Universitatea Babes-Bolyai

Facultatea de Ştiinţe Economice şi Gestiunea Afacerilor

*Predicția prețului unei locuințe*

Hurmuz Alicia-Maria (alicia.hurmuz@stud.ubbcluj.ro)

Iaroi Renata-Timea (renata.iaroi@stud.ubbcluj.ro)

Informatică Economică

An 3

**Introducere**

În prezent, numărul persoanelor care vor să își cumpere o locuință proprie, crește semificativ, în special în rândul tinerilor. Locuința este cel mai important factor din viața unei persoane deoarece influențează pozitiv starea de bine, ne oferă siguranță și protecție. De aceea, trebuie să înțelegem factorii care influențează tendința de creștere a prețurilor, prețul fiind important în decizia cumpărării unui imobil.

Am ales să ne concentrăm asupra întrebării:

*Cum influențează numărul dormitoarelor, suprafața în metri pătrați ( m2), numărul băilor, etajele și anul construcției imobilului prețul de cumpărare?*

Acestă întrebare de cercetare are ca și scop oferirea de informații valoroase despre cum factorii de mai sus, influențează stabilirea prețului. Mai mult, informațiile sunt un real interes in achiziționarea unui imobil, pentru a înțelege mai bine piața imobiliarelor și stabilirea deciziei finale în funcție de dorințele și necesitățile viitorilor cumpărători.

Cât de importanți sunt factorii studiați?

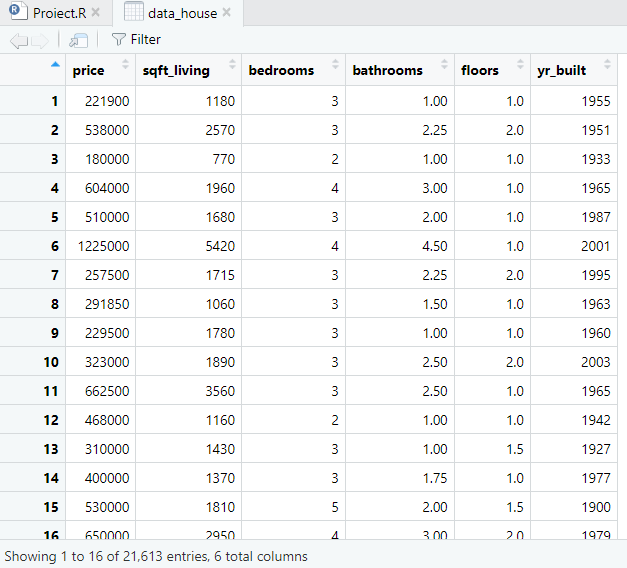
Numărul dormitoarelor este foarte evident, mai multe dormitoare însemnă preț mai mare. Confortul și spațiul suplimentar ridică prețul casei. Suprafața în metri pătrați(m2) este direct proporțională cu prețul. Numărul băilor cresc valoarea locuinței, crește și confortul pentru familii mari. Etajele, în special pentru apartamentele care sunt la nivele superioare prețul poate fi mult mai mare iar anul construcției are un impact semnificativ pentru că tehnologiile de construcție apărute în ultimii ani și modernizarea materialelor de construcție pot influența creșterea prețului.

**Setul de date**

Setul pe care l-am ales este https://www.kaggle.com/datasets/soylevbeytullah/house-prices-dataset. Acesta conține 21 de coloane, dar noi am lăsat doar variabilele relevante pentru înterebarea de cercetare aleasă, acestea fiind:

* **Price**: prețul casei,
* **Sqft\_living**: suprafața locuibilă a casei,
* **Bedrooms**: numărul de dormitoare,
* **Bathrooms**: numărul de băi,
* **Floors**: numărul de etaje,
* **Yr\_built**: anul când a fost construită construcția.

Am adus câteva modificări setului de date, precum eliminarea valorilor nule și apoi selectarea doar a coloanelor care sunt de inters pentru întrebarea noastră de cercetare. Aceste lucruri le-am facut în R Studio.

