

Universitatea Babeș-Bolyai  
Facultatea de Științe Economice și Gestiunea Afacerilor

# **PREDICȚIA PLASĂRII UNEI COMENZI DE CĂTRE CLIEȚII ONLINE**

Istrate Alexandra-Maria  
Sereșan Daria Emanuela  
Informatică Economică an III

**2022**

## I. Introducere

Unul dintre efectele pandemiei a fost creșterea bruscă a utilității comerțului electronic, precum și a vânzărilor prin intermediul acestuia. Astfel, comportamentul consumatorilor online a evoluat și el rapid. Această nouă eră a comerțului cu amănuntul este marcată de tranziția experienței de cumpărare a clienților de la cumpărăturile fizice la cumpărăturile online, având loc la o scară mult mai mare decât se preconiza.

Comportamentul consumatorilor online este modul în care consumatorii iau decizii de a achiziționa produse în comerțul electronic. Comportamentele în sine, cum este decizia de a face o achiziție, se bazează pe așteptări și nevoi în continuă evoluție.

În cadrul acestui studiu, am decis să analizăm setul de date astfel încât să putem răspunde la următoarele întrebări de cercetare:

1. Care dintre factorii de care dispunem în setul de date sunt mai relevanți pentru prezicerea comportamentului consumatorilor?
2. Cum se poate deduce din rata de ieșire, rata de respingere și media valorii paginii plasarea unei comenzi de către vizitatorii online?
3. Există vreo relație între rata de respingere și rata de ieșire a paginilor vizitate de clienți?
4. Care este comportamentul vizitatorilor online în funcție de rata de ieșire și tipul de utilizator?

Considerăm că răspunsurile la aceste întrebări pot contribui la îmbunătățirea site-urilor de către vânzători, pentru a capta atenția clienților și pentru a-i determina să plaseze comenzi online, menținând interesul acestora pentru produsele furnizate.

În cercetarea pe care am efectuat-o înainte de începerea studiului, am descoperit alte două analize efectuate asupra aceluiași set de date. Primul dintre acestea poate fi accesat la acest [link](#). Rezultatele acestui studiu spun că cele trei caracteristici cu cel mai mare impact sunt „PageValues”, „BounceRates” și „Administration\_Duration”. Cu cât media valorii paginii este mai mare, cu atât este mai mare șansa clientului de a face o achiziție. Cu cât este mai mare rata medie de respingere pentru pagina pe care o vizitează clientul, cu atât este mai mare șansa ca acesta să facă o achiziție. În schimb, cu cât un client petrece mai mult timp pe paginile administrative, cu atât este mai mică șansa ca acesta să finalizeze achiziția. Cel de-al doilea studiu găsit, care poate fi accesat la acest [link](#), prezintă următoarele concluzii: paginile cu rate de respingere mai mici s-au dovedit a fi un indicator puternic al faptului că vizita generează venituri, valoarea mai mare a paginii este un indicator puternic dacă vizita generează venituri și luna noiembrie generează cele mai favorabile raporturi de trafic, dar și alte luni, precum decembrie, au un trafic mare.

## II. Setul de date

Setul de date ales, intitulat „Online Shoppers Purchasing Intention” („Intenția de cumpărare a clienților online”) preluat de la adresa <https://www.kaggle.com/datasets/imakash3011/online-shoppers-purchasing-intention-dataset> conține date extrase din 12330 de sesiuni. Acesta a fost construit astfel încât fiecare sesiune să aparțină unui alt utilizator în decurs de un an. Aceste condiții au fost stabilite pentru a evita influențarea datelor de către o anumită campanie, utilizator sau perioadă. Conținutul setului de date este format din 18 attribute, dintre care 10 sunt numerice. Atributul „Revenue” reprezintă rezultatul analizei, și anume dacă un client a cumpărat sau nu în cadrul unei sesiuni, atribut care poate avea ca valori „True” sau „False”. Coloanele „Administrative”, „Administrative\_Duration”, „Informational”, „Informational\_Duration”, „ProductRelated”, „ProductRelated\_Duration”, reprezintă numărul de pagini vizitate de client în sesiunea respectivă și timpul total petrecut pe fiecare dintre aceste categorii de pagini.

Valorile acestor caracteristici sunt derivate din informațiile URL ale paginilor vizitate de utilizator și actualizate în timp real când un utilizator efectuează o acțiune, de exemplu, trecerea de la o pagină la alta. Atributele „BounceRates”, „ExitRates” și „PageValues” reprezintă valorile măsurate de Google Analytics pentru fiecare pagină de pe site-ul de comerț electronic. Valoarea funcției „BounceRates” pentru o pagină web se referă la procentul de vizitatori care intră pe site din acea pagină și apoi părăsesc site-ul fără a declanșa alte solicitări către serverul de analiză în timpul sesiunii respective. Pentru paginile web de pe care clienții aleg să părăsească site-ul se calculează funcția „ExitRates” care arată procentul vizionărilor acestora dintre toate paginile. „PageValues” reprezintă valoarea medie a venitului pentru o pagină web pe care un utilizator a vizitat-o înainte de a finaliza o tranzacție de comerț electronic. „SpecialDay” indică apropierea orei de vizitare a site-ului de o anumită zi specială (ziua mamei, ziua îndrăgostiților), pentru care probabilitatea ca sesiunile să fie finalizate cu o tranzacție este mai mare. Valoarea acestui atribut este determinată și în funcție de comerțul electronic și dinamica sa. Setul de date include, de asemenea, sistemul de operare, browserul, regiunea, tipul de trafic, tipul de vizitator ca vizitator care revine sau vizitator nou, o valoare booleană care indică dacă data vizitei este weekend și luna anului.

