PROIECT-

* FUNDAMENTE DE BIG DATA -

**Hotel Booking Demand**

Băghiuț Ioana-Roxana, Gr.2

Cuprins

[**1.** **Introducere** 3](#_Toc71734748)

[**2.** **Setul de date** 3](#_Toc71734749)

[**3.** **Rezultate și discuții** 6](#_Toc71734750)

[**4.** **Concluzii** 13](#_Toc71734751)

1. **Introducere**

În cadrul acestui proiect am analizat un set de date ce ține de industria ospitalității. Astfel, rezultate acestui studiu sunt utile persoanelor care lucrează în această zonă, mai exact celor care care lucrează în domeniul hotelier, sau persoanelor care doresc să înțeleagă care sunt factorii care influnețează oaspeții hotelurilor să renunțe la rezervările lor, dar și cum se poate prezice anularea acestora.

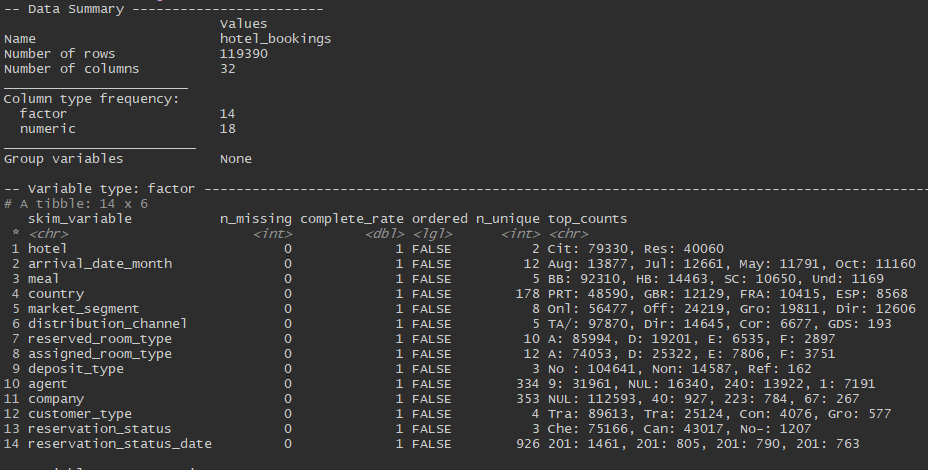
Prin acest proiect am dorit să descopăr cât de bine se poate prezice posibilitatea anulării rezervărilor, dar și ce anume influențează anularea lor. Alegerea acestor task-uri este relevantă, întrucât rezolvarea lor îi poate ajuta pe utilizatorii informațiilor (managerii celor două hotelurilor) să înțeleagă mai bine probleme legate de clienți sau rezervări.

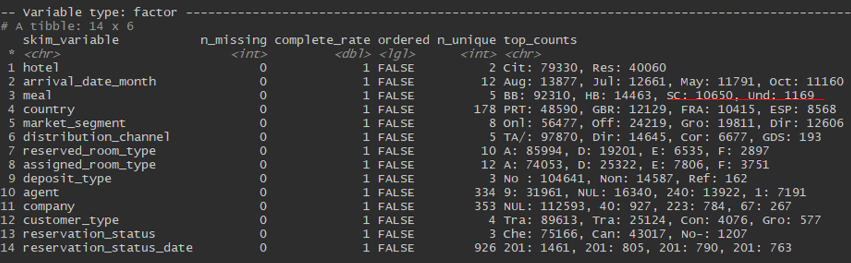
1. **Setul de date**

Setul de date utilizat în cadrul proiectului (*hotel\_bookings.cvs*) conține informații cu privire la rezervările a două hoteluri din categorii diferite (*resort hotel* și *city hotel*) pe perioada a aproximativ 4 ani. Amândouă hoteluri sunt situate în Portugalia.

