Proiect: Fundamente de Big Data

Analiza prețului mașinilor în funcție de dotările tehnice

Participanți:

Matiș Diana Rebeca

Introducere:

Industria auto este o forță industrială și economică majoră la nivel mondial. Produce 60 de milioane de mașini și camioane pe an, iar acestea sunt responsabile pentru aproape jumătate din consumul mondial de petrol, de asemenea are 4 milioane de angajați direct și mulți alții indirecți. Industria auto își păstrează o influență și o importanță foarte puternică, oferă, de asemenea, locuri de muncă bine plătite, are legături puternice cu industriile furnizori (ceea ce îi conferă un rol supradimensionat în dezvoltarea economică) și are o influență politică puternică.

Clienți devin din cel in cel mai pretențioși datorita multiplelor variante de alegere din această industrie si autovehiculele sunt din cel in cel mai performante ca să atragă clienți noi, însă parte din performanță si modernizare se datorează si nevoii de reinventare a unui concept prieten cu miedul înconjurător si cu resursele naturale care sunt tot mai puține. Cunoscând câteva atribute tehnice si prețul mașinilor, proiectul de față își propune să răspundă la câteva întrebări de cercetare:

1. Care este legătura intre factorii care influențează prețul mașinilor?
2. În cazul în care această legătură există, cât de puternică este aceasta?
3. Aceste variabile influențează în aceeași măsură sau în proporții diferite prețul mașinii?
4. Se poate face o estimare a prețului unei mașini in funcție de factorii aleși?
5. Există o sinergie intre factorii relevanți?

Principalele persoane interesate de rezultatele unei astfel de cercetări sunt potențialii clienți si producătorii auto. Consider că aceste întrebări sunt relevante in ambele direcții, atât pentru clienți cât si pentru producătorii auto. Este important pentru ambele părți de știut care sunt atributele tehnice care influențează prețul cel mai mult și in ce direcție îl influențează acestea. Clientul dorește o mașină performantă la preț acceptabil iar producătorul dorește un număr mare de vânzări mizând pe calitate si pe o cerere realistă.

1.Regresia

Pasul întâi in realizarea metodei regresiei este încărcarea datelor si observarea datelor folositoare si nefolositoare. Știind ca variabila dependenta de care ne vom folosi este „price”, prețul autovehiculelor, s-a realizat o selecție a celor mai eficienți predictori independenți pentru a fi folosiți in model. Aceasta selecție este realizata cu procedura de „Selecție înainte (forward)”.

```
> model<-lm(price~turbo+wheelbase+carlength+carwidth+carheight+curbweight+enginesize+boreratio+stroke+compressionratio+horsepower+peakrpm+citympg+highwaympg
+           data=car)
> FWDfit.p<-ols_step_forward_p(model,penter=.05)
> #This gives you the short summary of the models at each step
> FWDfit.p
```

Selection Summary

Step	Variable Entered	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	enginesize	0.7641	0.7630	99.5993	3974.8248	3889.4537
2	curbweight	0.7946	0.7925	62.7874	3948.4845	3638.6587
3	peakrpm	0.8182	0.8154	34.7463	3925.4972	3432.0296
4	stroke	0.8273	0.8239	25.0574	3916.8874	3352.7074
5	compressionratio	0.8364	0.8323	15.4967	3907.8245	3271.6409
6	carwidth	0.8409	0.8361	11.7921	3904.1374	3234.7110
7	horsepower	0.8453	0.8398	8.1229	3900.3249	3197.2610

Selecția înainte începe cu un model care nu conține predictori, apoi adaugă predictorii la model, pe rând, până când toți predictorii sunt în model. La fiecare pas se adaugă la model variabila care oferă cea mai mare îmbunătățire suplimentară potrivirii. In cazul de față variabila care oferă cea mai mare îmbunătățire este „enginesize” la care apoi se mai adaugă si variabilele „curbweight” si „peakrpm”.

O alternativă a Selecției înainte este următoarea funcție care alege cel mai potrivit subset de predictori pentru model:

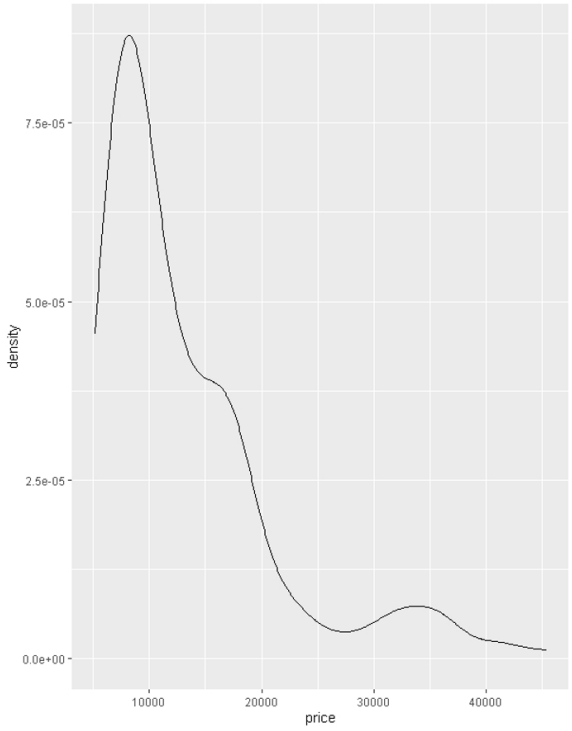
```
> model<-lm(price~wheelbase+carlength+carwidth+carheight+curbweight+enginesize+boreratio+stroke+compressionratio+horsepower+peakrpm+citympg+highwaympg,
+           data=car)
> modcompare<-ols_step_best_subset(model)
> modcompare
```

