factori?
4. Ce model prezintă acuratețea cea mai ridicată?
5. **SETUL DE DATE**

Pentru a realiza analiza problemei propuse, mai exact predicția preturilor pentru masinile Mercedes, am recurs la alegerea unui dataset care listează mașini Mercedes second hand împreună cu detaliile aferente acestora.

Sursa setului este https://www.kaggle.com/datasets/adityadesai13/used-car-dataset-ford-and-mercedes.

Setul de date conține 6 coloane:

* Year: anul fabricației
* Price: prețul în lire sterline
* Milage: rulajul( 1mil = 1,61 km)
* Tax: taxe
* Mpg: mile per gallon(1 galon=3,79 l)
* Engine size: mărimea motorului exprimată în l

Chart, scatter chart

Description automatically generated

Coloana de preț va fi coloana țintă în analiză noastră necesară predictiei dar pentru a realiza acest obiectiv avem nevoie și de celelalte variabile din cauza faptului că există depedente între acestea și preț .

1. **REZULTATE ȘI DISCUȚII**

Pentru a vizualiza legătura din variabile numerice și preț am realizat o matrice de corelație.

A picture containing calendar

Description automatically generated

Fiecare rând și coloana reprezintă câte o variabilă, iar fiecare valoare din matrice reprezintă coeficientul de corelație dintre rândul și coloana corespondentă. Dacă valoarea este aproape de -1, înseamnă că există o corelație negativă între valori. De exemplu: kilometraj și an. Desigur, kilometrajul crește pe măsură ce mașina este condusă. În mod similar, dacă valoarea este aproape de 1 înseamnă că există o corelație pozitivă între valori.

Corelația prețului cu variabilele independente:

* Prețul are o corelație pozitivă cu anul de fabricație, mărimea motorului și taxe.
* Prețul are o corelație negativă cu kilometrajul și consumul.

Constatăm că raportat la întrebările de cercetare, variabila de ieșire va fi numerică. Astfel, suntem într-o ipoteză în care încercăm să realizăm o predicție numerică.

În consecință, pentru a stabili dacă anul fabricației, rulajul, taxele , marimea motorului și consumul sunt variabile independente care influențeaza nivelul prețului unei mașini, precum și relația dintre acestea, am ales metoda celor mai mici patrate, respectiv regresia liniara. Această decizie este justificată de vizualizarea matricei de corelație, care ne-a dat o perspectivă asupra legaturilor dintre variabile, această legatură putând fi liniară. Scopul regresiei liniare este de a determina o dependență liniară între variabilele de intrare și de ieșire . Altfel spus, se obține o funcție f care determina relatia dintre Y(price) si X=(X1, X2,..., Xp) (year, mileage, tax, mpg, engineSize) astfel încât Y= f(X)+ε.

Vom realiza o presupunere asupra formei lui f, aceea că aceasta este liniară, astfel că Y = β0 + β1 X1 + ... + βp Xp. Vom face demersuri pentru a potrivi acest model pe datele de antrenament, urmând să obținem estimări pentru β0 , β1,..., βp.

Variabila dependentă va fi caracterizată de coloana de preț, cea care necesită a fi prezisă pe baza variabilelor independente: year, mileage, tax, mpg și engineSize.

Pentru început, am analizat prețul(price) în funcție de anul fabricației(year), considerând

price ≈ β0 + β1 x year, unde β0 reprezintă intercept, iar β1 panta.

Table

Description automatically generated

Astfel, dacă ținem cont doar de anul fabricației price=-5.257e+06 +2617.7805 x year.

Sunt stabilite și intervalele de încredere pentru parametrii βi, cu o încredere de 95%, utilizând formula βi [βˆ ∈ i − 2SE(βˆi ), βˆi + 2SE(βˆi )] :

* β0 : [-5.38e+06 , -5.13e+06]: prețul este situat în acest interval daca nu se ține cont de anul fabricației
* β1 : [2556.366 , 2679.195]: daca se ține cont de anul fabricației, se va determina o creștere a pretului între [2556.366-2679.195]

Având aceste informații, am trecut la testarea ipotezelor, respectiv:

* H0– Ipoteza nulă: Nu se identifică nicio relație între X (year) și Y (price) – testarea H0 utilizând presupunerea β1 = 0
* Ha – Există o relație între X (year) și Y (price) – testarea H1 utilizând presupunerea β1 =/= 0

Observăm că valoarea lui p value este 0, ceea ce infirmă ipoteza nula, astfel că relația dintre X (year) și Y (price) nu se datorează șansei.

Am continuat analizând prețul(price) în funcție de rulaj(mileage), considerând

price ≈ β0 + β1 x mileage, unde β0 reprezintă intercept, iar β1 panta.

A picture containing table

Description automatically generated

Astfel, dacă ținem cont doar de rulaj price=3.022e+04 -0.2776 x mileage, aceasta aratând o relație invers proporțională: cu cât rulajul crește, prețul va scădea.

