Housing

1. **Introducere**

Presupunem cazul unei companii imobiliare ce are la dispoziție setul de date luat spre analiză, ce conține prețurile proprietăților dintr-o anumită regiune. Aria de cercetare aleasă în vederea realizării acestui proiect este reprezentată deci, de piața imobiliară dintr-o anumită regiune.

Compania de imobiliare dorește să cunoască modul în care este afectat prețul imobilelor de factorii caracteristici precum: aria, numărul de dormitoare, numărul de băi, locuri de parcare disponibile, numărul de etaje, daca este sau nu situate imobilul la drumul principal, existenta subsolului etc.

Studiul acestor factori este relevant din două perspective. În primul rând, din prisma agenților imobiliari, deoarece aceștia doresc să cunoască factorii care contribuie la creșterea valorii unei proprietăți, în vederea unui profit cât mai mare. În al doilea rând, acest studiu este relevant pentru potențialii cumpărători de proprietăți, ce trebuie să aibă în vedere posibila evoluție a prețurilor în cazul în care doresc achiziționarea unei proprietăți, în funcție de caracteristicile definitorii ale viitoarei locuințe.

Domeniul de cercetare ales de noi a fost abordat de nenumărate ori, deoarece este un domeniu dinamic, aflat într-o continuă expansiune, însă pe diferite regiuni. Rezultatele studiilor similare au arătat o corelație între suprafață, zonă, confort etc. și prețul imobilelor, însă setul de date ales de noi are studiază și alte elemente.

Ne așteptăm să găsim o corelație între prețul imobilelor și suprafața acestora, însă ne dorim să descoperim și impactul altor aspecte prezente în setul de date ales de noi.

1. **Setul de date**

Setul de date ales de noi este Housing Dataset (<https://www.kaggle.com/ashydv/housing-dataset>)

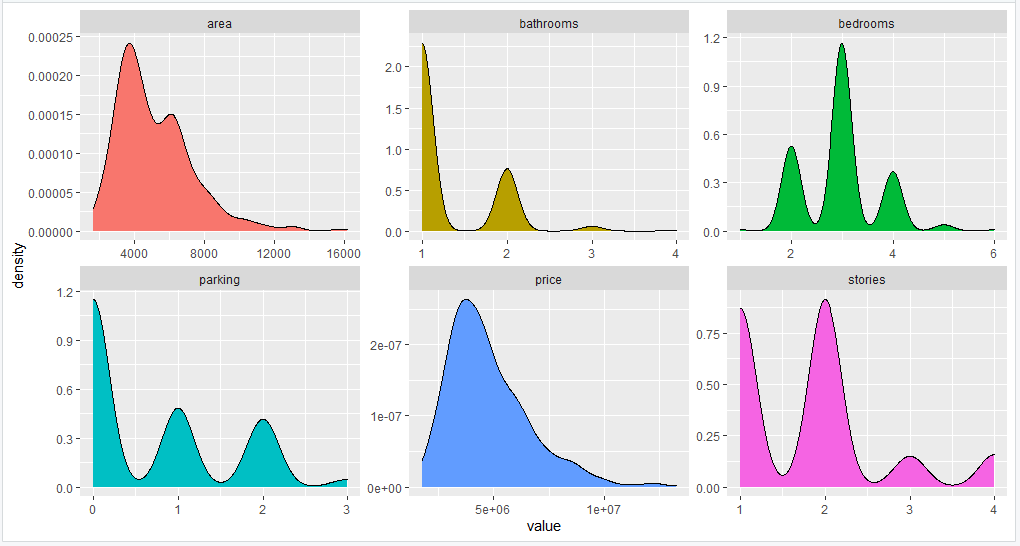
Acest set include atribute precum: preț, suprafață, număr de dormitoare, număr de băi, etaje, dacă imobilul este poziționat la drumul principal, dacă are cameră de oaspeți, dacă are subsol, etc.

Housing Dataset are 545 de instanțe a 13 variabile, atât nominale, cât și numerice.

Setul de date ales de noi este relevant pentru proiectul de cercetare deoarece prezintă atribute cheie care influențează valoarea prețului unui imobil.

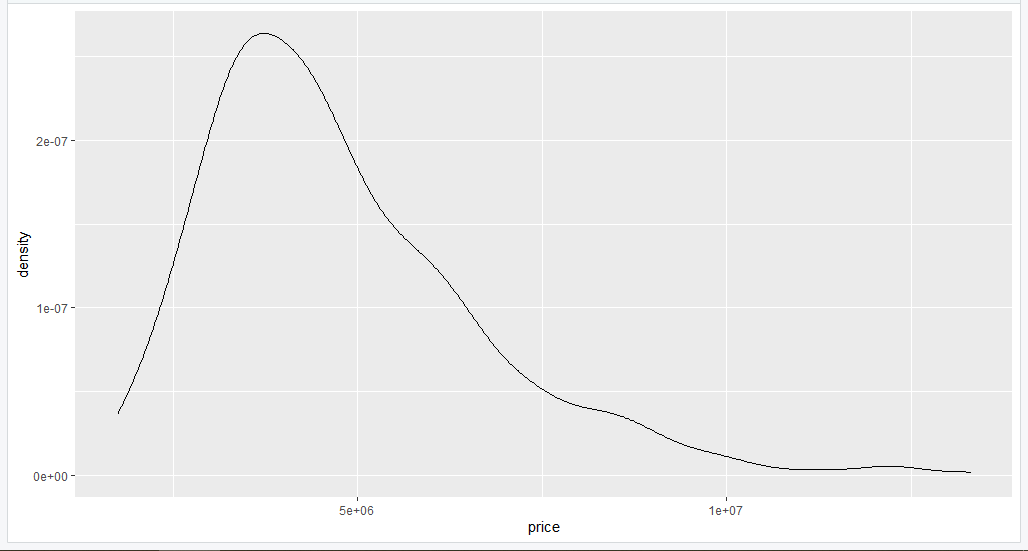
Datele din setul Housing Dataset nu au necesitat o curățare înainte de prelucrare întrucât toate valorile sunt corespunzătoare fiecărui atribut, neexistând excepții de forma N/A sau valori nule unde logica impune existența unor valori nenule.

Pentru început, vrem să vizualizăm atributele numerice într-o formă grafică:



După cum se poate observa și în graficul de mai sus, atributul suprafață (area) se concentrează în jurul valorii de 4000. Numărul de băi variază între valorile 1 și 4, majoritatea imobilelor dispunând de o singură baie, însă și numărul imobilelor care prezintă două băi nu este de neglijat. Numărul de dormitoare care predomină este 3 dormitoare per imobil, dar așa cum se observă și din grafic, o pondere ridicată o dețin acele imobile care prezintă doar 2 dormitoare, respectiv 4 dormitoare, numărul imobilelor cu 1, 5 și 6 dormitoare fiind redus. Cele mai multe imobile nu dispun de un loc de parcare, cu toate acestea numărul imobilelor care dispun de un singur loc de parcare este relativ egal cu numărul imobilelor care dispun de două, în timp ce foarte puține imobile au 3 locuri de parcare. Prețul imobilelor variază între 1.750.000 u.m. și 13.300.000 u.m. și se concentrează în jurul valorii de 3.750.000 u.m. Dacă ne raportăm la numărul de etaje, se poate observa faptul că majoritatea imobilelor prezintă 1 sau 2 etaje cu o pondere relativ egală, sau pe de altă parte 3 sau 4 etaje, tot cu o pondere relativ egală, cu toate acestea, imobilele dispuse pe suprafața a două etaje primează, fiind urmate de cele dispuse pe un singur etaj.