Best Subsets Regression

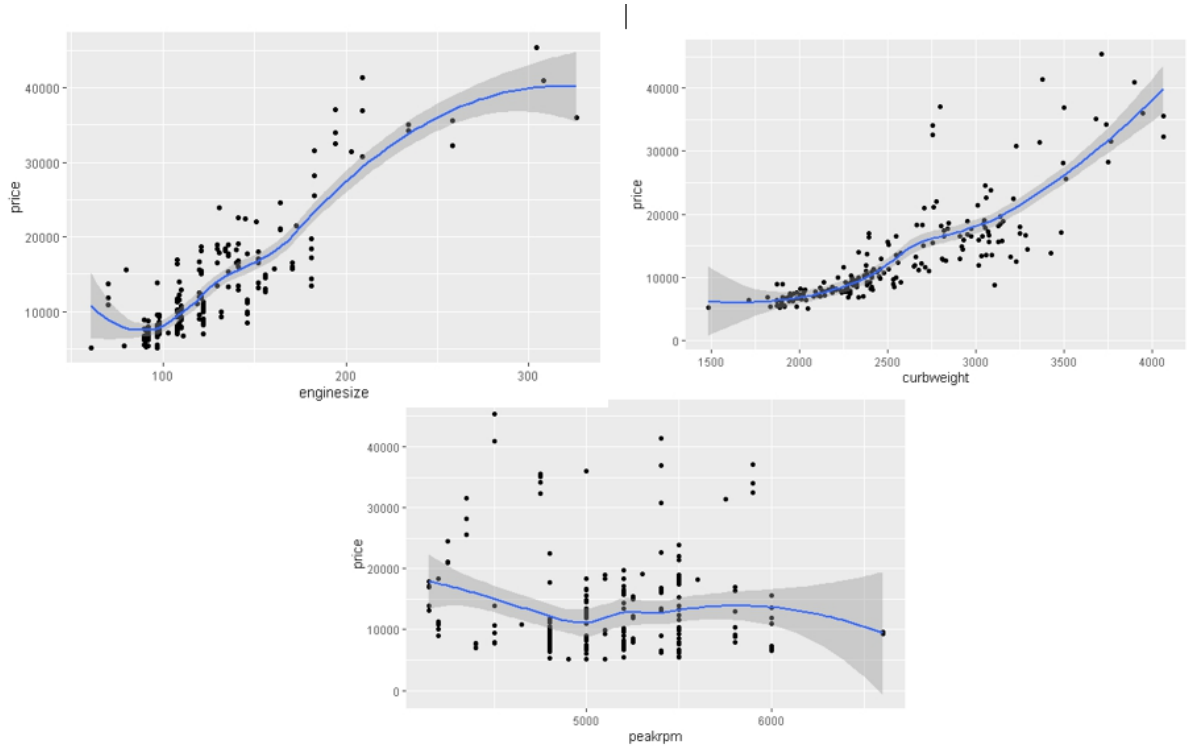
Model Index	Predictors
1	enginesize
2	curbweight enginesize
3	curbweight enginesize peakrpm
4	curbweight enginesize stroke peakrpm
5	curbweight enginesize stroke compressionratio peakrpm
6	carwidth enginesize stroke compressionratio horsepower peakrpm
7	carwidth curbweight enginesize stroke compressionratio horsepower peakrpm
8	carwidth carheight enginesize stroke compressionratio horsepower peakrpm citympg
9	carlength carwidth carheight enginesize stroke compressionratio horsepower peakrpm citympg
10	wheelbase carlength carwidth carheight enginesize stroke compressionratio horsepower peakrpm citympg
11	wheelbase carlength carwidth carheight enginesize stroke compressionratio horsepower peakrpm citympg highwaympg
12	wheelbase carlength carwidth carheight curbweight enginesize stroke compressionratio horsepower peakrpm citympg highwaympg
13	wheelbase carlength carwidth carheight curbweight enginesize boreratio stroke compressionratio horsepower peakrpm citympg highwaympg

Rezultatele sunt aceleași, arătând predictorii cei mai relevanți in funcție de câți predictori dorim in total pentru a fi folosiți in model. In continuarea proiectului așadar se va folosi variabila dependenta ”price” si variabilele independente ”enginesize”, ”curbweight” si ”peakrpm”.

Conform graficului de densitate se constată ca cele mai multe mașini au prețul situat aproximativ la 10.000 de dolari, iar cele mai puține mașini au prețul situat peste 40.000 de dolari.



In următoarea imagine se pot vizualiza relațiile dintre preț cu capacitatea cilindrica a motorului, greutatea proprie și a maximului rotaților pe minut.



1) Price vs. EngineSize: Cele mai multe instanțe se situează sub prețul de 20.000 de dolari cu o capacitatea cilindrica a motorului de aproximativ 100. Relația este destul de lineara.
2) Price vs. Curbweight: Cele mai multe instanțe se situează sub prețul de 20.000 de dolari cu greutatea mașinii de 350u.m. Relația este cel mai lineara dintre toate cele 3 variante.
3) Price vs. PeakRPM: Cele mai multe instanțe se situează sub prețul de 20.000 de dolari cu maximului a rotaților pe minut sub 6000.u.m. Relația nu este una liniară, însă tot influențează prețul.

Regresie liniara simpla – în regresia simpla vom lua fiecare dintre cei 3 parametri de sus individual. Ne va da o concluzie asupra relației dintre un predictor (*enginesize, curbweight, peakrpm*) si variabila dependenta (*price*), in condițiile ignorării celorlalți factori:

```
> mod_price_enginesize <- lm(data = car, price ~ enginesize)
> summary(mod_price_enginesize)

Call:
lm(formula = price ~ enginesize, data = car)

Residuals:
     Min       1Q   Median       3Q      Max
-10664.2  -2225.0   -482.4   1588.0  14271.5

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -8005.446    873.221  -9.168   <2e-16 ***
enginesize    167.698      6.539  25.645   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3889 on 203 degrees of freedom
Multiple R-squared:  0.7641,    Adjusted R-squared:  0.763
F-statistic: 657.6 on 1 and 203 DF,  p-value: < 2.2e-16

> confint(mod_price_enginesize)
                 2.5 %     97.5 %
(Intercept) -9727.1913 -6283.6997
enginesize    154.8047   180.5922
```

```
> mod_price_curbweight <- lm(data = car, price ~ curbweight)
> summary(mod_price_curbweight)

Call:
lm(formula = price ~ curbweight, data = car)

Residuals:
     Min       1Q   Median       3Q      Max
-11604.4  -2042.4   -361.1   1368.3  20618.6

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.948e+04  1.544e+03  -12.61   <2e-16 ***
curbweight   1.282e+01  5.921e-01   21.65   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4403 on 203 degrees of freedom
Multiple R-squared:  0.6977,    Adjusted R-squared:  0.6962
F-statistic: 468.6 on 1 and 203 DF,  p-value: < 2.2e-16

> confint(mod_price_curbweight)
                   2.5 %       97.5 %
(Intercept) -22520.12176 -16431.60300
curbweight      11.64881     13.98353
```

```
> mod_price_peakrpm <- lm(data = car, price ~ peakrpm)
> summary(mod_price_peakrpm)

Call:
lm(formula = price ~ peakrpm, data = car)

Residuals:
   Min     1Q Median     3Q    Max
 -8480  -5498  -2782   3101  31231

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 20595.935   6028.571   3.416 0.000766 ***
peakrpm        -1.428      1.171  -1.219 0.224141
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7979 on 203 degrees of freedom
Multiple R-squared:  0.00727,   Adjusted R-squared:  0.00238
F-statistic: 1.487 on 1 and 203 DF,  p-value: 0.2241

> confint(mod_price_peakrpm)
                   2.5 %       97.5 %
(Intercept) 8709.286882 3.248258e+04
peakrpm       -3.737467 8.812521e-01
```