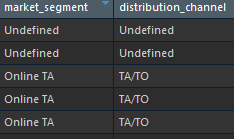
Hotel bookings conține 32 de coloane:

|  |  |  |
| --- | --- | --- |
| Atribut | Descriere | Tip |
|  |  |  |
| adr | = average daily rate  arată venitul mediu care se realizează pe cameră | numeric |
| adults | indică numărul de adulți | integer |
| agent | indică id-ul agenției de turism  obs: poate fi null, indicând astfel faptul ca rezervarea nu s-a realizat cu o agenție | categoric |
| arrival\_date\_day\_of\_month | indică ziua lunii în care au sosit oaspeții | integer |
| arrival\_date\_month | indică luna de cazare  obs: interval ianuarie-decembrie | categoric |
| arrival\_date\_week\_number | indică numărul săptămânii | integer |
| arrival\_date\_year | indică anul cazării | integer |
| assigned\_room\_type | indică tipul camerei rezevate | categoric |
| babies | indică numărul de bebeluși | integer |
| booking\_changes | numărul de schimbări asupra rezervării | integer |
| children | numărul de copii | integer |
| company | id-ul companiei care a făcut rezervarea/ a plătit pentru aceasta | categoric |
| country | țata de origine a oaspeților | categoric |
| customer\_type | tipul de rezervare  obs:  *contract* – când rezevarea are un contract asociat  *group* – rezervarea a fost făcută pentru un grup  *transient* – nu face parte dintr-un grup sau contract  *transient-party* – când rezervarea este de tip transient, dar e asociată cu o altă rezervare de tip transient | categoric |
| days\_in\_waiting\_list | numărul de zile din lista de așteptare până la confirmarea rezervării | integer |
| deposit\_type | indică dacă clientul a făcut un depozit pentru garantarea rezervării  obs:  *no deposit*  *non refund*  *refundable* | categoric |
| distribution\_channel | pe ce canal s-a făcut rezervarea  obs:  *corporate*  *direct*  *GSD*  *TA/TO*  *undefined* | categoric |
| hotel | tipul de hotel  obs:  *resort*  *city* | categoric |
| is\_canceled | dacă rezervarea a fost sau nu anulată (0 sau 1) | categoric |
| is\_repeated\_guest | dacă rezervarea a fost făcută de către un oaspete care a mai fost acolo (0 sau 1) | categoric |
| lead\_time | cu câte zile înainte s-a făcut rezervarea | integer |
| market\_segment | *aviation*  *complementary*  *corporate*  *direct*  *groups*  *offline TA/TO*  *undefined* | categoric |
| meal | *undefined/sc*  *bb – bed and breakfast*  *hb – half board*  *fb – full board* | categoric |
| previous\_bookings\_not\_cancelled | numărul de rezervări făcute până atunci care nu au fost anulate | integer |
| previous\_cancellations | numărul de rezervări făcute până atunci care au fost anulate | integer |
| required\_car\_parking\_spaces | numărul de locuri de parcare rezervate de client | integer |
| reservation\_status | *canceled*  *check-out*  *no-show* | categoric |
| reservation\_status\_date | data la care statusul rezervării a fost setat | date |
| reserved\_room\_type | codul tipului de cameră rezervată | categoric |
| stays\_in\_week\_nights | numărul de seri din timpul săptămânii din totalul de seri rezervate de client | integer |
| stays\_in\_weekend\_nights | numărul de seri de la sfârșitul săptămânii din totalul de seri rezervate de client | integer |
| total\_of\_special\_requests | numărul de cereri speciale făcute de către client | integer |

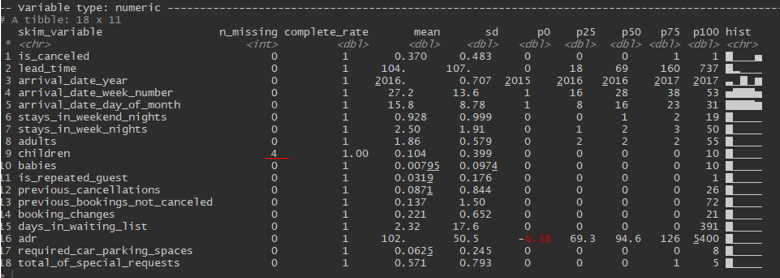
Întrucât setul de date are destul de multe coloane, pentru a avea o imagine de ansamblu asupra datelor am utilizat funcția **skim()**. Astfel, există 2 categorii de coloane în setul de date: numerice (18) șide tip factor (14).

În cadrul coloanelor de tip factor (din imaginea de mai jos) se observă faptul că nu sunt date lipsă. De asemenea, pentru fiecare coloană, se dă exact numărul de valori apărute. În cazul coloanei *meal*, aceasta ia valori și pentru *SC* și pentru *Undefined* care, în esență, înseamnă același lucru. Prin urmare, acestea ar putea fi combinate.