În continuare, am șters coloanele 10-15 și 11 (numărul ulterior al coloanei), deoarece unele attribute aveau valori pe care nu le cunoșteam (de exemplu: pentru „Region” valorile erau numerice, fără a fi specificată regiunea corespunzătoare; la „TrafficType” aveam, de asemenea, tot valori numerice pentru care semnificația era necunoscută) iar pe altele nu le-am considerat relevante pentru analiză („OperatingSystems”, „Browser”).

```
online_shoppers_intention[10:15]<- list(NULL)  
online_shoppers_intention[11]<- list(NULL)
```

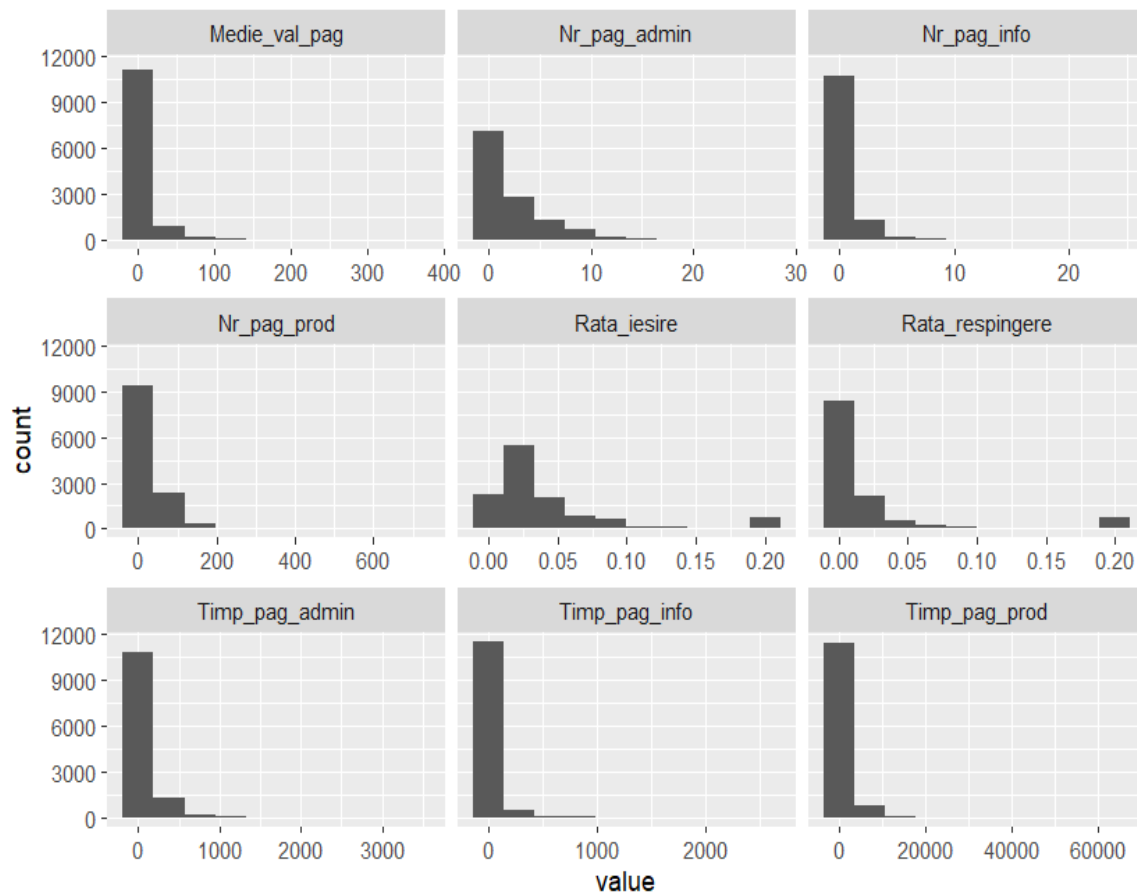
Pentru o înțelegere mai bună a datelor am redenumit coloanele după cum urmează: „Administrative” în „Nr\_pag\_admin”, „Administrative\_Duration” în „Timp\_pag\_admin”, „Informational” în „Nr\_pag\_info”, „Informational\_Duration” în „Timp\_pag\_info”, „ProductRelated” în „Nr\_pag\_prod”, „ProductRelated\_Duration” în „Timp\_pag\_prod”,

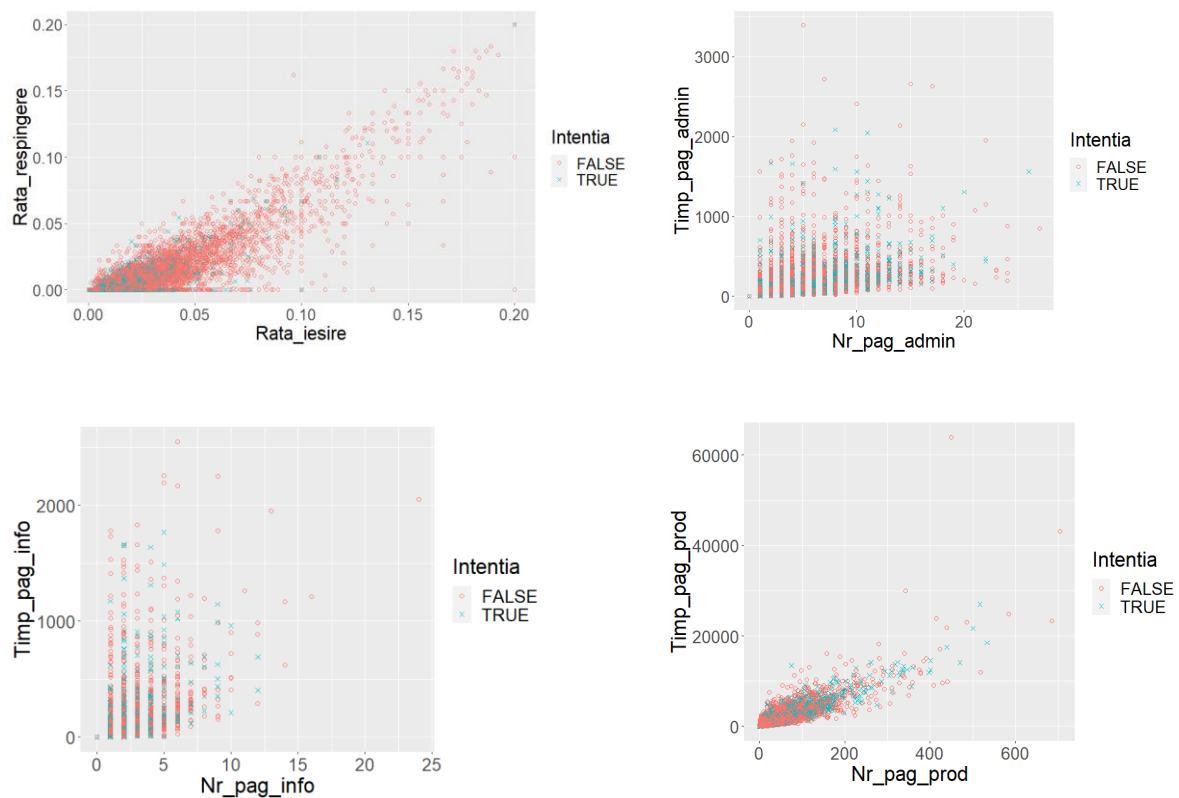
„BounceRates” în „Rata\_respingere”, „ExitRates” în „Rata\_iesire”, „PageValues” în „Medie\_val\_pag”, „VisitorType” în „Tip\_vizitator” și „Revenue” în „Intenția”.