price ≈ β0 + β1 × enginesize / price ≈ β0 + β1 × curbweight / price ≈ β0 + β1 × peakrpm

unde:

β0: intercept

β1: panta (slope)

Estimarile (≈) au fost calculate pe baza algoritmului de regresie si a datelor de intrare din baza de date.

	~ enginesize	~curbweight	~peakrpm
Std error (valoare absoluta) Cu cat estimarea parametrului diferă de valoarea reala, dorim sa fie cat de mic	β0 diferă cu 873.221 de valoarea reala β1 diferă cu 6.539 de valoarea reala	β0 diferă cu 1544 de valoarea reala β1 diferă cu 0.5921 de valoarea reala	β0 diferă cu 6028.571 de valoarea reala β1 diferă cu 1.171 de valoarea reala
T-value (valoare relativa la coeficient) ne spune numărul de deviații standard cu care estimarea parametrului βi se depărtează de valoarea reala	β0: -9.168 β1: 25.645	β0: -12.61 β1: 21.65	β0: 3.416 β1: -1.219
P-value daca p-value e mare înseamnă ca asocierea se datorează șansei si nu factorului, adică ϵ - daca p-value e mica înseamnă ca asocierea se datorează factorului ci nu șansei *p-value mic indica o relație	Este mai <0.0001 însemnând că asocierea se datorează factorului, indică că este o relație intre price si enginesize	Este mai <0.0001 însemnând că asocierea se datorează factorului, indică că este o relație intre price si curbweight	β0: 0.000766 β1: 0.224141 sunt valori mici însă indică o relație mai slabă decât in celelalte variante

Intervalul de încredere confint CI
Erorile standard sunt folosite pentru a calcula intervalele de încredere pentru parametri βi , cu o încredere de 95%.
Dorim ca intervalul de încredere sa fie cat mai mic.
	Ipoteza nula: Nu se identifica nicio relație intre X si Y, β1 = 0
	
	Exista o relatie intre X si Y, β1≠0
	valori mici ale lui p-values indica faptul ca **rejectam ipoteza nula respectiva**
	
	Intervalul e larg, nu mic.
	`> confint(mod_price_enginesize)`
	`2.5 % 97.5 %`
	`(Intercept) -9727.1913 -6283.6997`
	`enginesize 154.8047 180.5922`
	Noi dorim sa fie cat de mic, trebuie realizate tot mai multe modele ca sa fie tot mai mic, pana la ϵ epsilon, când nu se mai poate micșora.
		Ipoteza nula: Nu se identifica nicio relație intre X si Y, β1 = 0
		
		Exista o relatie intre X si Y, β1≠0
		valori mici ale lui p-values indica faptul ca **rejectam ipoteza nula respectiva**
		
		`> confint(mod_price_curbweight)`
		`2.5 % 97.5 %`
		`(Intercept) -22520.12176 -16431.60300`
		`curbweight 11.64881 13.98353`
		`>`
			Ipoteza nula: Nu se identifica nicio relație intre X si Y, β1 = 0
			
			Exista o relatie intre X si Y, β1≠0
			valori mici ale lui p-values indica faptul ca rejectam ipoteza nula respectiva.
			
			In cazul acesta deoarece valorile sunt p-value sunt mari exista posibilitatea ipotezei nule.
			
			`> confint(mod_price_peakrpm)`
			`2.5 % 97.5 %`
			`(Intercept) 8709.286882 3.248258e+04`
			`peakrpm -3.737467 8.812521e-01`
			`>`

RSE - Eroarea standard reziduala. Măsoară lipsa de potrivire, lack of fit a modelului. Acest RSE vrem sa fie cat mai mic, cu cat e mai mic, cu atâta potrivirea pe date e mai bună.	3889 on 203 degrees of freedom, este cea mai mică dintre cele 3 variante	4403 on 203 degrees of freedom	7979 on 203 degrees of freedom, este cea mai mare dintre cele 3 variante, ceea ce poate indica o lipsă de potrivire
R2 proporția din variabilitatea lui Y care poate fi explicata pe baza lui X. Dorim sa fie cat mai aproape de 1	0.7641 Are o potrivire puternica, valoarea este apropriată de 1	0.6977 Are o potrivire puternica, valoarea este apropriată de 1	0.00727 Are o potrivire mai slabă.
F statistics Ipoteza nula: este acceptata daca F statistics se apropie de 1 Ipoteza alternativa: este acceptata F statistics este (mult) mai mare decât 1	F-statistic: 657.6	F-statistic: 468.6	F-statistic: 1.487

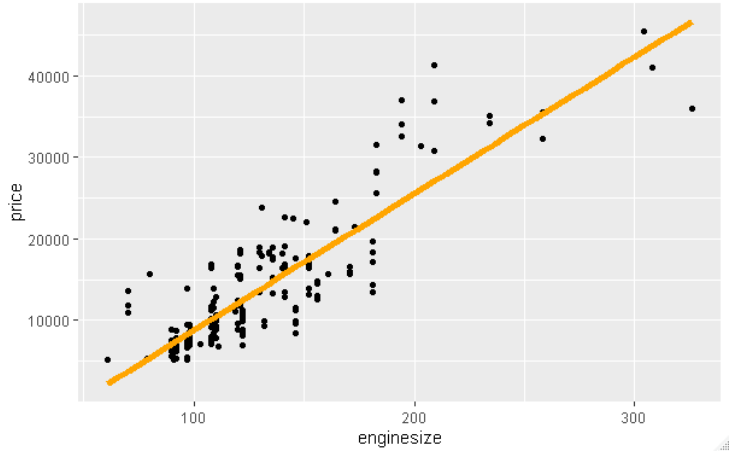
Desenarea datelor si a liniei de regresie:

S-a generat un nou set de date față de setul inițial.