Sunt stabilite și intervalele de încredere pentru parametrii βi, cu o încredere de 95%, utilizând formula βi [βˆ ∈ i − 2SE(βˆi ), βˆi + 2SE(βˆi )] :

* β0 : [3e+04 , 3.04e+04]: prețul este situat în acest interval dacă nu se ține cont de rulaj
* β1 : [-0.2776 , 0.003]: daca se ține cont de rulaj, se va determina o creștere a pretului între [-0.2776 , 0.003]

Având aceste informații, am trecut la testarea ipotezelor, respectiv:

* H0– Ipoteza nulă: Nu se identifică nicio relație între X (mileage) și Y (price) – testarea H0 utilizând presupunerea β1 = 0
* Ha – Există o relație între X (mileage) și Y (price) – testarea H1 utilizând presupunerea β1 =/= 0

Observăm că valoarea lui p value este 0, ceea ce infirmă ipoteza nula, astfel că relația dintre X (mileage) și Y (price) nu se datorează șansei.

Pentru a vedea cum influențează taxele(tax) nivelul prețului(price) am considerat price ≈ β0 + β1 x tax, unde β0 reprezintă intercept, iar β1 panta.

Table

Description automatically generated

Astfel, dacă ținem cont doar de taxe price=1.827e+04 +44.7594 x tax.

Sunt stabilite și intervalele de încredere pentru parametrii βi, cu o încredere de 95%, utilizând formula βi [βˆ ∈ i − 2SE(βˆi ), βˆi + 2SE(βˆi )] :

* β0 : [ 1.79e+04, 1.86e+04 ]: prețul este situat în acest interval dacă nu se tine cont de taxe
* β1 : [42.272, 47.246]: daca se ține cont de taxe, se va determina o creștere a prețului între [42.272, 47.246]

Având aceste informații, am trecut la testarea ipotezelor, respectiv:

* H0– Ipoteza nulă: Nu se identifică nicio relație între X (tax) și Y (price) – testarea H0 utilizând presupunerea β1 = 0
* Ha – Există o relație între X (tax) și Y (price) – testarea H1 utilizând presupunerea β1 =/= 0

Observam ca valoarea lui p value este 0, ceea ce infirmă ipoteza nulă, astfel că relația dintre X (mileage) și Y (price) nu se datorează șansei.

Analizând relația dintre consum(mpg) si pret(price) avem: price ≈ β0 + β1 x mpg, unde β0 reprezintă intercept, iar β1 panta.

Table

Description automatically generated

Astfel, dacă ținem cont doar de taxe price=3.926e+04 -273.9491 x mpg.

Sunt stabilite și intervalele de încredere pentru parametrii βi, cu o încredere de 95%, utilizând formula βi [βˆ ∈ i − 2SE(βˆi ), βˆi + 2SE(βˆi )] :

* β0 : [ 3.87e+04, 3.99e+04 ]: prețul este situat în acest interval dacă nu se ține cont de consum
* β1 : [-284.213, -263.685]: daca se ține cont de consum, se va determina o creștere a prețului între [-284.213, -263.685]

Având aceste informații, am trecut la testarea ipotezelor, respectiv:

* H0– Ipoteza nulă: Nu se identifică nicio relație între X (mpg) și Y (price) – testarea H0 utilizând presupunerea β1 = 0
* Ha – Există o relație între X (mpg) și Y (price) – testarea H1 utilizând presupunerea β1 =/= 0

Observăm că valoarea lui p value este 0, ceea ce infirmă ipoteza nula, astfel că relația dintre X (mpg) și Y (price) nu se datorează șansei.

Privind relația dintre mărimea motorului(engineSize) și preț(price) avem: price ≈ β0 + β1 x engineSize, unde β0 reprezintă intercept, iar β1 panta.

Table

Description automatically generated

Astfel, dacă ținem cont doar de marimea motorului price=7912.49 -7880.65 x engineSize.

Sunt stabilite și intervalele de încredere pentru parametrii βi, cu o încredere de 95%, utilizând formula βi [βˆ ∈ i − 2SE(βˆi ), βˆi + 2SE(βˆi )] :

* β0 : [ 7301.454, 8523.527]: pretul este situat în acest interval dacă nu se ține cont de marimea motorului
* β1 : [7592.281, 8169.024]: dacă se ține cont de mărimea motorului, se va determina o creștere a prețului între [7592.281, 8169.024]

Având aceste informații, am trecut la testarea ipotezelor, respectiv:

* H0– Ipoteza nulă: Nu se identifică nicio relație între X (mpg) și Y (price) – testarea H0 utilizând presupunerea β1 = 0
* Ha – Există o relație între X (engineSize) și Y (price) – testarea H1 utilizând presupunerea β1 =/= 0

Observam ca valoarea lui p value este 0, ceea ce infirmă ipoteza nulă, astfel ca relația dintre X (engineSize) și Y (price) nu se datorează șansei.