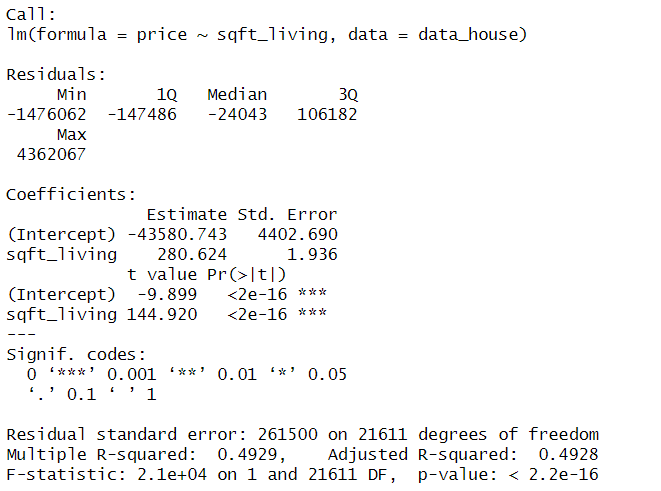


**Rezultate și discuții**

Pentru a răspunde la întrebarea ,,*Cum influențează numărul dormitoarelor, suprafața în metri pătrați ( m2), numărul băilor, etajele și anul construcției imobilului prețul de cumpărare?* vom folosi, în primul rând, **regresia liniară**.

Vom alege variabila dependentă ,,price” pentru toate regresiile pe care le vom efectua în cotinuare.

Prima regresie liniară are variabila independentă ,,sqft\_living”.



Formula: *price ≈ β0 + β1 × sqft\_living*

Interceptul(β0) sugerează că valoarea prețului atunci când sqt\_living ar fi 0.

Coeficientul sqft\_living(β1) de 280,624 sugerează cu câte unități monetare crește prețul casei, în vreme ce suprafața locuibilă crește.

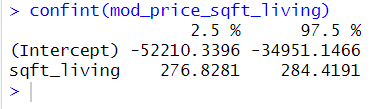
Std. Error ilustrează cu cât diferă interceptul de valoarea reală, adică cu 4402,69 de unități monetare. În cazul coeficientului sqft\_living, eroarea standard e de doar 1,936 unități monetare, fiind o eroare mică. Calculând aceste valori în valoare relativă pentru o concluzie mai clară:

(4402,690/43580,743)x100= 10,10% și (1,936/280,634)x100= 0,69%, sugerând că eroarea pentru coeficientul sqft\_living e mai mică.

T statstic ilustrează cu cât se diferențiază parametrul calculat β de 0.

P value arată probabilitatea ca între predictor și variabila independentă poate exista o asociere datorită șansei. Un p value mare indică faptul că asocirea se datorează șansei. În caz contrar, un p value mic înseamnă ca asocierea se datorează factorului. În cazul modelului nostru, observăm că atât interceptul cât și sqft\_living au o valorile p foarte mici (<2e-16), sugerând că sunt foarte semnificative în ceea ce constă prețul locuinței.

RSE-ul din cadrul acestei regresii, care e 261500, indică cu cât sqft\_living se distanșează de la dreapta regresiei, adică valorile reale(ale prețului), de valorile modelului(pe baza sqft\_living). Această valoare determină R-squared care este 0,4929, sugerând că modelul nostru explică 49,29% din fluctuațiile prețurilor locuințelor în funcție de dimeniunea imobilului.



Intervalele de încredere arată că interceptul se află în intervalul [-52210,34, -34951,15], iar coeficientul sqft\_living se află între [-276,83 , 284,42].

A doua regresie liniară va prelua numărul de dormitoare (,,bedrooms”) ca variabilă independentă.

A screenshot of a computer code

Description automatically generated

Formula: *price ≈ β0 + β1 × bedrooms.*

Interceptul(β0) sugerează că valoarea prețului atunci când bedrooms ar fi 0.