La o privire mai amănunțită asupra coloanelor *market\_segment* și *distribution\_channel* se observă faptul că există valori nedefinite, care vor putea fi șterse.



În cadrul coloanelor numerice există 4 valori lipsă pentru atributul children. Tot aici apar și valoarea medie, deviația standard, valorile quantile, alături de histograma fiecărei variabile. În cadrul histogramelor, se observă faptul că o mare parte dintre acestea sunt asimetrice, având o distribuție înclinată spre stânga, cu o coadă destul de lungă.

Astfel, în vederea curățării datelor am decis:

* Eliminarea înregistrărilor pentru care *market\_segment* sau *distribution\_channel* ia valori nedefinite;
* Adunarea coloanelor *children* și *babies* astfel încât aceasta să conțănă numărul total de copii pentru fiecare rezervare;
* Schimbarea coloanei *is­\_canceled* în „Yes” sau „No” în funcție de valorile de 0 sau 1 pe care aceasta le ia pentru fiecare înregistare;
* În cadrul coloanei *meal*, aceasta va lua valoarea *SC* în cazul în care aceasta are valoarea *Undefined*;
* Atributul *company* va lua valoarea „No” dacă aceasta este nulă, sau „Yes” dacă nu este, semnificând faptul că rezervarea respectivă a fost făcută printr-o companie sau nu;
* Adăugarea unei coloane *total\_nights\_stays* ca sumă între *stays\_in weekend\_nights* și *stays\_in\_week\_nights*, semnificând totalul de nopți rezervate;
* Adăugarea unei coloane noi *stay\_total\_cost* ca produs între coloana de mai sus, *total\_nights\_stays* și *adr*, semnificând costul total al rezervării;
* Eliminarea coloanelor *babies*, *stays\_in\_weeken\_nights*, *stays\_in\_week\_nights*, *reservation\_status\_date*, *reservation\_status*, *arrival\_date\_day\_of\_month*, *arrival\_date\_year*, *arrival\_date\_week\_number*, *assigned\_room\_type*, *country*, *total\_of\_special\_requests*, *booking\_changes*, *required\_car\_parking\_spaces*;

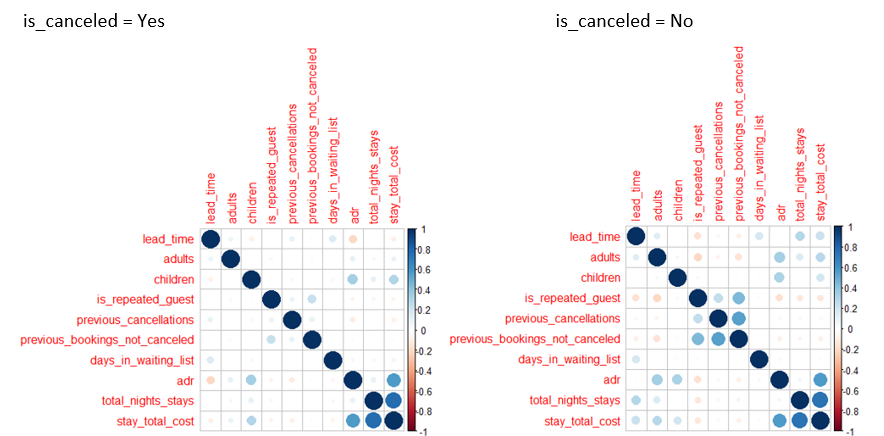
La final rămân 21 de coloane.

1. **Rezultate și discuții**

Pentru început am realizat matricile de corelație atât pentru rezervările neanulate, cât și pentru cele anulate (prezentate mai jos) pentru a vedea dacă atributele sunt independente între ele. Se observă astfel că atributele sunt mai corelate în cazul rezervărilor neanulate.