Pentru noua variabilă grid_*enginesize* am luat in considerare 60 de valori din cele inițiale ale lui *enginesize* pentru care vom afla care este valoarea variabilei *price*.

grid_enginesize	60 obs. of 2 variables

Luând in calcul noile valori si aplicând modelul mod_price_enginesize, datele calculate sunt ilustrate in graficul următor:



Așadar pe baza modelului realizat observăm faptul că se poate obține o predicție liniara în ceea ce privește prețul mașinilor introducând o variabilă nouă. Linia de regresie este ilustrată cu portocaliu.

Pentru noua variabilă grid_*curbweight* am luat in considerare 100 de valori din cele inițiale *curbweight* pentru care vom afla care este valoarea variabilei *price*.

grid_curbweight	100 obs. of 2 variables

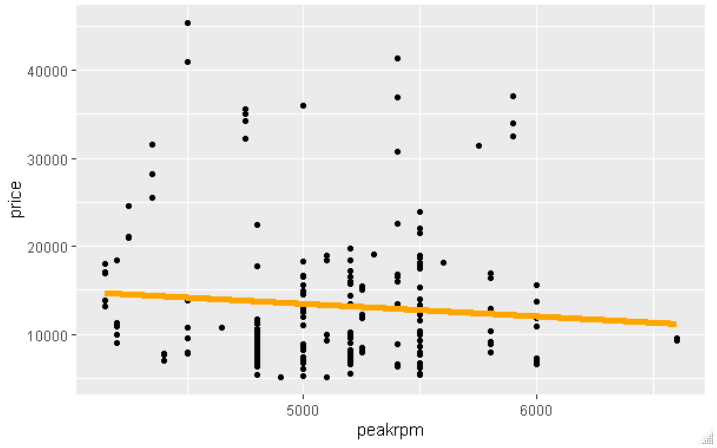
Luând in calcul noile valori si aplicând modelul mod_price_curbweight, datele calculate sunt ilustrate in graficul următor:

Așadar pe baza modelului realizat observăm faptul că se poate obține o predicție în ceea ce privește prețul mașinilor introducând o variabilă nouă. Linia de regresie este ilustrată cu portocaliu.

Pentru noua variabilă grid_*peakrpm* am luat in considerare 1000 de valori din cele inițiale *peakrpm* pentru care vom afla care este valoarea variabilei *price*.

grid_peakrpm	1000 obs. of 2 variables

Luând in calcul noile valori si aplicând modelul mod_price_peakrpm, datele calculate sunt ilustrate in graficul următor:



Așadar pe baza modelului realizat observăm faptul că se poate obține o predicție liniara în ceea ce privește prețul mașinilor introducând o variabilă noua. Linia de regresie este ilustrată cu portocaliu. Având in vedere ca linia de regresie merge in jos, ne indica faptul ca avem de a face cu o regresie negativa.

Regresie liniara multipla - luând in calcul că regresia simpla ne da o concluzie asupra relației dintre un predictor si variabila dependentă, in condițiile ignorării celorlalți factori consider necesar să aplicăm și regresia multiplă care ne da o concluzie asupra relației dintre un predictor si variabila dependenta, in condițiile fixării nivelului celorlalți factori, ca sa vedem diferența relațiilor dintre predictor si variabila in ambele situații.

```
> mod_price_all <- lm(data = car, price ~ enginesize + curbweight + peakrpm)
> summary(mod_price_all)

Call:
lm(formula = price ~ enginesize + curbweight + peakrpm, data = car)

Residuals:
    Min      1Q  Median      3Q     Max
-9479.3 -1781.2    40.1  1324.6 13196.4

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.940e+04  3.262e+03  -9.012  < 2e-16 ***
enginesize   1.156e+02  1.098e+01  10.523  < 2e-16 ***
curbweight   5.606e+00  8.834e-01   6.346 1.44e-09 ***
peakrpm      2.670e+00  5.230e-01   5.104 7.67e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3432 on 201 degrees of freedom
Multiple R-squared:  0.8182,    Adjusted R-squared:  0.8154
F-statistic: 301.4 on 3 and 201 DF,  p-value: < 2.2e-16
```

Model	R2	RSE	F-statistic
Price vs enginesize+peakrpm+curbweight	0.8182	3432	301.4
Având în vedere că valoarea F statistics este mult mai mare decât 1, fiind 301.4 iar p-value are o valoare foarte mică, acceptăm ipoteza alternativă și rejectând ipoteza nulă			

Se poate observa ca toate variabilele au un p-values mic, chiar si peakrpm care in regresia individuala avea p-value mare.			

```
> mod_price_enginesize_curbweight <- lm(data = car, price ~ enginesize + curbweight)
> summary(mod_price_enginesize_curbweight)

Call:
lm(formula = price ~ enginesize + curbweight, data = car)

Residuals:
    Min      1Q  Median      3Q     Max
-8789.9 -1913.3   -50.5  1292.4 14888.8

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.415e+04  1.388e+03 -10.192  < 2e-16 ***
enginesize   1.135e+02  1.163e+01   9.759  < 2e-16 ***
curbweight   5.092e+00  9.305e-01   5.472 1.31e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3639 on 202 degrees of freedom
Multiple R-squared:  0.7946,    Adjusted R-squared:  0.7925
F-statistic: 390.7 on 2 and 202 DF,  p-value: < 2.2e-16
```

Model	R2	RSE	F-statistic
Price vs enginesize+curbweight	0.7946	3639	390.7
Având în vedere că valoarea F statistics este mult mai mare decât 1, fiind 390.7 iar p-value are o valoare foarte mică, acceptăm ipoteza alternativă și rejectând ipoteza nulă			

```
> mod_price_enginesize_peakrpm <- lm(data = car, price ~ enginesize + peakrpm)
> summary(mod_price_enginesize_peakrpm)

Call:
lm(formula = price ~ enginesize + peakrpm, data = car)

Residuals:
     Min       1Q   Median       3Q      Max
-11655.8  -2137.3   -548.3   1991.2  13114.7

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.056e+04  3.224e+03  -6.378 1.20e-09 ***
enginesize   1.741e+02  6.504e+00  26.771  < 2e-16 ***
peakrpm      2.291e+00  5.678e-01   4.035 7.75e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3751 on 202 degrees of freedom
Multiple R-squared:  0.7817,    Adjusted R-squared:  0.7796
F-statistic: 361.7 on 2 and 202 DF,  p-value: < 2.2e-16
```

Model	R2	RSE	F-statistic
Price vs enginesize+peakrpm	0.7817	3751	361.7
Având în vedere că valoarea F statistics este mult mai mare decât 1, fiind 390.7 iar p-value are o valoare foarte mică, acceptăm ipoteza alternativă și rejectând ipoteza nulă			

In concluzie comparând toate modelele, cel mai bun este **Price vs enginesize+peakrpm+curbweight** in care R-square este cel mai mare si RSE cel mai mic.