Coeficientul bedrooms(β1) de 1299802 sugerează cu câte unități monetare crește prețul imobilului, în timp ce numărul de dormitoare ale unei locuințe crește..

Std. Error prezintă cu cât diferă interceptul de valoarea reală, adică cu 8932 de unități monetare. În cazul coeficientului bedrooms, eroarea standard e de 2554. Calculând aceste valori în valoare relativă pentru o concluzie mai clară

(8932/129802)x100= 6,88% și (2554/121716)x100= 2,10%, sugerând că eroarea pentru coeficientul bedrooms e mai mică.

T statstic cu cât se diferențiază parametrul calculat β de 0. Ambele valori t arată că sunt diferite de 0.

În cazul modelului nostru, observăm că atât interceptul cât și coeficientul bedrooms au o valorile p foarte mici(<2e-16), sugerând că sunt foarte semnificative în ceea ce constă prețul locuinței.

RSE-ul din cadrul acestei regresii, care e 349200, indică diferența dintre valorile reale(ale prețului), de valorile modelului(pe baza bedrooms). Această valoare determină R-squared care este 0.09508, sugerând că modelul nostru explică 9,508% din fluctuațiile prețurilor locuințelor în funcție de numărul de dormitoare, nefiind foarte sugestiv.

A number and text on a white background

Description automatically generated

Intervalele de încredere arată că interceptul se află în intervalul [112295.2, 147309.5], iar coeficientul bedrooms se află între [116709.5 , 126722.8].

A treia regresie are ca variabilă dependentă numărul de băi (,,bathrooms”).

A screenshot of a computer code

Description automatically generated

Formula: *price ≈ β0 + β1 × bathrooms.*

Interceptul(β0) sugerează că valoarea prețului atunci când bathrooms ar fi 0.

Coeficientul bathrooms (β1) de 250327 sugerează cu câte unități monetare crește prețul casei, în timp ce numărul de băi ale unei locuințe crește.

Valoarea t pentru intercept e 1,724 și nu este semnificativă. În schimb, valoarea t a coeficientului bathrooms este 90,714, fiind foarte semnificativă.

Std. Error prezintă cu cât diferă interceptul de valoarea reală, adică cu 6211 de unități monetare. În cazul coeficientului bathrooms, eroarea standard e de 2760. Calculând aceste valori în valoare relativă pentru o concluzie mai clară:

(6211/10708)x100= 58,02% și (2760/250327)x100= 1,01%, sugerând că eroarea pentru coeficientul bathrooms e cu mult mai mică.

În cazul modelului nostru, observăm că coeficientul bathrooms are valoarea p foarte mică, sugerând că este foarte semnificativ în ceea ce constă prețul locuinței. În schimb, valoarea p a interceptului nu este semnificativă, aceasta fiind 0,0847.

RSE-ul din cadrul acestei regresii, care e 312400, indică diferența dintre valorile reale(ale prețului), de valorile modelului(pe baza bathrooms). Această valoare determină R-squared care este 0.02758, sugerând că modelul nostru explică 27,58% din fluctuațiile prețurilor locuințelor în funcție de numărul de băi, fiind mai sugestiv decât modelul anterior (pe baza coeficientului bedrooms).

A number and numbers on a white background

Description automatically generated

Intervalele de încredere arată că interceptul se află în intervalul [-1465.06, 22881.68], iar coeficientul bedrooms se află între [244917.64, 255735.39].

Continuăm cu alegerea variabilei dependente ,,floors”(etaje).

A computer code with numbers and symbols

Description automatically generated

Formula: *price ≈ β0 + β1 × floors.*

Interceptul(β0) sugerează că valoarea prețului atunci când numărul de etaje ar fi 0.

Coeficientul floors (β1) de 174589 sugerează cu câte unități monetare crește prețul casei, în timp ce numărul de etaje ale unei locuințe crește.

Valorile t pentru intecept și pentru coeficientul floors sunt 39,31 și 39,06, fiind foarte semnificative.