Am vizualizat tabelul în varianta actuală:

	Nr_pag_admin	Timp_pag_admin	Nr_pag_info	Timp_pag_info	Nr_pag_prod	Timp_pag_prod	Rata_respingere	Rata_iesire	Medie_val_pag	Tip_vizitator	Intenția
1	0	0.00000	0	0.000	1	0.000000	0.200000000	0.200000000	0.000000	Returning_Visitor	FALSE
2	0	0.00000	0	0.000	2	64.000000	0.000000000	0.100000000	0.000000	Returning_Visitor	FALSE
3	0	0.00000	0	0.000	1	0.000000	0.200000000	0.200000000	0.000000	Returning_Visitor	FALSE
4	0	0.00000	0	0.000	2	2.666667	0.050000000	0.140000000	0.000000	Returning_Visitor	FALSE
5	0	0.00000	0	0.000	10	627.500000	0.020000000	0.050000000	0.000000	Returning_Visitor	FALSE
6	0	0.00000	0	0.000	19	154.216667	0.015789474	0.024561404	0.000000	Returning_Visitor	FALSE
7	0	0.00000	0	0.000	1	0.000000	0.200000000	0.200000000	0.000000	Returning_Visitor	FALSE
8	1	0.00000	0	0.000	0	0.000000	0.200000000	0.200000000	0.000000	Returning_Visitor	FALSE
9	0	0.00000	0	0.000	2	37.000000	0.000000000	0.100000000	0.000000	Returning_Visitor	FALSE
10	0	0.00000	0	0.000	3	738.000000	0.000000000	0.022222222	0.000000	Returning_Visitor	FALSE
11	0	0.00000	0	0.000	3	395.000000	0.000000000	0.066666667	0.000000	Returning_Visitor	FALSE
12	0	0.00000	0	0.000	16	407.750000	0.018750000	0.025833333	0.000000	Returning_Visitor	FALSE
13	0	0.00000	0	0.000	7	280.500000	0.000000000	0.028571429	0.000000	Returning_Visitor	FALSE
14	0	0.00000	0	0.000	6	98.000000	0.000000000	0.066666667	0.000000	Returning_Visitor	FALSE
15	0	0.00000	0	0.000	2	68.000000	0.000000000	0.100000000	0.000000	Returning_Visitor	FALSE
16	2	53.00000	0	0.000	23	1668.285119	0.008333333	0.016312636	0.000000	Returning_Visitor	FALSE
17	0	0.00000	0	0.000	1	0.000000	0.200000000	0.200000000	0.000000	Returning_Visitor	FALSE
18	0	0.00000	0	0.000	13	334.966667	0.000000000	0.007692308	0.000000	Returning_Visitor	FALSE
19	0	0.00000	0	0.000	2	32.000000	0.000000000	0.100000000	0.000000	Returning_Visitor	FALSE
20	0	0.00000	0	0.000	20	2981.166667	0.000000000	0.010000000	0.000000	Returning_Visitor	FALSE

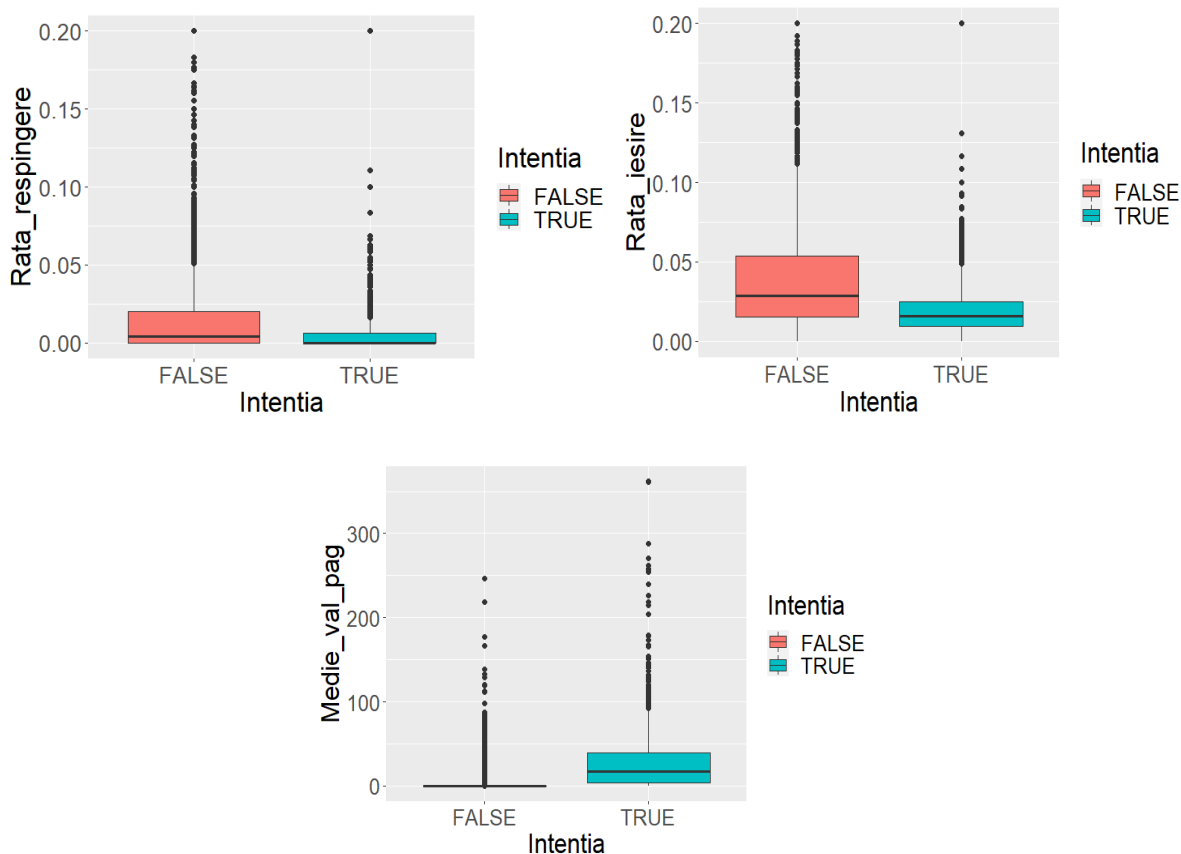
Pentru a ne putea forma o opinie de ansamblu despre influențele variabilelor asupra intenției, am realizat următoarea histogramă care cuprinde toate variabilele numerice din setul de date prelucrat.