Există anumite corelații pozitive între :

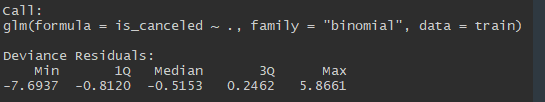
* *children* și *adr* - semnificând faptul că rezervările care au în componență copii sunt mai scumpe, afectând totodată și costul total al rezervării;
* *previous\_bookings\_not\_cancelled* și *is\_repeated\_guest –* arătând că oaspeții care nu și-au anulat rezervările sunt de obicei oaspeți care s-au mai cazat și în trecut;
* *adr* și *stay\_total\_cost –* relația este evidentă, cea de-a doua este calculată pe baza primeia, prin urmare costul total crește pe baza creșterii tarifului zilnic;
* *stay\_total\_cost* și *total\_nights\_stays* *–* la fel ca și mai sus, prima este calculată pe baza primeia, costul total al rezervării crește pe baza creșterii numărului total de nopți cazate;

În vederea realizării modelelor de clasificare am ales utilizarea: regresiei logistice, Naive Bayes, arborii de clasificare (atât cei prin abordarea CART, cât și utilizând entropia) și arborii avansați de clasificare (Bagging și Random Forest).

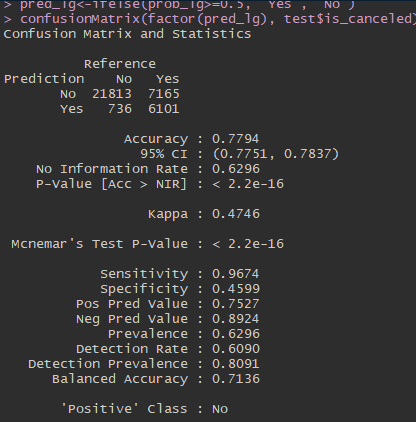
Am împărțit setul de date în set de antrenament și de test în proporții de 70% și 30% și stratificare după variabila *is\_canceled,* întrucât aceasta este și cea care mă intereasează.

1. **Regresia Logistică**

În vederea creării modelului de regresie logistică am utilizat toate cele 21 de atribute.



Mai departe, am realizat o predicție și am realizat matricea de confuzie.

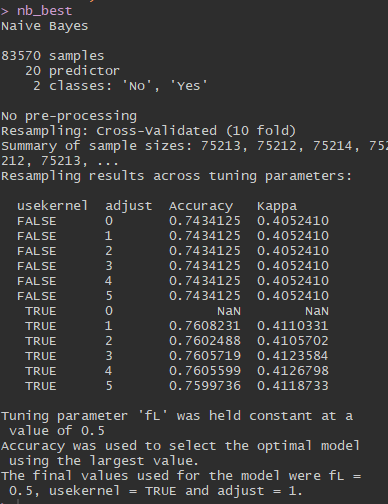
Se observă faptul că, modelul are o acuratețe de 77,94% și un P-value destul de bun. Totodată, no information rate este 0,6296 arătând faptul că există o șansă de 62,96% să facem o clasificare corectă dacă nu am avea nicio informație.

Clasa pozitivă este NO, adică sunt mai puține cazurile de anulare a rezervărilor.

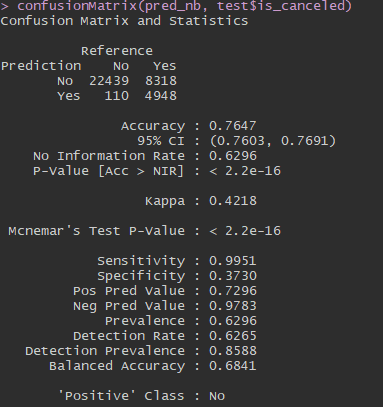
1. **Naive Bayes**

În vederea realizării unui model Naive Bayes, am realizat mai întâi un set *features* în care am păstrat toate atributele care nu sunt *is\_canceled.* Mai departe, am utilizat metoda de validare *10 folds cross validation* și am alcătuit modelul.

Pentru a găsi modelul optim am realizat o căutare extensivă, rezultatul fiind cel de mai jos.

Se observă faptul că modelul optim are o acuratețe de 76,08%, acesta fiind cel care care utilizează kernel și adjust=1.