Std. Error prezintă cu cât diferă interceptul de valoarea reală, adică cu 7102 de unități monetare. În cazul coeficientului floors, eroarea standard e de 4470. Calculând aceste valori în valoare relativă pentru o concluzie mai clară:

(7102/279199)x100= 2,57% și (4470/279199)x100= 2,56%, sugerând că erorile sunt asemănătoare.

În cazul modelului nostru, observăm că atât interceptul cât și coeficientul floors au o valorile p foarte mici, sugerând că sunt foarte semnificative în ceea ce constă prețul locuinței.

RSE-ul din cadrul acestei regresii, care e 354800, indică diferența dintre valorile reale(ale prețului), de valorile modelului(pe baza floors). Această valoare determină R-squared care este 0.06594, sugerând că modelul nostru explică 6,59% din fluctuațiile prețurilor locuințelor în funcție de numărul de etaje, nefiind foarte sugestiv.

A number of numbers on a white background

Description automatically generated

Intervalele de încredere arată că interceptul se află în intervalul [265278.4, 293118.7], iar coeficientul floors se află între [165827.2, 183350.8].

Vom continua prin alegerea variabilei dependente ,,yr\_built”(anul constructiei).

A computer screen shot of a number

Description automatically generated

Formula: *price ≈ β0 + β1 × yr\_built.*

Coeficientul yr\_built (β1) de 675,1 sugerează cu câte unități monetare crește prețul casei, în timp ce anul de construcșie ale unei locuințe crește.

Valorile t pentru intecept și pentru coeficientul floors -7,723 și 7,952, fiind semnificative și diferite de 0.

Std. Error prezintă cu cât diferă interceptul de valoarea reală, adică cu 167350,2 de unități monetare. În cazul coeficientului yr\_built, eroarea standard e de 84,9. Calculând aceste valori în valoare relativă pentru o concluzie mai clară:

(167350.2/675,1)x100= 21,17% și (84,9/675,1)x100= 12,57%, sugerând că erorile standard pentru yr\_built e mai mică.

În cazul modelului nostru, observăm că atât interceptul cât și coeficientul floors au o valorile p destul de senmificative.

RSE-ul din cadrul acestei regresii, care e 366600, indică diferența dintre valorile reale(ale prețului), de valorile modelului(pe baza anului construcției). Această valoare determină R-squared care este 0.002917, sugerând că modelul nostru explică doar 0,2917% din fluctuațiile prețurilor locuințelor în funcție de anul construcșiei, nefiind sugestiv.

A close up of numbers

Description automatically generated

Intervalele de încredere arată că interceptul se află în intervalul [-1118496,6760, -462459.0697], iar coeficientul yr\_built se află între [508.6662, 841.4734].

Cele mai relevante grafice ale regresiilor sunt

A graph showing a red line

Description automatically generated A graph with a line going up

Description automatically generated

A graph of a line graph

Description automatically generated

Se poate remarca în mod cert faptul că cel mai puternic predictor este sqft\_living, adică dimensiunea locuinței. Apoi, al doilea cel mai relevant grafic este relația dintre preț și numărul de băi, remarcându-se o relație pozitivă și clară. Nu în ultimul rând, avem relația dintre preț și numărul de camere, având o relație vizibilă, dar nu așa puternică ca celelalte două.

În continuare, vom adauga toate cele 5 variabile independente în regresie și vom analiza rezultatul.

A screenshot of a computer code

Description automatically generated

Interpretari:

* Sqft\_living- Pentru un nivel fix al al celorlalte variabile, coeficientul este de 299.99, însemnând prețul unei locuințe va crește cu 299,29 de unități monetare, pentru fiecare creșterea cu o unitate a suprafeței locuinței.
* Bathrooms- Pentru un nivel fix al al celorlalte variabile, coeficientul este de 68590.53, însemnând prețul unei locuințe va crește cu 68590.53 de unități monetare, pentru fiecare creștere cu o unitate a numărului de băi.
* Floors- Pentru un nivel fix al al celorlalte variabile, coeficientul este de 55863.67, însemnând prețul unei locuințe va crește cu55863.67 de unități monetare, pentru fiecare creștere cu o unitate a numărului de etaje ale locuinței.
* Yr\_built- Pentru un nivel fix al al celorlalte variabile, coeficientul este de -3385.62, însemnând prețul unei locuințe va scădea cu -3385.62 de unități monetare, pentru fiecare an care trece de la construcția locuinței.
* Bedrooms- Pentru un nivel fix al al celorlalte variabile, coeficientul este de -67222.30, însemnând prețul unei locuințe va scădea cu 67222,30 de unități monetare, pentru fiecare creștere cu o unitate a numărului de dormitoare.