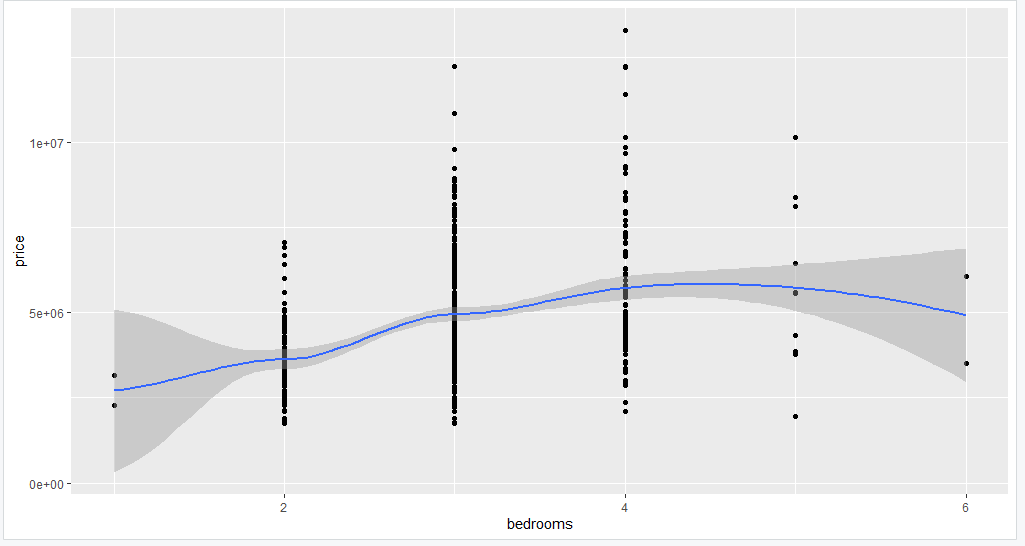
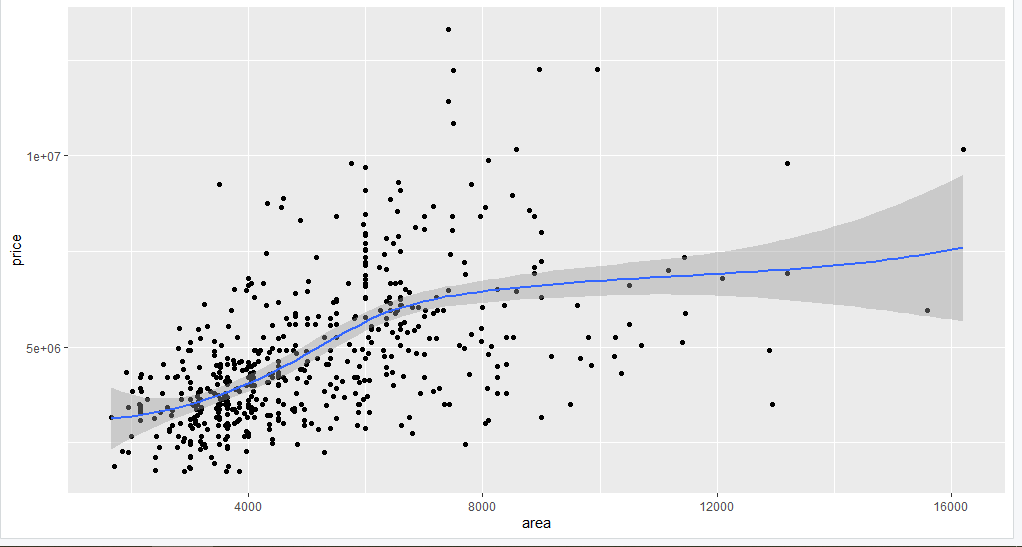
Deoarece ne raportăm la preț ca fiind variabila dependentă, dorim o vizualizare mai amănunțită asupra acestuia:



Cu ajutorul graficului de mai sus, putem vizualiza prețurile imobilelor. Calculăm o unitate de măsură raportată la grafic, aceasta fiind egală cu 2.500.000, deci se observă că majoritatea prețurilor au valorile în intervalul 2.500.000 u.m și 5.000.000 u.m.

1. **Rezultate și discuții**

Așa cum am descris în partea de introducere, ne așteptăm ca prețul să se afle într-o legătură liniară cu suprafața sau/și numărul de dormitoare. Pentru a valida această premise, am folosit grafice pentru a vedea dependența prețului de cele două attribute menționate. Rezultatele sunt următoarele:



În urma vizualizării graficelor, se poate concluziona faptul că percepția noastră cu privire la dependența prețului de suprafață și/sau numărul de dormitoare a fost eronată deoarece legătura nu este una liniară. Din acest considerent, am apelat la utilizarea arborilor de decizie, care pot fi utilizați atunci când suprafața de regresie nu este netedă[[1]](#footnote-1).

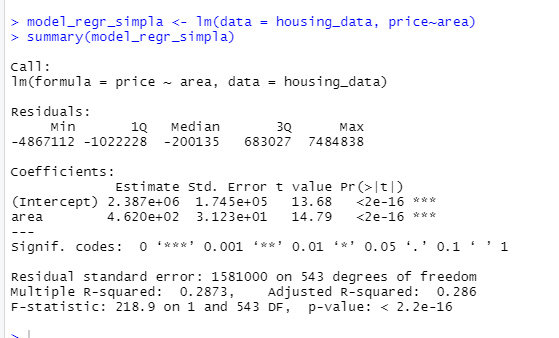
Chiar dacă nu s-a concretizat o legătură liniară între variabila dependentă (preț) și variabilele independente studiate, vom cerceta în cele ce urmează legătura dintre preț și suprafață.

**Regresia liniară simplă**

Pentru a decide care sunt variabilele importante am ales să realizăm o secție mixtă, adică o combinație între selecția forward (înainte), care presupune studierea fiecărei variabile independente separat, și selecția backward (înapoi) unde se începe cu o regresie multiplă ce conține toți predictorii, urmând să fie excluși cei care nu sunt importanți.

**model\_regr\_simpla <- lm(data = housing\_data, price~area)**

**summary(model\_regr\_simpla)**



Conform rezultatelor afișate, dacă presupunem că suprafața este 0, atunci prețul ar avea valoarea afișată a coeficientului ꞵ0 (Intercept), caz imposibil deoarece nu există logică în această afirmație: nu se poate propune spre vânzare o proprietate a cărei suprafață este 0.