Așadar pe baza R2 si RSE, decidem ca modelul cu enginesize, peakrpm și curbweight este cel cu acuratețea cea mai buna.

Model	R2	RSE
Price vs enginesize+peakrpm+curbweight	**0.8182**	**3432 on 201**
Price vs enginesize+curbweight	0.7946	3639 on 202
Price vs enginesize+peakrpm	0.7817	3751 on 202

Realizarea predicțiilor:

În calculele următoare se realizează o predicție în care:

enginesize = 350, în care capacitatea cilindrica la motor sa fie de 350u.m

curbweight = 2910, în care greutatea proprie sa fie de 2910u.m

peakrpm = 4750, în care RPM sa fie de 4750u.m

Rezultate:

	Potrivire	Limita inferioară	Limita superioară
Încredere	40041.58	35695.7	44387.46
Predicție	40041.58	31998.92	48084.24

Pentru a răspunde la întrebarea, cât de bună e predicția aceasta ne vom raporta la Y=f(X) obstinând astfel estimarea de încredere cu intervalele aferente.

Dacă ne raportam la Y=f(X)+ ε vom obține intervalul de predicție, care este unul mult mai larg, obținând estimarea de predicție cu intervalele aferente.

Intervalul de încredere se situează intre limita inferioara 35695.7 si 44387.46 limita superioara. Măsoară incertitudinea existenta asupra **valorii medii** a variabilei dependente, adică incertitudinea preturilor medii. Cu o încredere de 95%, valoarea (medie) a preturilor va fi situata intre 35695.7 si 44387.46.

Intervalul de predicție se situează intre limita inferioara 31998.92 si 48084.24 limita superioara. Măsoară incertitudinea existenta asupra unei **valori punctuale** a variabilei dependente. Cu o încredere de 95%, valoarea vânzărilor va fi situata intre 31998.92 si 48084.24.

Intervalul de predicție este mai larg in comparație cu cel de încredere, deoarece reflecta incertitudinea luându-l in calcul pe ε, a predicției realizate pentru o instanță particulara in comparație cu valoarea medie a variabilei dependente, pentru mai multe instanțe.

Adăugarea unui termen de interacțiune:

Prin adăugarea unui termen de interacțiune putem concluziona ca dintre toate modelele testate, cel mai bun model este cel in cate toți factorii contribuie in mod sinergic, nu doar prin condițiile fixării nivelului celorlalți factori.

```
> mod_price_enginesize_curbweight_peakrpm_interaction <- lm(data = car, price ~ enginesize * curbweight * peakrpm)
```

Model	R2	RSE	F-statistic
Price vs enginesize+peakrpm+curbweight	0.8182	3432	301.4
Price vs enginesize+curbweight	0.7946	3639	390.7
Price vs enginesize+peakrpm	0.7817	3751	361.7
Price vs enginesize+peakrpm+curbweight+interacțiune	**0.826**	**3391**	**133.6**

```
lm(formula = price ~ enginesize * curbweight * peakrpm, data = car)

Residuals:
    Min      1Q  Median      3Q     Max
-9121.2 -1902.2  -136.7  1143.1 13661.7

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  8.612e+04  5.181e+04   1.662   0.0981 .
enginesize                  -9.380e+02  4.479e+02  -2.094   0.0375 *
curbweight                  -2.682e+01  1.867e+01  -1.436   0.1525
peakrpm                     -1.865e+01  1.025e+01  -1.820   0.0703 .
enginesize:curbweight        2.967e-01  1.407e-01   2.109   0.0362 *
enginesize:peakrpm           1.921e-01  8.857e-02   2.169   0.0313 *
curbweight:peakrpm           6.099e-03  3.752e-03   1.626   0.1056
enginesize:curbweight:peakrpm -5.475e-05  2.838e-05  -1.929   0.0552 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3391 on 197 degrees of freedom
Multiple R-squared:  0.826,     Adjusted R-squared:  0.8198
F-statistic: 133.6 on 7 and 197 DF,  p-value: < 2.2e-16
```

Dintre toate modelele testate, cel mai bun model este cel in cate toți factorii contribuie in mod sinergic, ci nu prin condițiile fixării nivelului celorlalți factori.

In concluzie in urma realizării a trei regresii individuale, a trei regresii multiple si a unei regresii cu interacțiune putem răspunde la următoarele întrebări de cercetarea menționate în introducerea proiectului:

1. Care este legătura intre factorii care influențează prețul mașinilor?

Există o legătură între capacitatea cilindrica a motorului (enginesize), greutatea proprie a mașinii (curbweight) si rotațiile per minut maxime (peakrpm) in ceea ce privește prețul unei mașini. Alți factori din datele de intrare precum caii putere, consumul de mile per galon, lățimea, lungimea, înălțimea mașinii contribuie la aceasta însă nu sunt factori decisivi, influențând foarte slab prețul mașinii.

2. În cazul în care această relatie există, cât de puternică este aceasta?

Relatia dintre capacitatea cilindrica a motorului (enginesize), greutatea proprie a mașinii (curbweight) si rotațiile per minut maxime (peakrpm) in ceea ce privește prețul unei mașini are un R2 de aproape 80%, cu RSE de 3432. De asemenea F-statistic de 301.4 ne rejectează ipoteza nulă, F statistics este (mult) mai mare decât valoarea 1. Adica este o relatie puternica.

3. Aceste variabile influențează în aceeași măsură sau în proporții diferite prețul mașinii?

Variabilele influențează în proporții diferite prețul mașinii. In regresia individuala peakrpm are cea mai slabă influentă față de enginesize si curbweight. In regresia multiplă enginesize este cea mai

influentă, iar curbweight scade din influenta comparativ cu regresia simplă însă rămâne relevanta cu o valoare p-value mică.