În final, putem trage următoarele conzluzii:

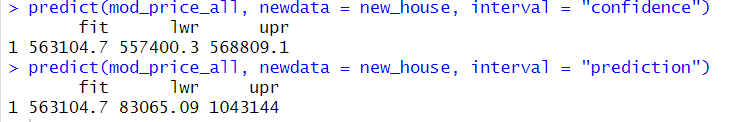
1. Modelul se potrivește în proporție de 55,51% pe datele disponibile, lucru rezultat din valoarea lui R-squared.
2. Datorită valorilor p foarte mici (<2e-16), constatăm că variabilele sunt relevante pentru a prezice variabila dependentă price.
3. În explicarea variabilei dependente price, toți predictorii au o valoare explicativă bună, remarcându-se cel mai mult și pozitiv variabilele sqft\_living, floors și barhrooms.

Până acum am construit 6 modele, 5 regresii simple și una multiplă.

Așadar, putem observa că cel mai bun model până acum este cel de regresie multiplă, având cel mai mare R-squared, cea mai mica valoare RSE și un F-statistic de 5393, valoare îndepărtată de 1, indicând o relație puternică.

Vom continua raportul prin realizarea predicțiilor pentru un set de date nou, utilizând modelul de regresie antrenat anterior. Noul set de date conține următoarele date:

* sqft\_living = 2500,
* bedrooms = 4,
* bathrooms = 2.5,
* floors = 2,
* yr\_built = 2005.



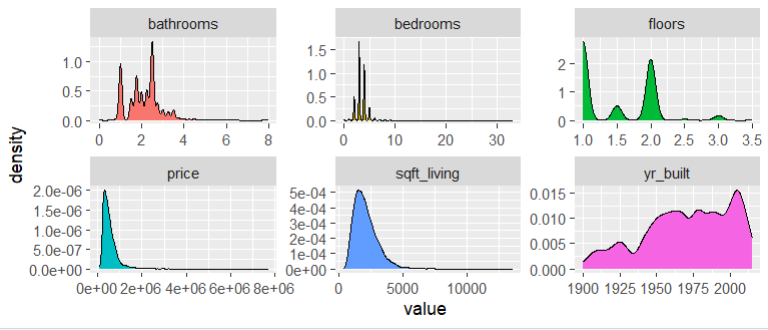
Intervalul de încredere se situează între limita inferioară(lwr) de 557500,3 și limita superioară (upr) de 568809,1. Se estimează pe baza acestui interval de încredere (de 95%), prețul mediu al locuinței pe acest set nou de date, va fi 563104,7.

Intervalul de predicție se situează între limita inferioară(lwr) de 563104,7 și limita superioară (upr) de 83065,09. Se estimează pe baza acestui interval de predicție, cu o încredere de de 95%, prețul unei locuințe individuale pe acest set nou de date, va fi 563104,7. Aici, intervalul este mai larg deoarece măsoară incertitudinea unei valori individuale asupra prețului.

Mai mult, valoarea RMSE pentru această regresie este 247177.7 sugerând cu cât diferă valorile reale de predicțiile făcute de model, aceasta fiind o valoare medie.