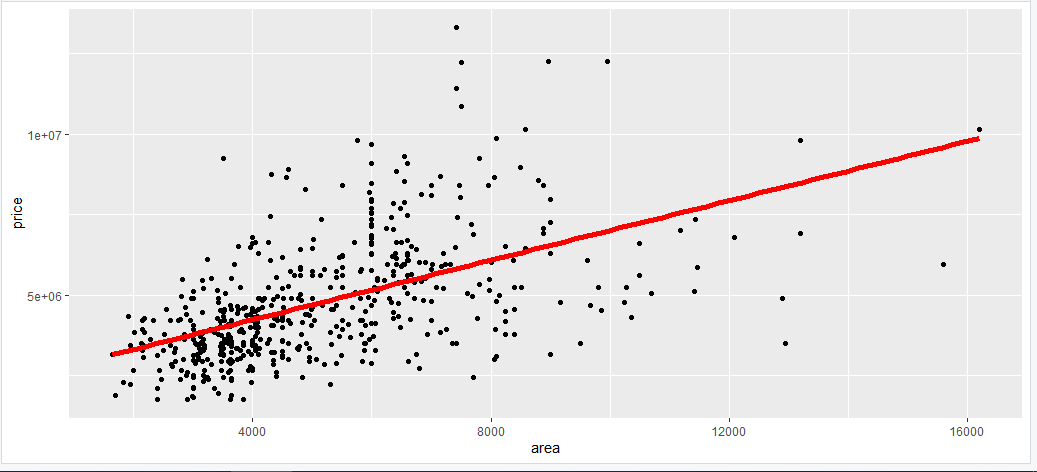
Pentru suprafață, dacă aceasta va crește cu o unitate, atunci prețul va crește cu 4.620e+02 (ꞵ1).

Standard error ne spune cât estimarea parametrului diferă de valoarea reală, în medie.

P-value reprezintă probabilitatea să facem o asociere între predictori și variabila dependentă datorită șansei. O valoare mică ne permită să tragem concluzia că există o asociere între predictor (suprafața) și variabila dependentă(prețul)[[2]](#footnote-2). Concluzionăm că există o relație, adică suprafața este semnificativă pentru determinarea prețului, deci respingem ipoteza nulă.

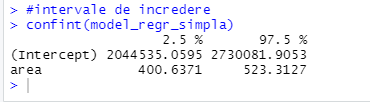
Residual standard error este mărimea medie cu care variabila dependentă va devia de la linia de regresie, R-squared ne indică proporția din variabilitatea prețului care poate fi explicată pe baza suprafeței (în cazul de față 28,73%)[[3]](#footnote-3), adică putem explica prețul în proporție de 28,73% pe baza suprafeței.

În continuare vom genera linia de regresie pentru predicția relației dintre variabila dependentă (preț) și variabila independentă (suprafață)



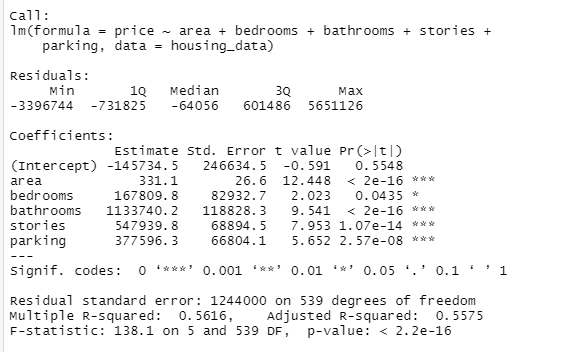
Se observă faptul că există distanță între punctele de pe grafic și linia de regresie. Această distanță este RSE – Eroarea standard reziduală menționată mai sus, valoare care este foarte mare, rezultând că potrivirea nu este una bună pe date.

Intervale de încredere pentru suprafață în modelul regresiei liniare simple:



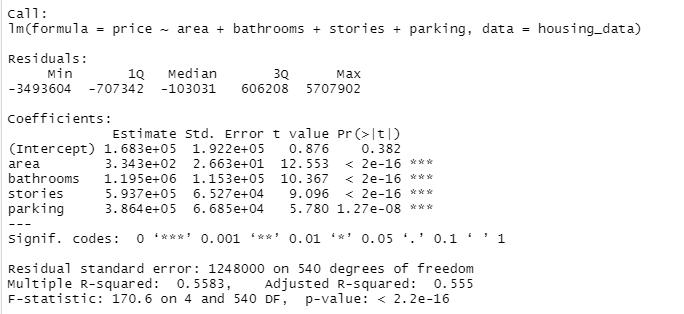
Cu ajutorul acestui model am văzut cum evoluează prețurile în funcție de suprafață.

**Regresia liniară multiplă**

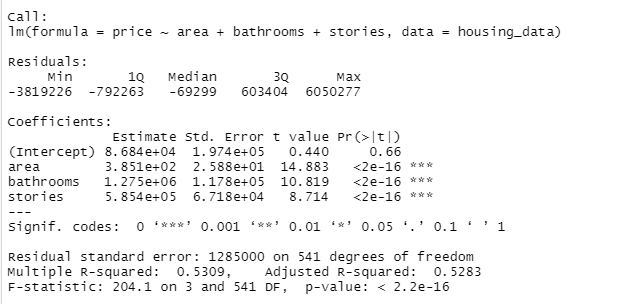
****

Regresia lineară multiplă ne ajută să vedem cât de puternică este influența variabilelor independente asupra variabilei dependente. De exemplu, pentru un nivel fixat al suprafeței, numărului de dormitoare, numărului de etaje și numărului de parcări, creșterea cu o unitate a numărului de băi, determină o creștere a prețului cu 1133740,2 u.m.

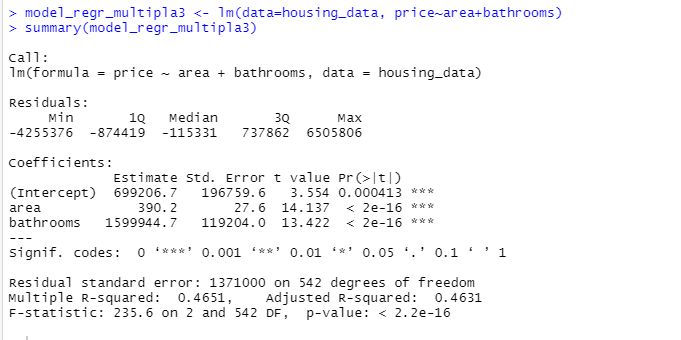
Observăm că doar anumiți factori sunt importanți pentru determinarea prețului, iar în consecință vom lua doar acești factori spre studiu, în diferite combinații:



\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

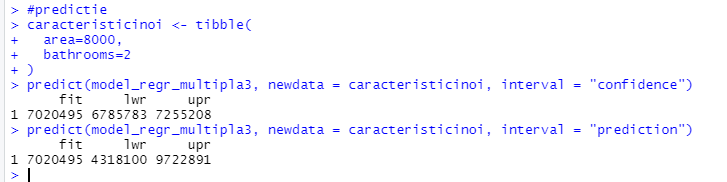


\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_



În urma acestui studiu, se poate observa că F-statistic are valoarea cea mai depărtată de 1 în cazul în care luăm spre analiză suprafața și numărul de băi. Așadar, acesta este cazul cel mai satisfăcător, însă se explică doar 46,51% din variabilitatea prețului pe baza acestor doi factori.