4. Se poate face o estimare a prețului unei mașini in funcție de factorii relevanți?

Da, am exemplificat o astfel de predicție în care am dat valori noi pentru estimare factorilor relevanți. Pentru prezicerea unei valori medii folosim intervalul de încredere iar pentru prezicerea unui răspuns particular folosim intervalul de predicție.

5. Există o sinergie intre factorii relevanți?

Da, prin adăugarea unui termen de interacțiune obținem un R2 mai mare si RSE mai mic, care depășește valorile regresiei multiple, iar valoarea lui p-values indică prezenta unei astfel de sinergii.

2. Arborii de decizie

In primul model de arbore de decizie setul de date a fost împărțit in set de antrenare (70%) si set de test (30%). Algoritmul se va construi pe 143 de instanțe si îl vom testa pe 62 de instanțe.

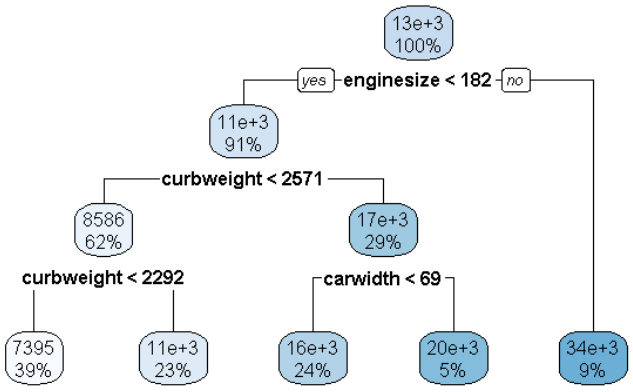
Data	
car	205 obs. of 15 variables
car_split	List of 4
car_test	62 obs. of 15 variables
car_train	143 obs. of 15 variables

S-a folosit metoda ”anova” pentru predicție numerică, luând atributul ținta *price* ~ (against) toate atributele din setul de date.

In funcție de *price* in total de 13.220.310 dolari, va apărea prima condiție *enginesize* care va împărți setul de arborele in doua condiții in funcție de capacitatea cilindrica a motorului, daca este <182 sau >=182. Rezultatul la condiția YES este de 11.105.040 de dolari si la condiția NO de 34373.040, respectiv 91% si 9%. Valorile au fost preluate din afișarea arborelui in mod text.

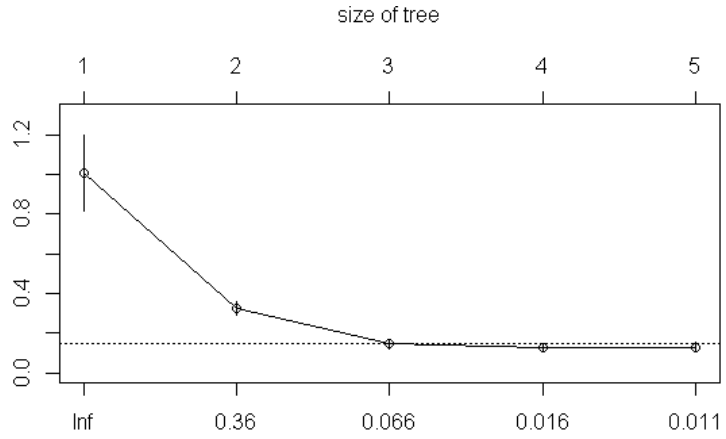
```
n= 143

node), split, n, deviance, yval
      * denotes terminal node

 1) root 143 9339371000 13220.310
   2) enginesize< 182 130 2611626000 11105.040
     4) curbweight< 2571 89  418740500  8586.073
       8) curbweight< 2291.5 56   72993230  7395.205 *
       9) curbweight>=2291.5 33  131561300 10606.940 *
     5) curbweight>=2571 41  402299400 16573.050
      10) carwidth< 68.6 34  257596900 15852.650 *
      11) carwidth>=68.6 7   41351590 20072.140 *
   3) enginesize>=182 13  329385200 34373.040 *
```

Algoritmul a ales care e cea mai buna varianta pe baza metricii folosite cu cel mai mic SSE (sau deviance). După ce s-a găsit cea mai potrivită medie, setul de instanțe este împărțit in cele 2 subseturi, si algoritmul continua recursiv pe fiecare subset. Arborele final are in total 8 noduri si 3 ramuri. Acest arbore ne mai indică si faptul că cele mai importante atribute de care depinde *price* sunt *enginesize, curbweight* deoarece se pozitionează la începutul arborelui.



In continuare putem observa cum valoarea erorii inițiale de 1.0 scade de-a lungul verificării. Pentru nodul întâi ponderea de eroare fiind inițial la 1.0. apoi scade pana la 0.0891

```
          CP nsplit  rel error    xerror       xstd
1 0.68509535      0 1.00000000 1.0088166 0.19122905
2 0.19172451      1 0.31490465 0.3236085 0.03513425
3 0.02293367      2 0.12318015 0.1440728 0.02303909
4 0.01106615      3 0.10024648 0.1267546 0.02217064
5 0.01000000      4 0.08918033 0.1295013 0.02223042
```

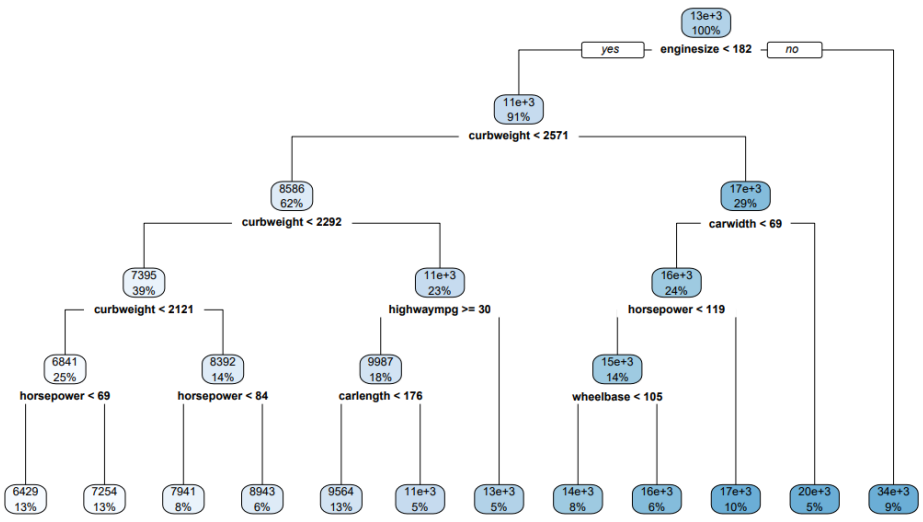


Valoarea cea mai mică a SSE de 0.1267 se înregistrează atunci când α=0.01, iar dimensiunea arborelui este de 4.