Mai jos este prezentată matricea de confuzie realizată în urma predicției:

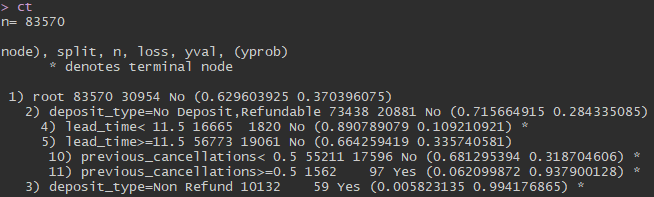


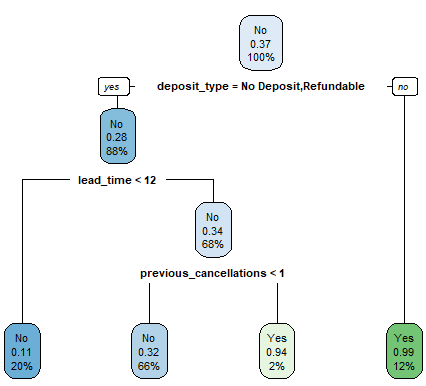
Modelul optim are o acuratețe de 76,47% asupra setului de test, mai slab decât regresia logistică.

1. **Arbori de clasificare**

**c1) Arbori utilizând abordarea CART**

Am realizat arborele folosind funcția **rpart**, având ca formulă atributul *is\_canceled* în funcție de restul atributelor din setul de date și metoda *class*, având în vedere faptul că aceasta va realiza o clasificare. Arborele rezultat este mai jos:



Se observă faptul că *deposit\_type* joacă un rol foarte important, acesta fiind așezat în vârful arborelui, împărțind setul în două, acesta clasifică clasa majoritară (care este *NO* în acest caz). Pe primul nivel este plasat *deposit\_type (No deposit/ Refundable)*, iar pe următorul nivel este *deposit\_type (No refund)*.

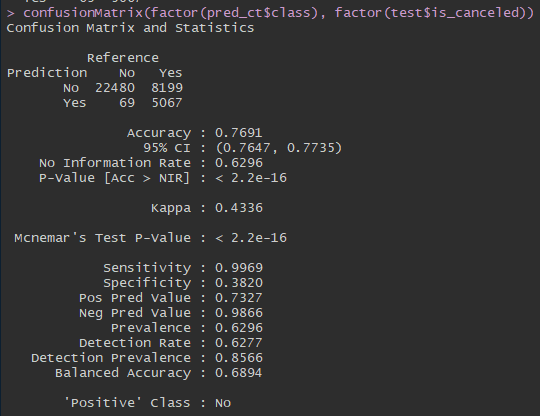
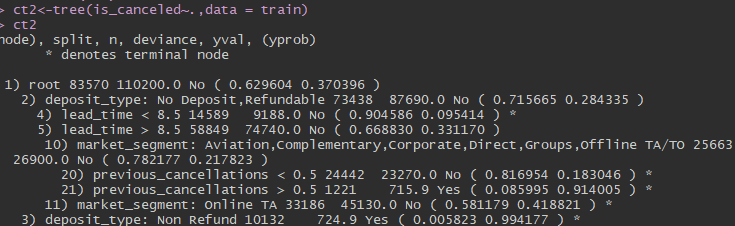
Următorul criteriu este *lead\_time*, timpul de așteptare care nu depășește 11.5 zile și, respectiv, care depășește 11.5 zile. Ultimul criteriu selectat este *previous\_cancellations.*

Putem concluziona astfel: dacă rezervarea este de tip *No deposit/Refundable* și timpul de așteptare nu depășește 12 zile, atunci rezervarea nu va fi anulată, sau dacă aceasta depășește 12 zile, iar presoana care face rezervarea nu și-a mai anulat rezervări în trecut, atunci nu o va anula nici atunci, însă, dacă aceasta și-a mai anulat rezervari atunci, cel mai probabil o va face și de data aceasta. Dacă rezervarea este de tip *Non Refund* atunci, aceasta va fi anulată.

Modelul are o acuratețe de 76,91% și un P-value bun, arătând că acesta este o metodă de clasificare destul de bună. Totuși, nu se deosebește prea mult de modelele anterioare, no information rate și P-value au aceeași valoare cu modelul Naive Bayes, iar acuratețea nu este decât cu foarte puțin (0.47) mai mare.

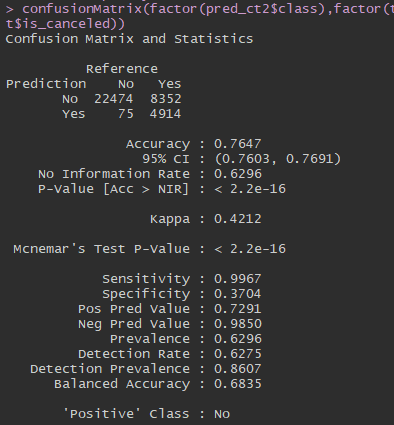
**c2) Arbori utilizând entropia**

Am creat un nou arbore de clasificare folosind entropia. Astfel, cu ajutorul comenzii **tree** și utilizând formula *is\_canceled* în funcție de restul atributelor.