**Predicție**



Din imagine, se poate observa valoarea prezisă a prețului (7.020.495) în funcție de valorile alese de noi pentru suprafață și numărul de băi. În intervalul de încredere, această valoare se situează între 6.785.783 și 7.255.208, iar în intervalul de predicție, această valoare este situată între 4.318.100 și 9.722.891. Intervalul de predicție este mai larg deoarece pentru intervalul de încredere nu luăm în considerare Ԑ.

Pentru intervalul de încredere, dacă suprafața unui imobil este 8000 și numărul de băi este egal cu 2, valoarea medie a prețului, va fi situată între 6.785.783 și 7.255.208, cu o încredere de 95%.

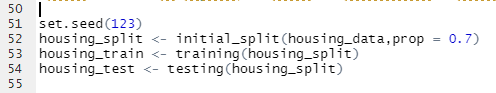
Pentru intervalul de predicție, dacă suprafața unui imobil este 8000 și numărul de băi este egal cu 2, valoarea prețului, va fi situată între 4.318.100 și 9.722.891, cu o încredere de 95%.

**O altă abordare**

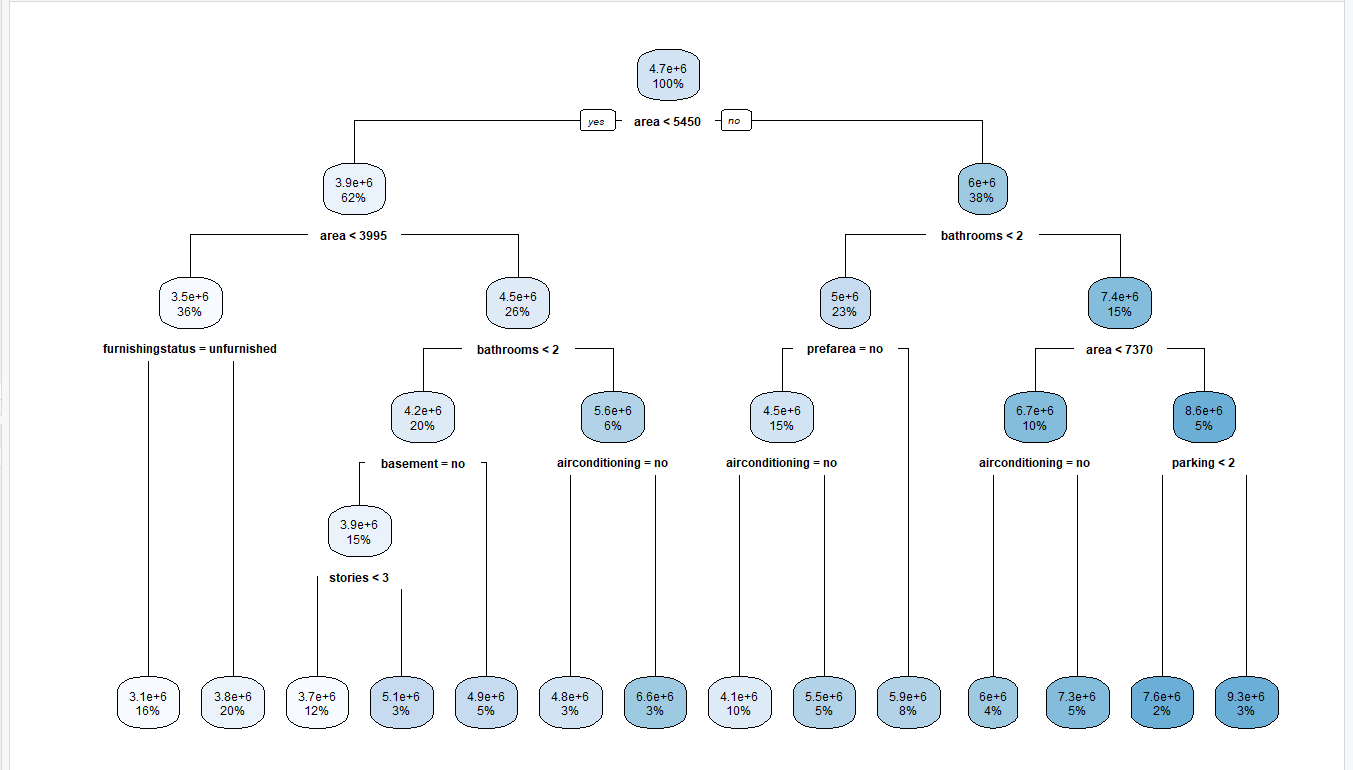
Având în vedere interesele unei agenții imobiliare, arborii de decizie se pliază pe aria de interes, aceștia fiind populari în mediul de afaceri datorită ușurinței de a fi interpretați și explicați și de către persoane fără cunoștințe de analytics[[4]](#footnote-4).

În ciuda acestor avantaje, trebuie să menționăm și faptul că arborii de decizie sunt mai puțin competitivi decât alte metode și pot fi lipsiți de robustețe, ca dezavantaje principale ale acestei metode. Cu toate acestea, este posibil ca acuratețea predicției să fie îmbunătățită prin combinarea unui număr mare de arbori.

Pentru început vom dezvolta abordarea CART. Aceste presupune împărțirea setului de date în două părți: setul de test și setul de antrenament, cu proporțiile 30%, respectiv 70%. Pentru a stabiliza rezultatele vom folosi funcția set.seed și îi vom fixa parametrul 123. Nu are importanță valoarea parametrului, acesta este relevant pentru a asigura reproductibilitatea modelului în momentul unei utilizări ulterioare.