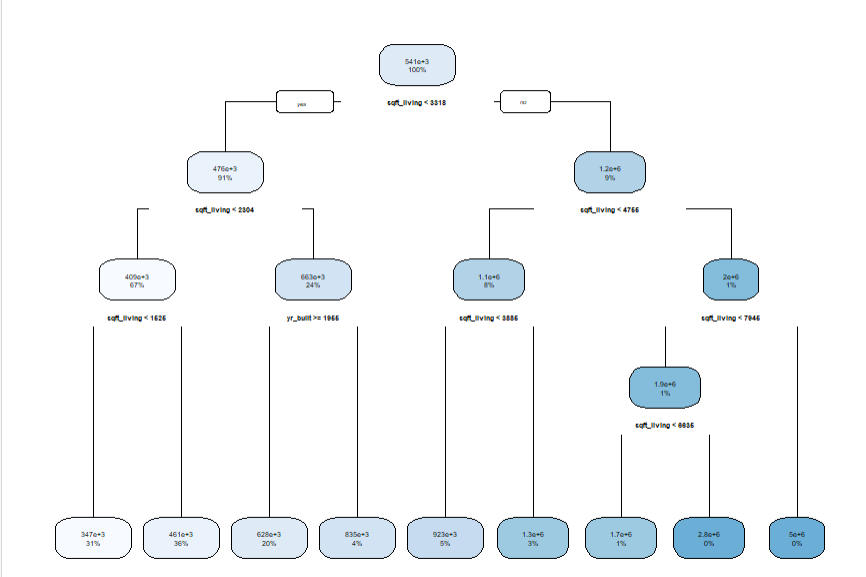
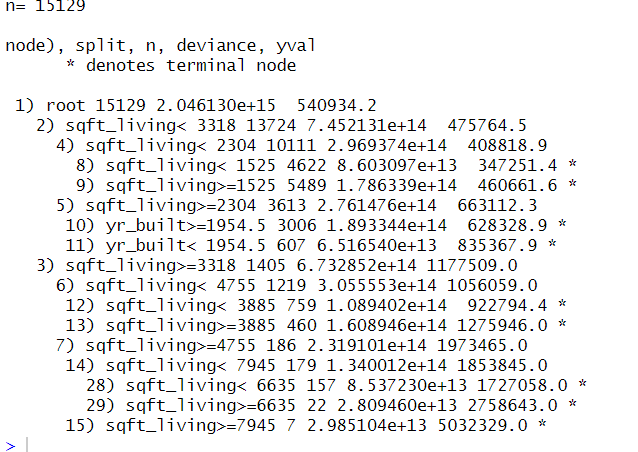
Al doilea tip de model ce urmează a fi prezentat este arborele de decizie.

Graficul densităților prezintă distribuția valorilor variabilelor numerice din setul de date, în datele de antrenament.

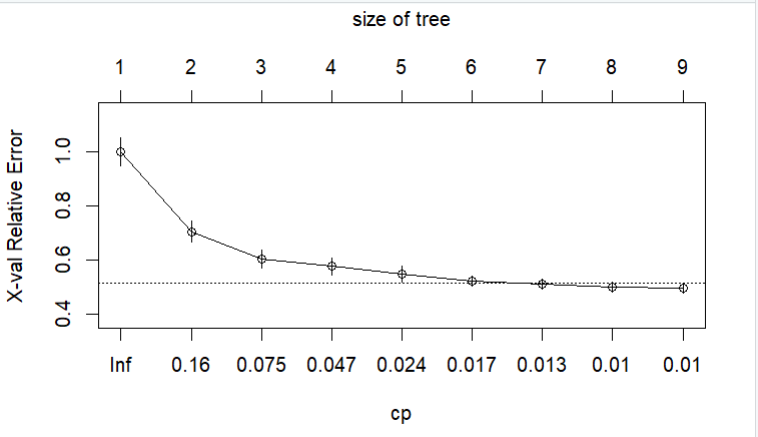


Se pot observa diferite aspecte, precum faptul că cele mai frecvente locuințe au una sau 2 băi. Totodată, numărul de dormitoare variază între 2 și 4. Majoritatea locuințelor au 1 sau 2 etaje, dar se observă și locuințe cu 3 etaje. Graficul distribuției prețurilor arată că majoritatea locuințelor au prețuri mai mici. De asemenea, putem vedea că cele mai multe construcții de locuințe au fost construite între 1950 și 2000.