Chart, scatter chart

Description automatically generated

Informațile culese din dezvoltarea modelelor de mai sus au determinat concluzia că există o legătură între variabilele independente și nivelul prețului. În consecință, am consifderat a fi utilă  folosirea metodei de regresie liniara cu mai multi predictori(year, mileage, tax, mpg, engineSize). Astfel, functia va lua forma: price ≈ β0 + β1 x year + β2 x mileage + β3 x tax + β4 x mpg+ β5 x engineSize, obținând valorile:

Table

Description automatically generated

În continuare am testat ipotezele, acestea fiind:

* H0 – Ipoteza nulă: Nu se identifică nicio relație între predictori și variabila dependentă – β1 = β2= β3=β4= β5
* Ha – Ipoteza alternativă – există cel puțin unul dintre βi care îndeplinește condiția βi =/= 0

Luând în considerare valoarea lui F statistics care este mult mai mare că 1 (este 6492) și valorile lui p value fiind 0, respingem ipoteza nulă și o acceptăm pe cea alternativă.Astfel, funcția va lua forma: Y(price)=-4.08e+06+2030.3729 x year -0.1276 x mileage -3.1236 x tax -125.7980 x mpg + 8964.5823 x engineSize.

Am încercat să prezicem valoarea prețului unei mașini din anul 2020, cu rulaj de 1000, taxa de 140, consum 29.2 și mărimea motorului de 4.0, astfel încât am obținut valoarea de 53030.88.

Graphical user interface, text, application

Description automatically generated

Având la dispoziție informațiile la care ne-am referit anterior, putem formulă răspunsurile la întrebările formulate:

1.      Există o legătură între anul fabricației, rulaj, taxe, consum, mărimea motorului și nivelul prețului.

2.      Aceste variabile influențează într-o proporție diferită valoarea prețului.

3.      Se poate realiza o estimare de preț în condițiile în care se cunosc anul fabricației, rulaj, taxe, consum și mărimea motorului.

În continuarea am încercat și alți algoritmi utili pentru prezicerea prețului.

Necesar pașilor următori este împărțirea setului de date. Train-test split este o tehnică de evaluare a performanței unui algoritm de învățare automată. Procedura presupune luarea unui set de date și împărțirea lui în două subseturi. Primul subset este utilizat pentru a se potrivi modelului și este denumit setul de date de antrenament. Al doilea subset nu este folosit pentru a antrena modelul, în schimb, elementul de intrare al setului de date este furnizat modelului, apoi se fac predicții și se compară cu valorile așteptate. Acest al doilea set de date este denumit setul de date de testare. Datele vor fi împărțite in 0.33% pentru testare, și 67% pentru antrenament.

Graphical user interface, text, application, email

Description automatically generated

Regresia arborelui de decizie observă caracteristicile unui obiect și antrenează un model în structura unui arbore pentru a prezice date în viitor pentru a produce rezultate continue semnificative. Ieșirea continuă înseamnă că ieșirea/rezultatul nu este discret, adică nu este reprezentat doar de un set discret și cunoscut de numere sau valori.

Graphical user interface, text, application

Description automatically generated

Am încercat prezicerea prețului folosind arborele de decizie, această determinând o acuratețe de 85,14% și o eroare medie pătratică de 3704.54.

Coeficientul de determinare (R²) este o măsură statistică într-un model de regresie care determină proporția de varianță a variabilei dependențe care poate fi explicată prin variabilă independentă. Cu alte cuvinte, coeficientul de determinare spune cât de bine se potrivesc datele cu modelul. Pe de altă parte RMSE ne arată dacă răspunsurile prezise sunt aproape de răspunsurile reale, cu cât este mai mică cu atât potrivirea este mai aproape de setul de date.

Următorul algoritm încercat este Random Forest. Random Forest Regression este un algoritm de învățare supravegheat care utilizează metodă de învățare prin ansamblu pentru regresie.

Metoda de învățare prin ansamblu este o tehnică care combină predicții de la mai mulți algoritmi de învățare automată pentru a face o predicție mai precisă.

Algoritmul Random Forest urmează un proces în doi pași:

* Construiește n regresori ai arborelui de decizie (estimatori).

Numărul de estimatori se regaseste dupa cum se observa mai jos sub forma n\_estimators.

* Predicția medie între estimatori

Fiecare regresie din arbore de decizie prezice un număr ca rezultat pentru o intrare dată. Regresia Random Forest ia media acelor predicții ca rezultat „final”.

Graphical user interface, text, application, email

Description automatically generated

Acesta prezintă o acuratețe de 90,41% și o eroare medie pătrățica de 2975.13.

Analizând și cu rezultatele obținute anterior, putem spune că acesta este modelul cel mai potrivit pentru setul de date, prezentând cea mai mare acuratețe și cea mai mică eroare medie pătratică.

**CONCLUZII**

Așadar, în urmă aplicării mai multor metode pe setul de date (regresie liniară, arbore de decizie, random forest) am constat că există o relație între variabilele independente respectiv anul fabricației, rulaj, taxe, consum, mărimea motorului și nivelul prețului, reușind prezicerea prețului prin intermediul metodelor prezentate anterior. Modelele realizate reprezintă un instrument util pentru a explică datele din setul supus spre analiză.

Astfel, persoanele care doresc să achiziționeze o mașină Mercedes pot cunoaște faptul că, anul fabricației, rulaj, taxe, consum, mărimea motorului vor influență puternic nivelul de preț al mașinii.