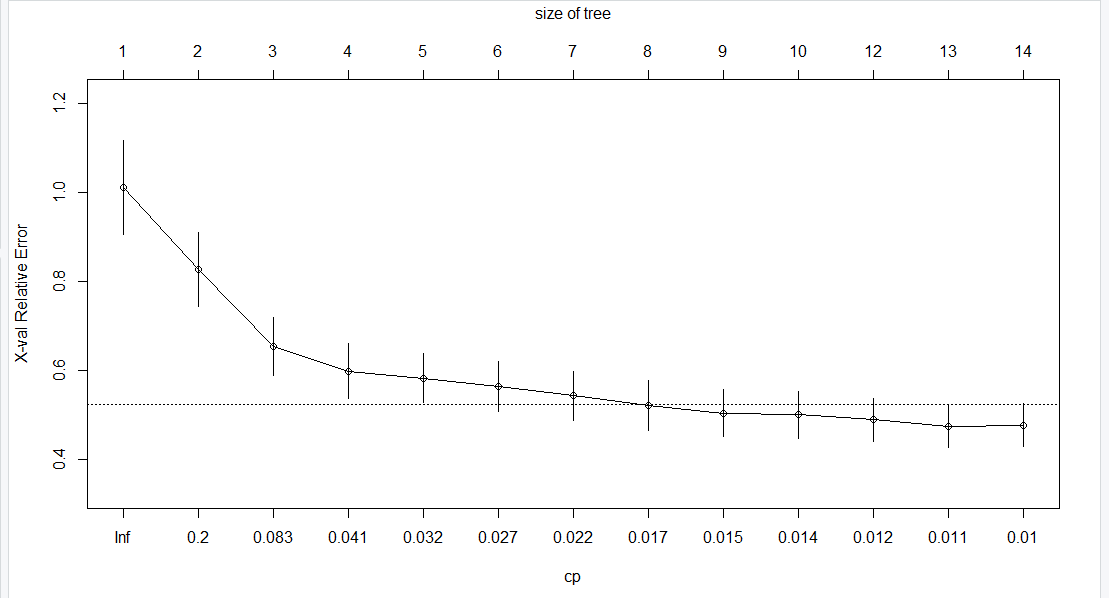


Crearea primul model s-a realizat utilizând valorile implicite (minsplit=20, maxdepth=30), oferind doar setul de date ca parametru și specificând metoda ca fiind anova. Astfel rezultatul afișat este următorul:



Se poate observa că în vârful arborelui, atributul care împarte setul nostru de date în două părți este suprafața (area). Împărțirea se realizează pe două ramuri din punct de vedere al suprafeței, prima ramură conține valorile mai mici decât 5450, iar a doua ramură valorile mai mari. Apoi aceste valori, la rândul lor, se împart din nou, astfel încât în ramura din partea stângă, împărțirea se va face tot prin suprafață, în imobile cu o suprafață mai mică decât 3995 și mai mare decât această valoare, iar în ramura din partea dreaptă, unde avem imobilele cu suprafața mai mica decât 5450, împărțirea se va face luând în considerare numărul de băi. Analog, se coboară în arbore tot mai adânc, specificându-se atributul cu ajutorul căruia se va lua decizia de împărțire, pănă la nodurile frunze. Pe fiecare nod apare prețul ca o medie pentru fiecare împărțire a atributelor în funcție de atributul după care s-a făcut decizia, precum și procentul pe fiecare nivel ierarhic.

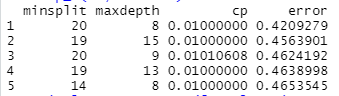
În continuare dorim să minimizăm SSE + α | T | , unde SSE reprezintă suma erorilor pătrate, α reprezintă costul de complexitate și | T | reprezintă numărul de noduri terminale în arbore.



Testăm mai multe valori alpha, pentru fiecare scade de la o valoare mare la o valoare cât mai mică, realizându-se câte un arbore și încercând să se scadă valoarea SSE-ului. De la un anumit punct valoarea SSE-ul nu mai scade, în cazul de față valoarea oprindu-se la 0.01.

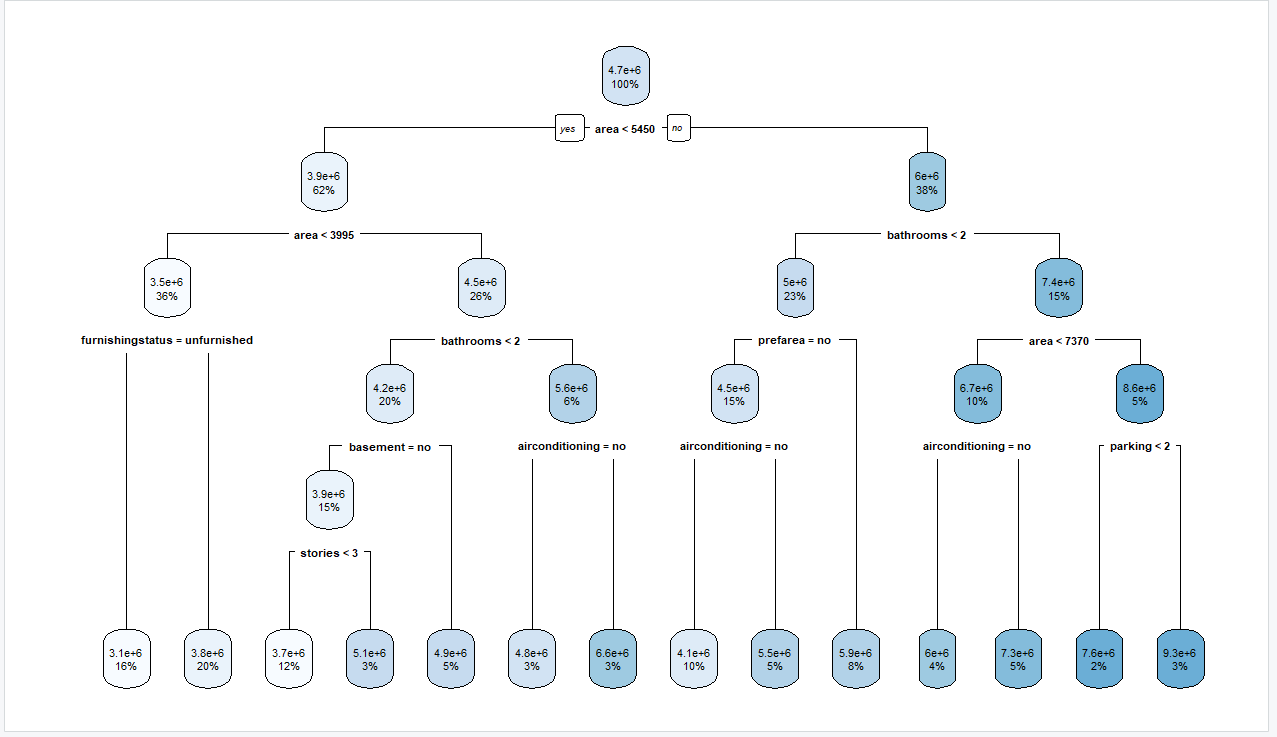
După cum se observă în grafic, în punctul în care alpha ia valoarea 0.017 și mărimea arborelui este 8, SSE + α | T | este minimă (vezi linia trasată). Prin urmare, 0.017 este valoarea alpha aleasă pentru care s-a minimalizat formula precedentă.