Am creat patru vizualizări pentru a deduce câteva relații preliminare. În prima vizualizare observăm că există o relație strânsă între rata de respingere și rata de ieșire întrucât valorile acestora formează o conglomerare care ia forma unei drepte. Atât persoanele care plasează comenzi cât și persoanele care părăsesc site-urile fără a cumpăra ceva vizitează de obicei pagini care au ambele rate mici, însă este evident faptul că o mică parte din persoanele care nu cumpără vizitează și pagini care au ambele rate mari. În cea de-a doua vizualizare este redat faptul că indiferent de intenția de cumpărare, majoritatea persoanelor petrec cel mult 800 de minute și vizualizează cel mult 15 pagini de tip administrativ pe an. Relația dintre variabilele *Timp\_pag\_info* și *Nr\_pag\_info* arată că, indiferent de intenția de cumpărare, clienții care obișnuiesc să citească mai puține pagini de tip informațional tind să petreacă mai mult timp pe acestea. Majoritatea persoanelor care vizitează pagini legate de produse accesează cel mult 200 de pagini pe an și petrec maxim 10000 minute pe acestea (aproximativ 167 de ore) pe an.

Din aceste patru grafice se constată că doar relația dintre rata de respingere și rata de ieșire influențează intenția de cumpărare a clienților online.



Dorim să prezicem dacă o persoană va plasa o comandă online pe baza ratei de respingere, a ratei de ieșire și a mediei valorii paginii. Majoritatea clienților plasează o comandă pe pagini cu o valoare mai mică de  $\sim 0,008$  a ratei de respingere. Există clienți care au intenția de a cumpăra și pe paginile cu o rată de respingere mai mare de 0,008, însă aceștia sunt în număr foarte mic. Pentru majoritatea vizitatorilor care nu efectuează o comandă, rata de respingere a paginilor poate atinge pragul de 0,02, însă există și în acest caz un număr mic de utilizatori care accesează pagini cu o rată de respingere mult mai mare. Valoarea ratei de ieșire pentru majoritatea clienților care nu efectuează o comandă se situează între valorile  $\sim 0,02$  și  $\sim 0,06$ , iar pentru clienții care plasează o comandă, valoarea ratei de ieșire este cuprinsă între  $\sim 0,01$  și  $\sim 0,025$ . Pentru ambele cazuri există și utilizatori, mult mai puțini însă, care vizitează pagini cu o rată de ieșire mai mare. În cazul utilizatorilor care nu au intenția de a plasa o comandă online nu există o tendință generală a mediei valorii paginilor. Pentru majoritatea clienților care cumpără, paginile vizitate au o medie a valorii de cel mult  $\sim 43$  de unități monetare, însă există comenzi și de pe pagini a căror medie a valorii depășește acest număr.

### III. Rezultate și discuții

#### Metoda regresiei logistice

Am ales metoda regresiei liniare întrucât aceasta permite înțelegerea relațiilor dintre variabile. Analiza prin regresie indică ce variabile de tip predictor dintr-un model sunt semnificative pentru variabila țintă din punct de vedere statistic.

Am creat un tibble 2 x 2 pentru numărarea cazurilor „True” și „False” al atributului „Intentia”. Avem ca rezultate 10422 de înregistrări cu valoarea „False” (vizitatorii paginilor nu au încheiat sesiunea cu plasarea unei comenzi) și 1908 de înregistrări cu valoarea „True” (utilizatorii au plasat o comandă online). Raportul dintre cele două valori este unul disproporționat, numărul cazurilor cu valoarea „False” fiind de cinci ori mai mare decât al cazurilor cu valoarea „True”.

```
> by_Intentia <- group_by(online_shoppers_intention, Intentia)
> summarize(by_Intentia, count=n())
# A tibble: 2 x 2
  Intentia count
  <lgl>      <int>
1 FALSE    10422
2 TRUE      1908
```