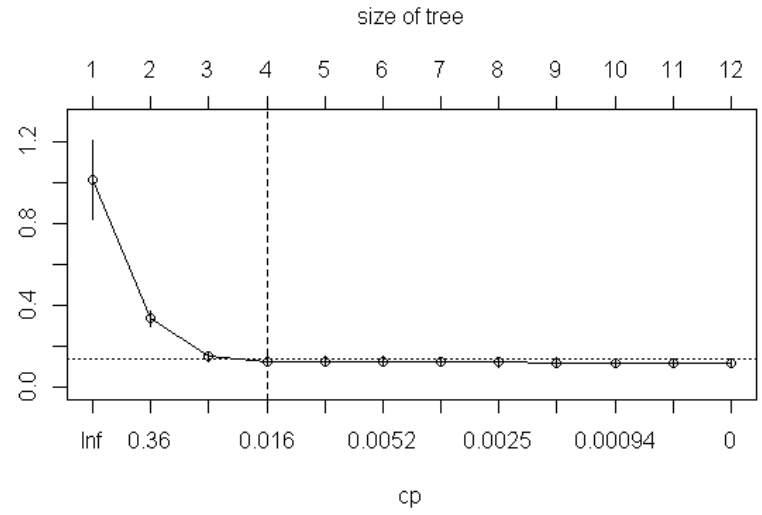
$$minimize\ \{SSE + \alpha|T|\}$$

```
> m1$cptable
          CP nsplit  rel error    xerror       xstd
1 0.68509535      0 1.00000000 1.0088166 0.19122905
2 0.19172451      1 0.31490465 0.3236085 0.03513425
3 0.02293367      2 0.12318015 0.1440728 0.02303909
4 0.01106615      3 0.10024648 0.1267546 0.02217064
5 0.01000000      4 0.08918033 0.1295013 0.02223042
```

In al doilea model de arbore pe care îl voi prezenta este cu modificarea parametrilor folosind cross validation de 10 in care se creste arborele pana la obținerea valorii zero pentru parametrul alpha.



Arborele obținut este unul mult mai mare ca cel anterior, cu un total de 12 împărțiri.




```
> m2$cptable
           CP nsplit  rel error    xerror       xstd
1  0.6850953473      0 1.00000000 1.0094066 0.19189915
2  0.1917245057      1 0.31490465 0.3338160 0.03670035
3  0.0229336665      2 0.12318015 0.1462092 0.02466120
4  0.0110661513      3 0.10024648 0.1246418 0.02386118
5  0.0054575774      4 0.08918033 0.1245025 0.02385675
6  0.0050376855      5 0.08372275 0.1245211 0.02392029
7  0.0033091934      6 0.07868507 0.1220173 0.02335415
8  0.0019243566      7 0.07537587 0.1202197 0.02336173
9  0.0013575401      8 0.07345152 0.1167039 0.02295730
10 0.0006567321      9 0.07209398 0.1148567 0.02291677
11 0.0005317074     10 0.07143724 0.1152167 0.02292549
12 0.0000000000     11 0.07090554 0.1140364 0.02293685
```

Prin aceasta metoda, valoarea SSE scade fata de eroarea inițială fiind de 0.114, iar α=0.00, dimensiunea arborelui fiind de 12.

In al treilea model de arbore pe care îl voi prezenta este cu modificarea parametrilor minsplit si maxdepth in funcție de cele mai optime valori.

```
hyper_grid <- expand.grid(
  minsplit = seq(3, 20, 1),
  maxdepth = seq(6, 15, 1)
)
head(hyper_grid)
models <- list()
for (i in 1:nrow(hyper_grid)) {
  minsplit <- hyper_grid$minsplit[i]
  maxdepth <- hyper_grid$maxdepth[i]
  models[[i]] <- rpart(
    formula = price ~. ,
    data = car_train,
    method = "anova",
    control = list(minsplit = minsplit, maxdepth = maxdepth)
  )
```

Pentru a afla cele mai optime valori s-a construit in hyper_grid care face toate combinațiile posibile intre parametrii minsplit de la 3-20 si maxdepth de la 6-15, rezultând in 180 de observații in hyper_grid.

hyper_grid	180 obs. of 2 variables

Vizualizarea lui:

	minsplit	maxdepth
1	3	6
2	4	6
3	5	6
4	6	6
5	7	6
6	8	6
7	9	6
8	10	6
9	11	6
10	12	6
11	13	6
12	14	6
13	15	6
14	16	6
15	17	6

Pentru fiecare combinație minsplit si maxdepth de mai sus, algoritmul va construi un arbore, in total vor fi 180 de modele.
Pentru fiecare dintre cele 180 de modele se va lua parametrul cp (get_cp) si eroarea minima (get_min_error) prin construirea a doua funcții.

```
get_cp <- function(x) {
  min <- which.min(x$cptable[,"xerror"])
  cp <- x$cptable[min, "CP"]
}
get_min_error <- function(x) {
  min <- which.min(x$cptable[, "xerror"])
  xerror <- x$cptable[min, "xerror"]
}
```

	minsplit	maxdepth	cp	error
1	3	6	0.01106615	0.1184129
2	4	6	0.01106615	0.1159816
3	5	6	0.01106615	0.1196787
4	6	6	0.01106615	0.1187259
5	7	6	0.01106615	0.1282000
6	8	6	0.01106615	0.1202658
7	9	6	0.01106615	0.1161615
8	10	6	0.01000000	0.1128072
9	11	6	0.01106615	0.1208212
10	12	6	0.01106615	0.1259820
11	13	6	0.01106615	0.1154918
12	14	6	0.01106615	0.1181704
13	15	6	0.01106615	0.1172176
14	16	6	0.01106615	0.1155216
15	17	6	0.01106615	0.1194977
16	18	6	0.01106615	0.1166986
17	19	6	0.01106615	0.1111910