În continuare, căutăm arborele optim cu valorile pentru minsplit și maxdepth cele mai bune. În acest sens, minsplit va lua valori între 5 și 20, iar maxdepth între 8 și 15. Astfel se va creea un arbore de decizie pentru fiecare combinație dintre minsplit și maxdepth, rezultând 128 de combinații. Creem o listă de modele pentru fiecare combinație dintre minsplit și maxdepth, apoi calculăm cp și eroarea, urmând să adaugăm aceste valori în tabelul anterior pentru a vedea pentru fiecare cobinație acești parametrii. Pentru a obține valorile optime, vom vizualiza primele 5 înregistrări din tabel, ordonate după eroare, descrescător. Astfel opținem următoarea combinație optima: minsplit=20, maxdepth=8, cp 0.01, error=0.4209279.



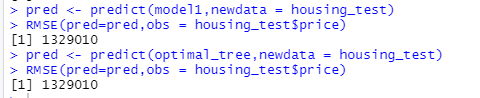
După ce am aflat combinația optima, vom construi arborele optim cu aceste valori (vezi pagina următoare).

Așa cum se poate observa din arborele optim, acesta este similar cu primul arbore unde minsplit și maxdepth au avut valori standard. Similaritatea se explică prin faptul că minsplit optim este 20, așa cum este și standard, iar maxdepth optim este 8, la fel cum am văzut anterior în cazul minimizării SSE + α | T | pe modelul creat cu valori implicite, mărimea arborelui fiind 8.



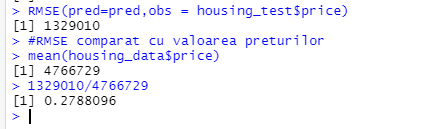
**Predicții**

Pentru modelul 1, eroarea medie RMSE = 1.329.010, implicit și pentru arborele optim, datorită faptului că arborii sunt similari, așa cum am demonstrat anterior.



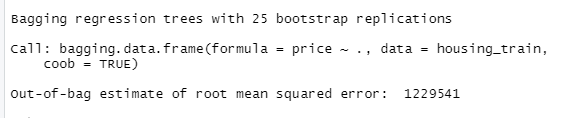
Observam valoarea RMSE destul de mare, însă aceasta trebuie comparată cu media preturilor.

În imaginea următoare se poate observa raportul dintre RMSE și media prețurilor:



BAGGING

Se creează modelul

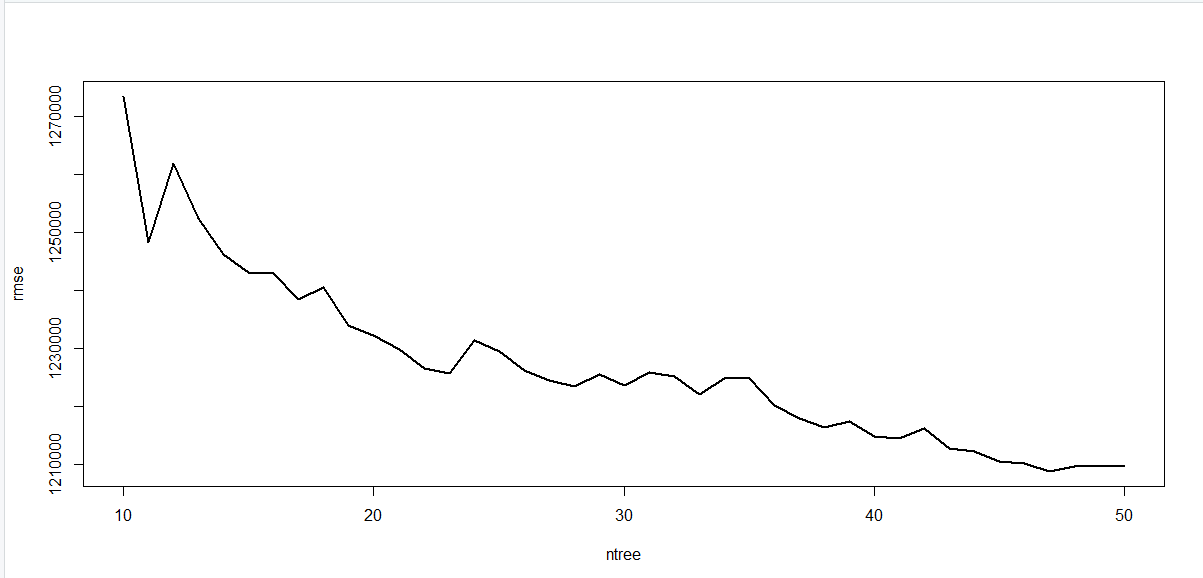


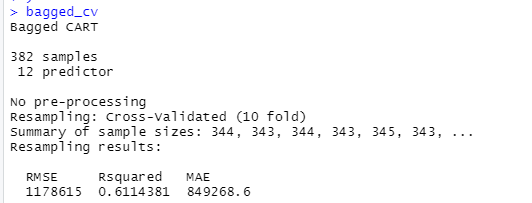
Comparativ cu RMSE precedent, valoarea s-a îmbunătățit (1.229.541)

Este o abordare pesimistă deoarece rămân multe instanțe neluate în considerare.

Vrem să modificăm valoarea standard pentru bagging(25), luând valori de la 10 până la 50, să vedem dacă se îmbunătățește RMSE-ul.