În urma observării graficelor, am ales să continuăm analiza prin realizarea de tibble-uri, regresii simple și tribble-uri pentru attributele „Rata\_iesire”, „Rata\_respingere” și „Medie\_val\_pag”.

```
> mod <- glm (data = online_shoppers_intention, Intentia ~ Rata_iesire, family = binomial)
> summary(mod)

Call:
glm(formula = Intentia ~ Rata_iesire, family = binomial, data = online_shoppers_intention)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8879 -0.6805 -0.5163 -0.1713  3.9248

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.72734    0.04226  -17.21  <2e-16 ***
Rata_iesire  -34.87057    1.61324  -21.61  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

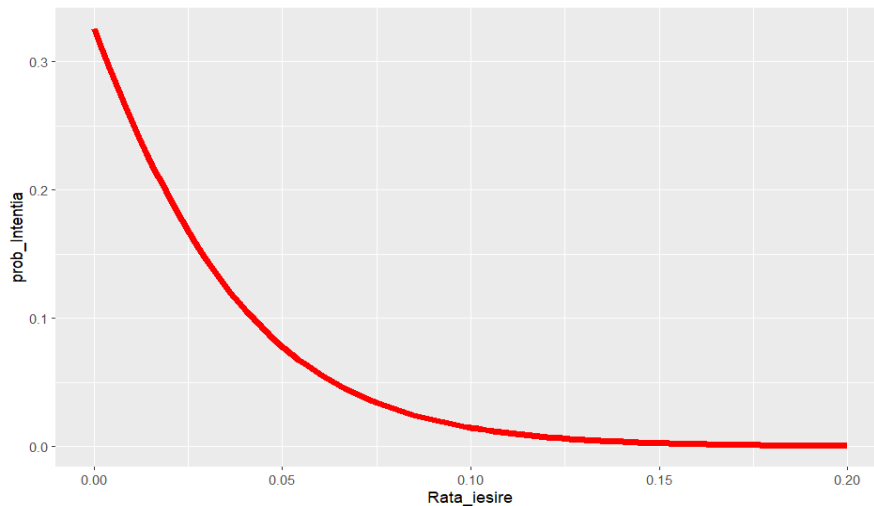
    Null deviance: 10624.8  on 12329  degrees of freedom
Residual deviance: 9665.7  on 12328  degrees of freedom
AIC: 9669.7

Number of Fisher Scoring iterations: 6
```

Estimarea coeficientului indică variația medie a log odds ale variabilei „Intentia” asociate cu o creștere de o unitate a ratei de ieșire. În acest caz, o creștere de o unitate a ratei de ieșire este asociată cu o modificare medie de -34,87057 în log odds pentru variabila „Intentia” cu o valoare de 1. Valorile mai mari ale ratei de ieșire sunt asociate cu o probabilitate mai mică pentru ca variabila „Intentia” să ia o valoare de 1. Eroarea standard oferă o idee despre variabilitatea asociată cu estimarea coeficientului. Valoarea z se obține împărțind estimarea coeficientului la eroarea standard. Valoarea z pentru variabila ratei de ieșire este  $-34,87057 / 1,61324 = -21.61$ . Valoarea p  $\Pr(>|z|)$  indică probabilitatea asociată cu o anumită valoare z. Acest lucru redă cât de bine variabila ratei de ieșire este capabilă să prezică valoarea variabilei „Intentia” din model. Abaterea nulă arată cât de bine poate fi prezisă variabila „Intentia” de către un model cu doar un termen de interceptare. Abaterea reziduală redă cât de bine poate fi prezisă variabila „Intentia” în model. Cu cât această valoare este mai mică, cu atât modelul este mai capabil să prezică valoarea variabilei „Intentia”.

```
> ggplot() +
+ geom_line(data = grid, aes(Rata_iesire, probab_Intentia), color = "red", size = 2)
> nd <- tribble(~Rata_iesire, 0.00, 0.05, 0.10, 0.15, 0.20)
> predicted <- predict(mod, newdata = nd, type = "response")
> predicted
```

```
1 2 3 4 5
0.3257779897 0.0779255824 0.0145658874 0.0025785960 0.0004519637
```



În acest grafic sunt reprezentate probabilitățile unui utilizator de a plasa o comandă pentru fiecare valoare a ratei de ieșire. Se observă că începând cu valoarea ratei de ieșire de 0, probabilitatea plasării unei comenzi scade. În acest sens, pentru o valoare de 0,05 a ratei de ieșire, probabilitatea este de 7%, pentru o valoare de 0,10, probabilitatea este de 2%, iar începând cu valoarea de 0,15, probabilitatea tinde să fie nulă.

```
> mod_Rata_respingere <- glm(data = online_shoppers_intention, Intentia ~ Rata_respingere, family
= binomial)
> summary(mod_Rata_respingere)
```

```
Call:
glm(formula = Intentia ~ Rata_respingere, family = binomial,
    data = online_shoppers_intention)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6717 -0.6717 -0.5944 -0.2426  3.9622
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.3741     0.0284  -48.39  <2e-16 ***
Rata_respingere -32.3758     2.3333  -13.88  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)
```

```
Null deviance: 10625 on 12329 degrees of freedom
Residual deviance: 10088 on 12328 degrees of freedom
AIC: 10092
```

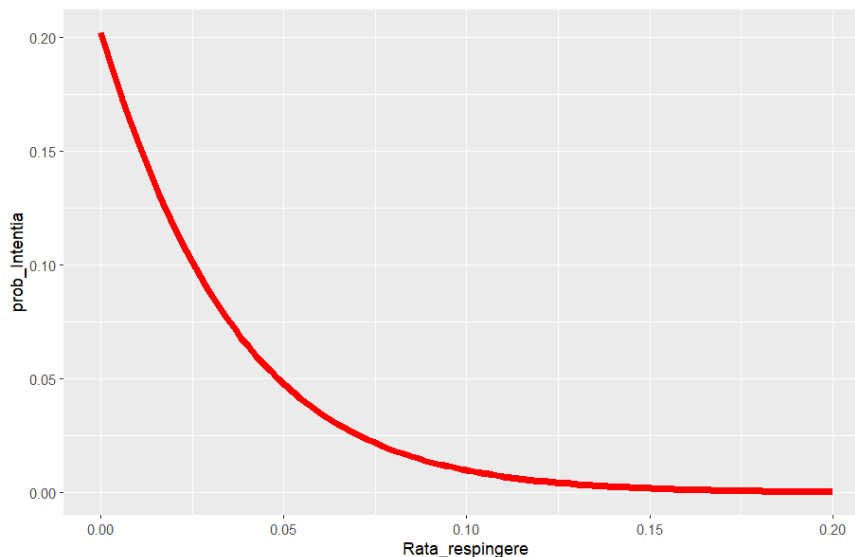
```
Number of Fisher Scoring iterations: 7
```

O creștere de o unitate a ratei de respingere este asociată cu o modificare medie de -32,3758 în log odds pentru variabila „Intentia” cu o valoare de 1. Probabilitatea valorii 1 a variabilei „Intentia”, scade pentru valorile mai mari ale ratei de respingere. Valoarea z pentru variabila ratei de respingere este  $-32,3758 / 2,3333 = -13,88$ .



```
> ggplot() +
+ geom_line(data = grid, aes(Rata_respingere, prob_Intentia), color = "red", size = 2)
> nrata_resp <- tribble(~Rata_respingere, 0.00, 0.05, 0.10, 0.15, 0.20)
> predicted <- predict(mod, newdata = nrata_resp, type = "response")
> predicted
```

```
1 2 3 4 5
0.2019495777 0.0477457669 0.0098368952 0.0019645659 0.0003898707
```