Arborele rezultat este puțin mai adânc de data aceasta. Similar cu cel anterior, *deposit\_type* și *lead\_time* sunt plasate pe primele două nivele, însă de data aceasta perioada de timp pentru *lead\_time* este de 8.5 zile și nu de 11.5. Pe locul trei ca importanță se situează atributul *market\_segment* și abia după aceea *previous\_cancellations.*

Astfel, dacă rezervarea este nu este de tip *No Deposit/Refundable* atunci, aceasta va fi anulată. În schimb, dacă aceasta este de acel tip și dacă timpul de așteptare nu depășește 9 zile, atunci nu va fi anulată. Dacă se depășesc 9 zile și segmentul de marketing este nu este unul definit, atunci, din nou, aceasta nu va fi anulată. În schimb, dacă segmentul de marketing este unul definit și persoana care face rezervarea și-a mai anulat rezervări în trecut, atunci aceasta va anula rezervarea și acum, iar dacă nu a mai anulat nimic în trecut, atunci nu va anula nici acum.

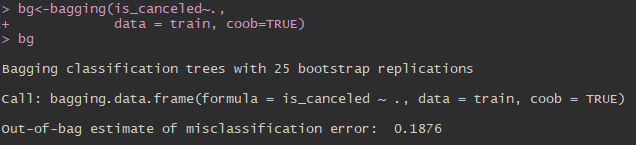


Acuratețea modelului nu este mai mare decât cea a arborelui anterior, ba chiar, este identică cu modelul Naive Bayes.

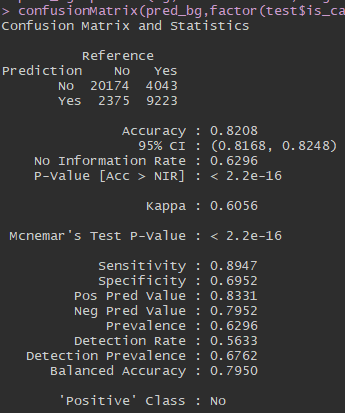
1. **Arbori de clasificare avansați**

**d1) Bagging**

În vederea creării acestui model am utilizat formula *is\_canceled* în funcție de restul atributelor și am setat *coob=TRUE,* adică va face estimarea performanțelor tuturor arborilor pe care îi va construi.

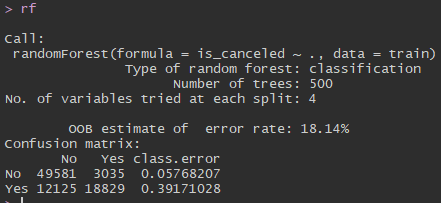


Astfel, acesta realizează 25 de bag-uri, având o eroare out-of-bag de 0.1876, ceea ce înseamna o acuratețe de 82%.

Modelul are o acuratețe de 82.02%, mult mai bună decât modelele anterioare. De asemenea, specificitatea este și ea mult mai bună decât cea a modelelor realizate până acum.

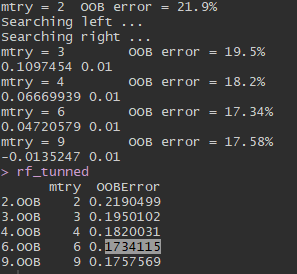
**d2) Random Forest**

La fel ca la modelele anterioare am utilizat formula *is\_canceled* în funcție de restul variabilelor.

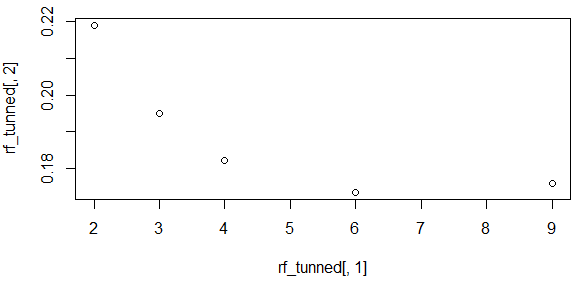
Astfel, cu ajutorul funcției **randomForest** au fost creați 500 de arbori, 4 variabile la fiecare nod.