Vom obține cele mai mici 5 erori din combinația cu gridul inițial si eroarea cea mai mica.

```
  top_n(-5, wt=error)
minsplit maxdepth          cp     error
       9        7 0.01000000 0.1092584
      10        8 0.01000000 0.1098536
      12        9 0.01106615 0.1109229
      19        6 0.01106615 0.1111910
       4       10 0.01106615 0.1112437
```

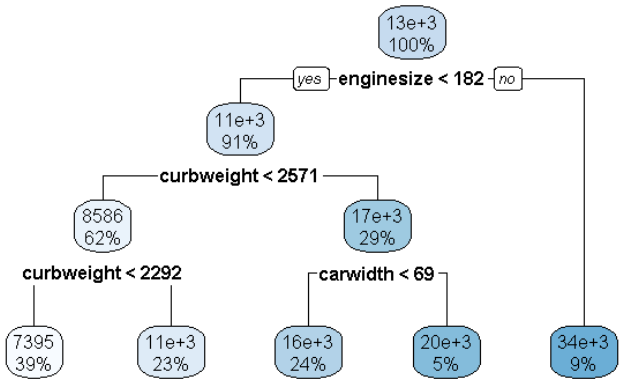
Așadar am obținut valorile de minsplit si maxdepth pentru arborele optim, minsplit=9 si maxdepth=7.

La realizarea predicției a arborelui obținut pe setul de test rezultă:

> RMSE(pred = pred, obs = car_test$price)

[1] 3.078,653 – Predicția este cu o eroare medie de 3.078,653 pe cele 62 de observații din setul de test, nu este o acuratețe foarte bună.

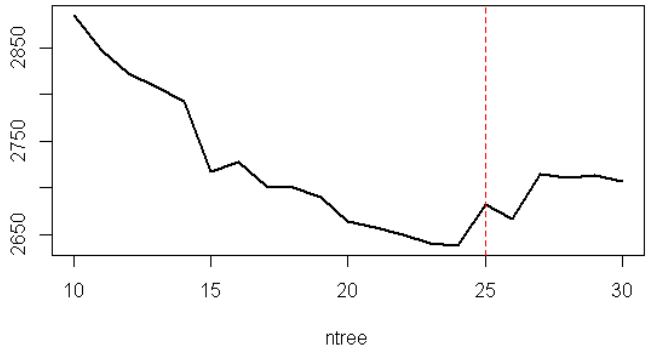
```
> optimal_tree
n= 143

node), split, n, deviance, yval
      * denotes terminal node

 1) root 143 9339371000 13220.310
   2) enginesize< 182 130 2611626000 11105.040
     4) curbweight< 2571 89  418740500  8586.073
       8) curbweight< 2291.5 56   72993230  7395.205 *
       9) curbweight>=2291.5 33  131561300 10606.940 *
     5) curbweight>=2571 41  402299400 16573.050
      10) carwidth< 68.6 34  257596900 15852.650 *
      11) carwidth>=68.6 7   41351590 20072.140 *
   3) enginesize>=182 13  329385200 34373.040 *
```



Valoarea RMSE poate fi diminuata cu comanda Bagging cu scopul de a o reduce:

```
> bagged_m1

Bagging regression trees with 25 bootstrap replications

Call: bagging.data.frame(formula = price ~ ., data = car_train, coob = TRUE)

Out-of-bag estimate of root mean squared error:  2681.425

> pred <- predict(bagged_m1, newdata = car_test)
> RMSE(pred = pred, obs = car_test$price)
[1] 2937.751
```

Se poate observa o îmbunătățire a valorii RMSE de 2.937,751 față de cea inițială de 3.078,653. Pentru calcularea acesteia s-a folosit Baggig cu 25 de bags, pe care ii consider suficienți deoarece eroarea se stabilește si se află in cel mai inferior nivel la un număr <25 de bags, nemaifiind relevant rezultatul din optimizarea procedurii de bagging cu >50 de bags.



In concluzie in urma realizării a trei modele de arbori de decizie, putem răspunde la următoarele întrebări de cercetarea menționate în introducerea proiectului:

1. Care este legătura intre factorii care influențează prețul mașinilor?

Există o legătură între capacitatea cilindrica a motorului (enginesize), greutatea proprie a mașinii (curbweight) in ceea ce privește prețul unei mașini. Alți factori din datele de intrare precum peakrpm, caii putere, consumul de mile per galon, lățimea, lungimea, înălțimea mașinii contribuie la aceasta însă nu sunt factori decisivi, influențând foarte slab prețul mașinii.

2. În cazul în care această relație există, cât de puternică este aceasta?

Relatia dintre capacitatea cilindrica a motorului (enginesize), greutatea proprie a mașinii (curbweight) si in ceea ce privește prețul unei mașini este puternică situându-se in primele noduri ale arborilor.

3. Aceste variabile influențează în aceeași măsură sau în proporții diferite prețul mașinii?

Variabilele influențează în proporții diferite prețul mașinii.

4. Se poate face o estimare a prețului unei mașini in funcție de factorii relevanți?

Da, am exemplificat o astfel de predicție cunoscând enginesize, curbweight si carwidth. Predicția este cu o eroare medie de 3.078,653 pe cele 62 de observații din setul de test, nu este o acuratețe foarte bună.

5. Există o sinergie intre factorii relevanți?

Aceasta metoda nu ne permite verificarea sinergiei.

Concluzie:

În urma metodei regresiei liniare si a metodei arborilor de decizie concluzionăm faptul că ambele indică atributele capacității cilindrice a motorului și greutatea proprie ca fiind cei mai influențabili predictori în ceea ce privește prețul mașinii din toate atributele tehnice prezente in setul de date. Ambele metode sunt valide însă cu o marjă de eroare foarte mare, RSE si RMSE foarte mari, nefiind cele mai fiabile rezultate, ducând ca predicțiile construite pe ambele metode de asemenea să nu fie cele mai fiabile.

Consider ca ambele metode sunt foarte utile în analiza datelor de intrare si extragerea concluziilor utile de afaceri, în special metoda arborilor de decizie care poate fi folosită ca și explicație si unui public necunoscător de data analytics. Sunt o modalitate utilă de a căuta prin cantitatea mare de date și a descoperi relațiile necunoscute.