Pentru o valoare a ratei de respingere de 0 probabilitatea plasării unei comenzi este de 20%. Aceasta se modifică la o valoare de 5% pentru o rată de respingere de 0,05. Probabilitatea este de 1% pentru o valoare de 0,10 a ratei de respingere. Pentru valorile de 0,15 și 0,2, probabilitatea tinde spre 0.

```
Call:
glm(formula = Intentia ~ Medie_val_pag, family = binomial, data = online_shoppers_intention)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-6.2531	-0.4135	-0.4135	-0.4135	2.2369

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-2.416447	0.034421	-70.20	<2e-16 ***
Medie_val_pag	0.089023	0.002341	38.03	<2e-16 ***

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

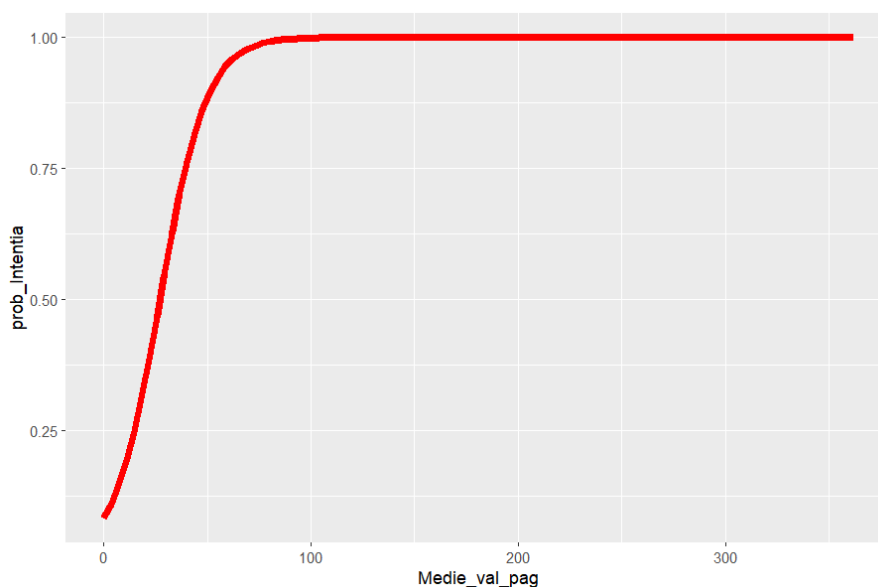
Null deviance: 10624.8 on 12329 degrees of freedom  
Residual deviance: 7888.8 on 12328 degrees of freedom  
AIC: 7892.8

Number of Fisher Scoring iterations: 5

O creștere de o unitate a mediei valorii paginii este asociată cu o modificare medie de 0,089023 în log odds pentru variabila „Intentia” cu o valoare de 1. Probabilitatea valorii 1 a variabilei „Intentia” crește pentru valorile mai mari ale mediei valorii paginii. Valoarea z pentru variabila media valorii paginii este  $0,089023 / 0,002341 = 38,03$ .

```
> ggplot() +
+ geom_line(data = grid, aes(Medie_val_pag, prob_Intentia), color = "red", size = 2)
> nmedie_pag <- tribble(~Medie_val_pag, 0, 100, 200, 300)
> predicted <- predict(mod, newdata = nmedie_pag, type = "response")
> predicted
```

```
1 2 3 4
0.08192707 0.99847750 0.99999979 1.00000000
```



Pentru o valoare a mediei valorii paginii de 0 probabilitatea plasării unei comenzi este de 8%. În continuare, probabilitatea crește brusc, atingând probabilitatea de plasare a unei comenzi de 100% pentru valorile de 100 u.m., 200 u.m. și 300 u.m. pentru media valorii paginii.

Am realizat o regresie multiplă cu attributele „Rata\_iesire”, „Rata\_respingere” și „Medie\_val\_pag” întrucât le-am considerat a fi cele mai relevante.

```
> mod_all <- glm(data = online_shoppers_intention, Intentia ~ Rata_respingere + Rata_iesire +
+ Medie_val_pag, family = binomial)
> summary(mod_all)
```