Eroareai *out-of-bag* are valoare de 18.14%, rezultând o acuratețe de 82% asupra seturilor rămase.

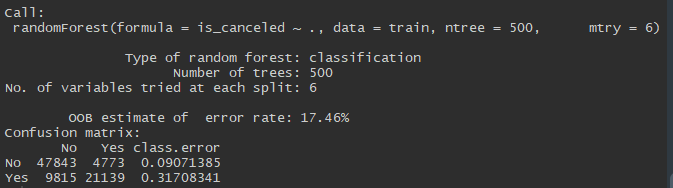
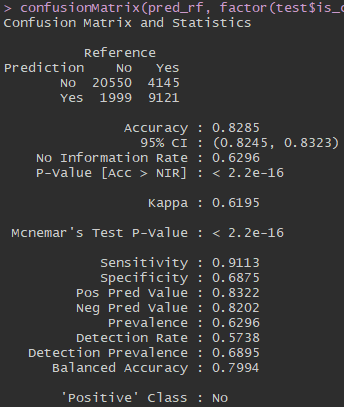
Mai departe, pentru a găsi cel mai bun model, am utilizat comanda **tuneRF** luând ca parametrii setul *features*, utilizat și la modelul Nave Bayes, realizând 500 de arbori, care vor începe cu 5 variabile și vor crește cu un *stepFactor=1.5*, oprindu-se atunci când performanța scade sub 1%.



Se observă că modelul cu eroarea OOB cea mai mică este pentru *mtry=6.*

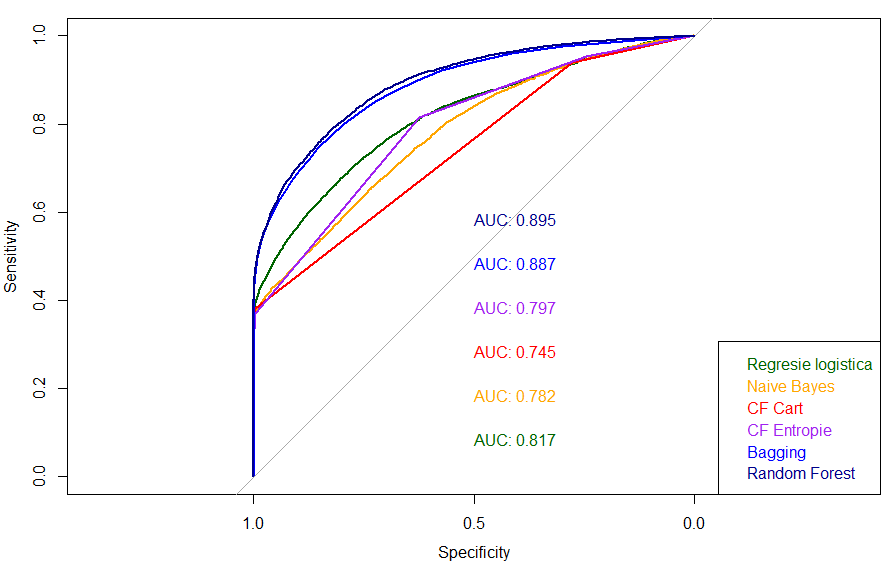


Mai departe, am construit cel mai bun model *random forrest* utilizând *mtry=6.*

**

Cel mai bun model *random forest* are o rată a erorii OOB de 17.46%.

Acuratețea pe setul de test este de 82.85%, având totodată senzitivitatea și specificitatea mai bune decât modelele anterioare.

La final, am realizat un grafic cu toate curbele ROC aferente modelelor realizate:

Modelul *Random Forest* are ocupă cea mai multă suprafață din grafic, întrucât acesta reușește să descopere 89.5% dintre rezervările anulate, acesta fiind urmat la o diferență mică de *Bagging*.

1. **Concluzii**

În urma rezultatelor obținute se poate răspunde am reușit să ating obiectivelel propuse la începutul proiectului. Astfel, se poate ajunge la concluzia că rezervările sunt în mare parte influențate de:

* Tipul de depozit
* Timpul de așteptare
* Anulările anterioare
* Segmentul de marketing

De asemenea, acestea vor pute fi prezise cu o acuratețe de 82,85% cu ajutorul modelului *Random* *Forest*.