Observăm că pe măsură ce numărul de bag-uri crește, scade RMSE-ul

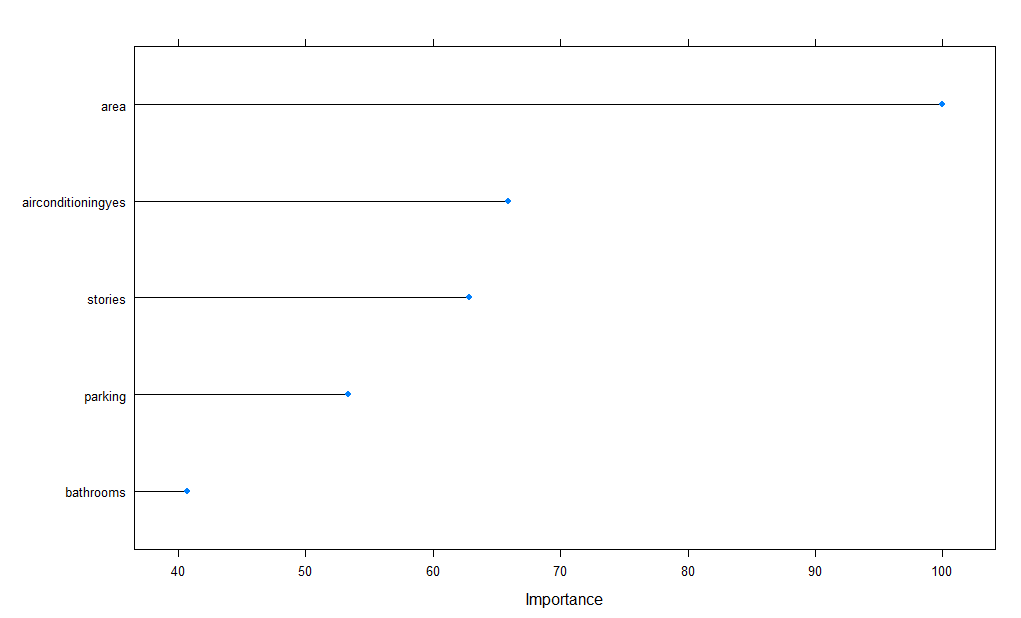




Se observă că RMSE-ul s-a îmbunătățit,

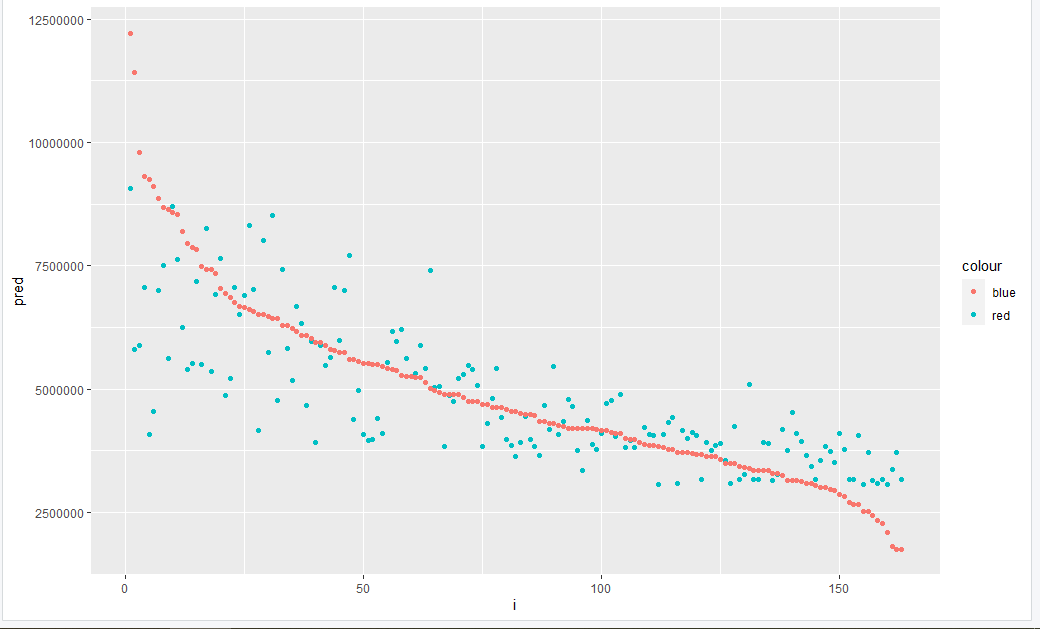
Atributele cele mai importante sunt agregate pe toți arborii construiți. Astfel, prin bagging, nu avem doar un singur arbore, iar ca observație se poate vedea că în toți arborii atributul area este primul. Faptul că un imobil are aer condiționat reduce eroarea cu aproximativ 67%.

Ordinea de importanță a atributelor:

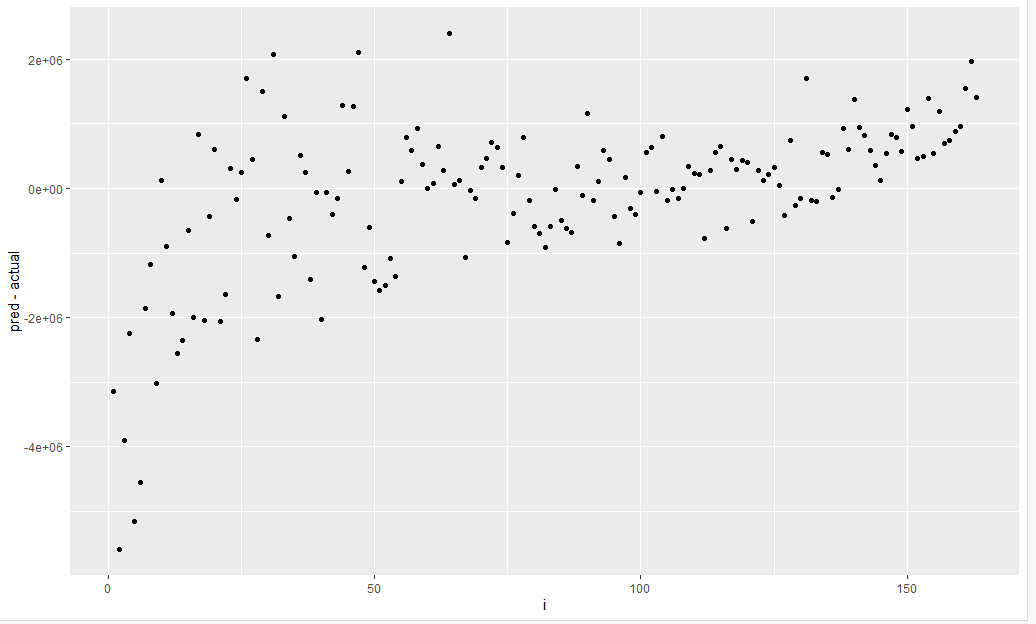


Predicție: RMSE= 1253765

Predicția este reprezentată cu roșu, iar valorile reale cu albastru.



Dorim să vizualizăm diferența dintre predicție și valoarea reală. Valorile se situează în jurul valorii 0.



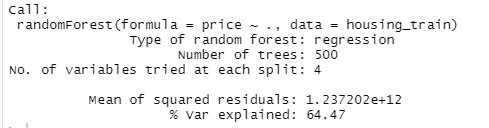
**RANDOM FOREST**

Anterior, am luat primele 5 atribute, însă acum apare întrebarea că dacă există și alte atribute pe care nu le-am luat în considerare, dar care totuși sunt importante.

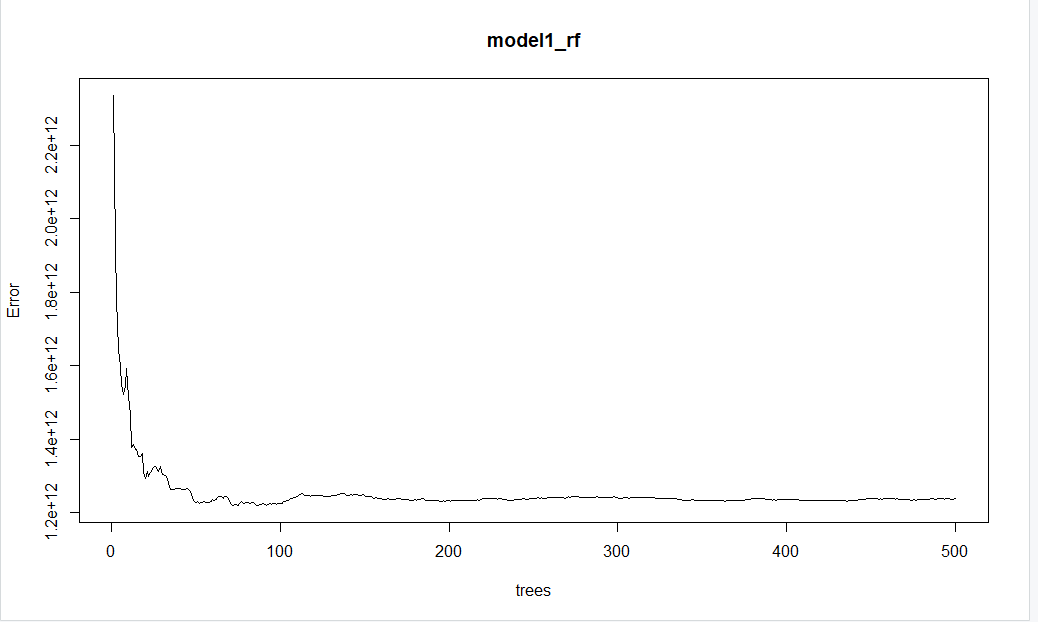
Astfel menționăm:

m - atribute extrase aleatoriu, pe care se vor face teste (vedem dacă și alte atribute devin importante), care este egal cu radical din numărul de atribute (aici 13), adica m este 3.