Pentru început am împărțit setul de date in două, unul set pentru antrenament(70%) și unul pentru test (30%), apoi am creat arborele de antrenament.



Setul de date este împărțit în mai multe subseturi de dimensiuni mai mici. Avem nodul de început, nodul rădăcina= 15129 care conține instanțele. Urmează a se împărți setul de date în două, pe baza variabilei criteriului sqft\_living<3318 pentru minimizarea SSE(devierea). Nodurile care conțin predicția finală sunt prezente în primul grafic (text) având \* la final. Pentru evitarea overfitting-ului se folosește cp( parametru de complexitate), pentru graficul al doilea.



Putem observa din graficul de mai sus că cea mai mică valoare SSE (deviere) este atunci când α=0,1 și avem 9 noduri terminale.

Am creat apoi al doilea arbore cu cp=0, ceea ce înseamnă că toate împărțirile din setul de date sunt fară prunare, rezultând un arbore foarte complex pentru a vizualiza toate împărțirile posibile. Acesta prezintă cum în vreme ce complexitatea arborelui crește, eroarea relativă scade.

A graph of a tree

Description automatically generated

Locul unde se află cea mai mică eroare relativă este prezentat prin linia verticală punctată.

În continuare se caută cele mai bune valori pentru maxdepth si minsplit, de care depinde complexiateta arborelui nostru. Am folosit un hyper grid pentru a găsi cei mai buni parametrii, folosind maxdepth între 8 și 15 și minsplit între 5 și 20.

A screenshot of a computer code

Description automatically generated

Din imaginea prezentată vedem următoarele:

* dacă arborele are minim 5 observații pentru fiecare împărțire, va avea o adâncime de 8 și asa mai departe.

A number of numbers and letters

Description automatically generated with medium confidence

Se observă că arborele optim trebuie să aiba cel puțin 14 instanțe și o adâncime de 15 niveluri. Pentru prevenirea overfittingului, trebuie să avem un cp de 0,01. Se vede și că în acest caz, cea mai mică eroare relativă (cross-validare) este de 0,4910224.

De asemenea, am obținut valoarea RMSE-ului de 254806,5, indicând faptul că modelul are o precizie medie în ceea ce consta prețul locuințelor în funcție de suprafața locuinței, numărul de băi, numărul de camere, numărul de etaje și anul construcției. Așadar, modelul e unul relativ bun, dar trebuie să ținem cont că predicțiile nu sunt întotdeauna reale.

**Concluzie**

Acest proiect s-a concentrat asupra întrebării din introducere și anume, ,,*Cum influențează numărul dormitoarelor, suprafața în metri pătrați ( m2), numărul băilor, etajele și anul construcției imobilului, prețul de cumpărare? “.*

Am realizat cercetări pe baza variabilelor sqft\_living (suprafața locuibilă), bedrooms (numărul de dormitoare), bathrooms (numărul de băi), floors (numărul de etaje) și yr\_built (anul construcției imobilului). Am folosit un modele de regresie simplă, un model de regresie multiplă și un model de arbore de decizie. În urma realizării modelului de regresie multiplă, s-a constatat că factorii ce influențează cel mai mult prețul unei locuințe sunt suprafața locuibilă, numărul de băi și numărul de etaje. Având un R-squared de 55,51% înseamnă că modelul explică prețul locuințelor cu variabilele alese în proporție de 55,51%. Acesta are si un RMSE de 247177,7 aceasta fiind o diferența între prețurile prezise de model și cele reale. În cadrul arborelui de decizie, putem observa că cea mai importană variabilă e, în mod sigur, suprafața locuinței, iar RMSE-ul aflat e de 254806,5. Acesta indică o valoare mai puțin mai mare decât cea de la modelul de regresie, indicând un model mai slab.

În concluzie, modelul de regresie multiplă este mai eficient și are performanțe mai bune.