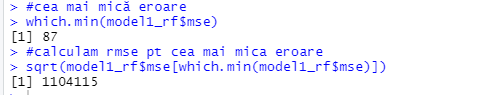
Observăm că: modelul este pentru regresie, numărul de arbori este 500, numărul de variabile încercate la fiecare împărțire este 4. Acest model explică 64.47% din informații. Media rezidurilor obținute este foarte mare.



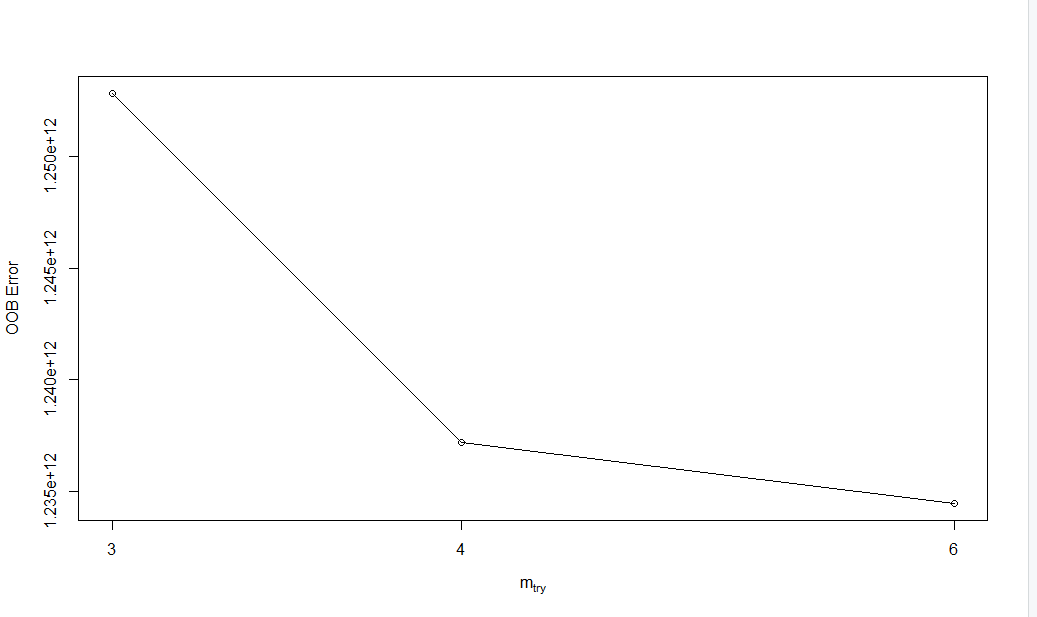
Se observă că eroarea scade pe măsură ce se construiesc arborii.

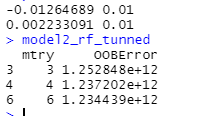


Observăm că cea mai mica eroare se află pe poziția 87, iar RMSE-ul aferent este 1.104.115



**Random forest tunning**





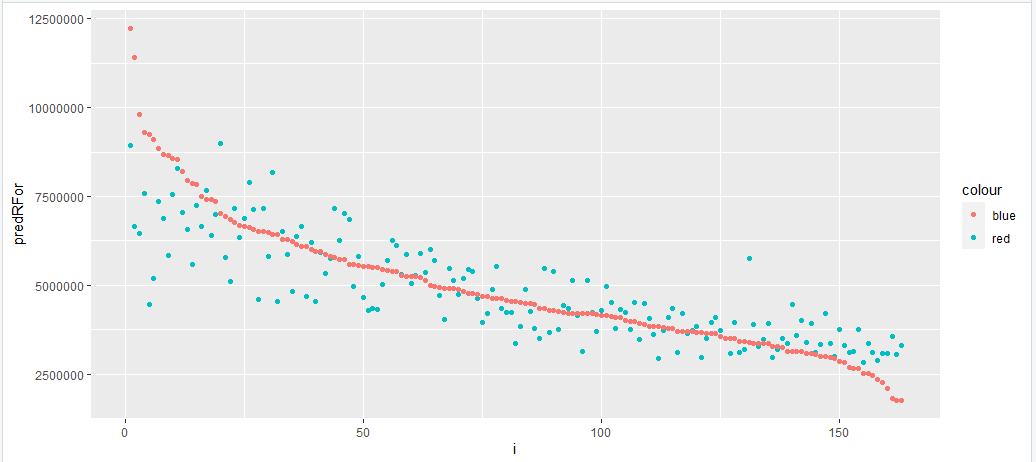
Se poate observa că eroarea cea mai mica este la 6 pentru mtry

RMSE-ul rezultat este 1.111.053



Modelul nu este mai bun deoarece RMSE-ul este mai mare decât RMSE-ul pentru model1\_rf.

Predicția pentru modelul găsit ca fiind cel mai bun, adică modelul unde RMSE-ul este cel mai mic este reprezentată în imaginea următoare:



Menționăm că i reprezintă observațiile din setul nostru de date.

1. **Concluzia**

Luând în considerare întrebările de la care am pornit acest studiu, și anume cum este afectat prețul de atributele existe în setul de date, am ajuns la concluzia că, cel mai bun model de analiză este modelul Random Forest, deoarece în acest caz am obținut cea mai mică valoare pentru RMSE, adică 1.104.115.

În urma acestui model, s-a realizat predicția pentru atributele setului de date. Graficul pe baza căruia se pot face interpretări datorită evoluției se poate vizualiza în imaginea de mai sus, cu mențiunea că predicția este reprezentată cu roșu, iar valorile reale cu albastru.

1. Curs 4 – Arbori de decizie – GC Silaghi (UBB) [↑](#footnote-ref-1)
2. Curs 1 – Regresia – GC Silaghi (UBB) [↑](#footnote-ref-2)
3. Curs 1 – Regresia – GC Silaghi (UBB) [↑](#footnote-ref-3)
4. Curs 4 – Arbori de decizie – GC Silaghi (UBB) [↑](#footnote-ref-4)