```
Call:
glm(formula = Intentia ~ Rata_respingere + Rata_iesire + Medie_val_pag,
    family = binomial, data = online_shoppers_intention)
```

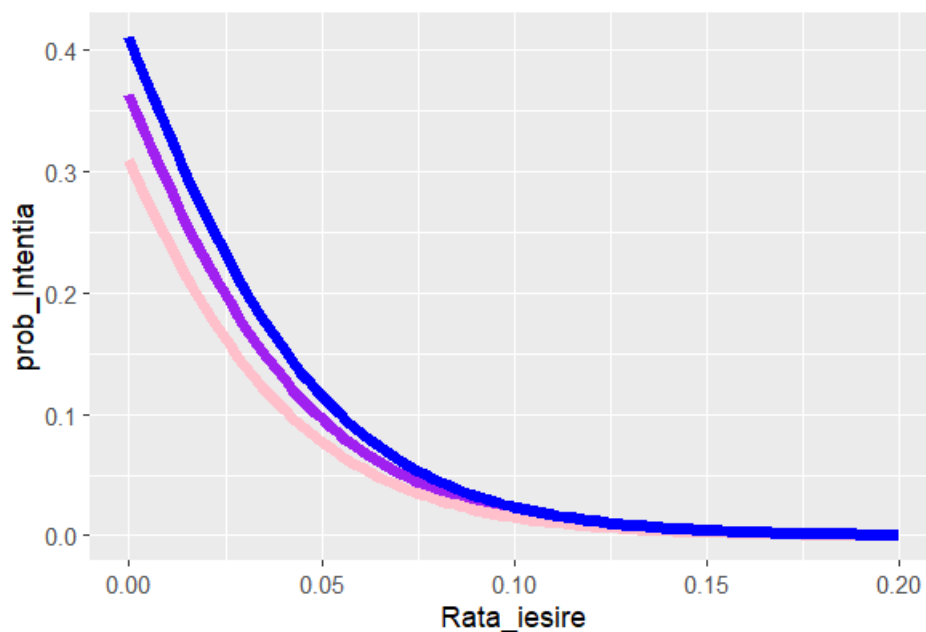
```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.9610  -0.4952  -0.4086  -0.1996   3.5006
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.711293   0.055068  -31.076  <2e-16 ***
Rata_respingere -0.236360   3.137409  -0.075    0.94
Rata_iesire    -21.830925   2.247061  -9.715  <2e-16 ***
Medie_val_pag   0.079028   0.002312  34.179  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)
```

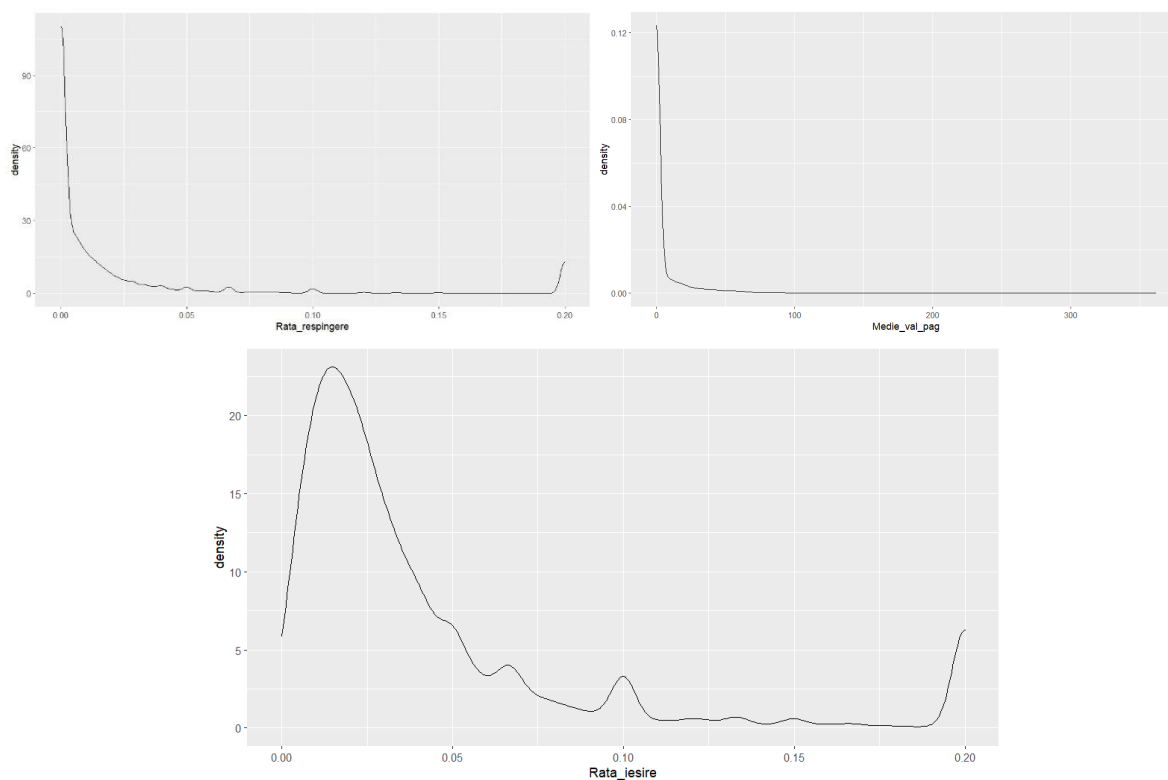
```
Null deviance: 10624.8 on 12329 degrees of freedom
Residual deviance: 7555.9 on 12326 degrees of freedom
AIC: 7563.9
```

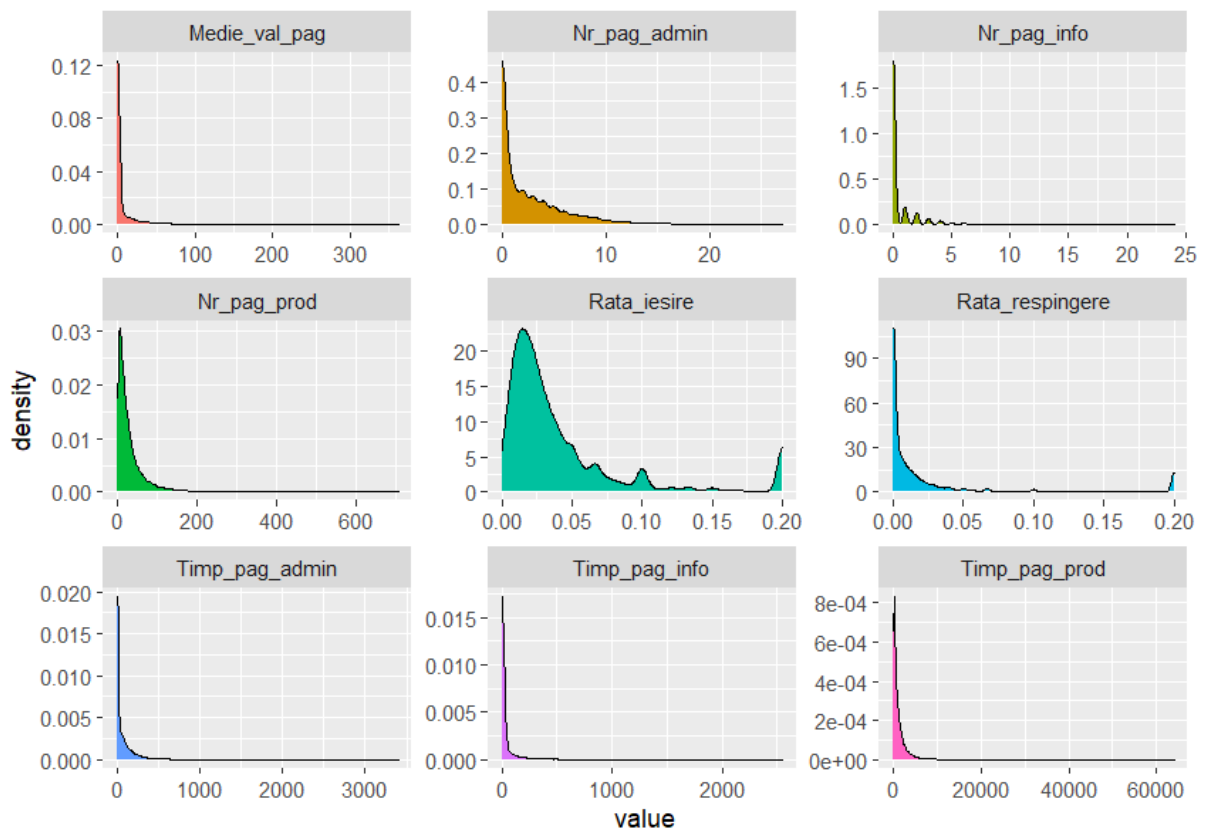
```
Number of Fisher Scoring iterations: 7
```



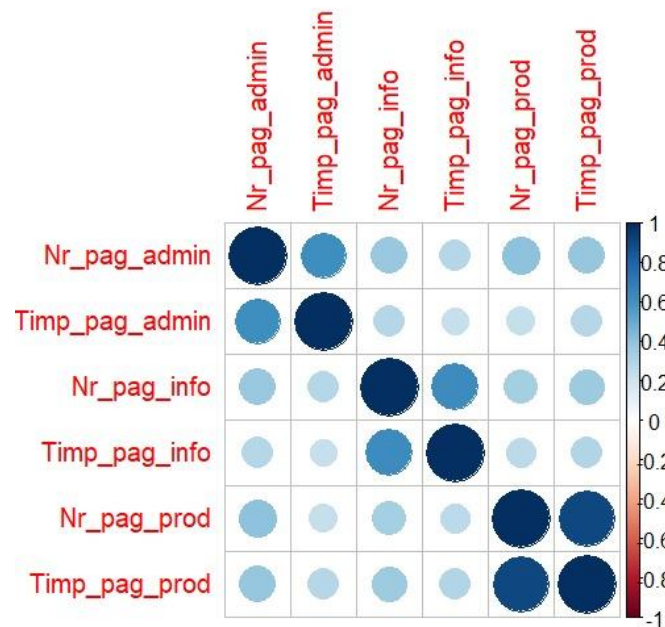
Din acest grafic reiese că probabilitatea de a cumpăra este mai mare în cazul unui utilizator necunoscut (marcat cu culoarea albastră), față de cea pentru un utilizator nou (culoarea mov) și aceasta la rândul ei, este mai mare decât cea pentru un utilizator vechi (culoarea roz). Rata de ieșire are valori apropiate în cele 3 cazuri.

## Metoda Naive Bayes





Corelația dintre variabile:



```

> table(train$Intentia)

FALSE TRUE
7295 1335
> table(test$Intentia)

FALSE TRUE
3127 573

> modNbSimpleCV
Naive Bayes

8630 samples
10 predictor
2 classes: 'FALSE', 'TRUE'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 7767, 7768, 7768, 7767, 7767, 7767, ...
Resampling results across tuning parameters:

  usekernel Accuracy Kappa
FALSE      0.8107738 0.1645780
TRUE       0.7625693 0.2692368

Tuning parameter 'fL' was held constant at a value of 0
Tuning
parameter 'adjust' was held constant at a value of 1
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were fL = 0, usekernel = FALSE and
adjust = 1.
>

> confusionMatrix(modNbSimpleCV)
Cross-Validated (10 fold) Confusion Matrix

(entries are percentual average cell counts across resamples)

      Reference
Prediction FALSE TRUE
FALSE    77.6 12.0
TRUE     6.9 3.5

Accuracy (average) : 0.8108

> modNbCVSearch$results %>%
+ top_n(5, wt = Accuracy) %>%
+ arrange(desc(Accuracy))
  usekernel fL adjust Accuracy Kappa AccuracySD KappaSD
1 FALSE 0.5 0 0.8028925 0.1969886 0.0122766 0.03350492
2 FALSE 0.5 1 0.8028925 0.1969886 0.0122766 0.03350492
3 FALSE 0.5 2 0.8028925 0.1969886 0.0122766 0.03350492
4 FALSE 0.5 3 0.8028925 0.1969886 0.0122766 0.03350492
5 FALSE 0.5 4 0.8028925 0.1969886 0.0122766 0.03350492
6 FALSE 0.5 5 0.8028925 0.1969886 0.0122766 0.03350492

> confusionMatrix(modNbCVSearch)
Cross-Validated (10 fold) Confusion Matrix

(entries are percentual average cell counts across resamples)

      Reference
Prediction FALSE TRUE
FALSE    75.8 11.0
TRUE     8.7 4.4

Accuracy (average) : 0.8029

```

```

> modNbCVSearch$results %>%
+ top_n(5, wt = Accuracy) %>%
+ arrange(desc(Accuracy))
  usekernel fl adjust Accuracy Kappa AccuracySD KappaSD
1 FALSE 0.5 0 0.8028925 0.1969886 0.0122766 0.03350492
2 FALSE 0.5 1 0.8028925 0.1969886 0.0122766 0.03350492
3 FALSE 0.5 2 0.8028925 0.1969886 0.0122766 0.03350492
4 FALSE 0.5 3 0.8028925 0.1969886 0.0122766 0.03350492
5 FALSE 0.5 4 0.8028925 0.1969886 0.0122766 0.03350492
6 FALSE 0.5 5 0.8028925 0.1969886 0.0122766 0.03350492

> predProb
      FALSE      TRUE
3 9.999995e-01 5.364129e-07
4 9.999300e-01 6.995786e-05
6 9.971117e-01 2.888335e-03
10 9.987178e-01 1.282229e-03
19 9.999545e-01 4.545966e-05
23 9.995120e-01 4.879679e-04
24 9.999545e-01 4.546672e-05
28 9.999937e-01 6.281361e-06
29 9.998469e-01 1.531320e-04
34 9.999866e-01 1.340666e-05
35 9.985754e-01 1.424567e-03

```

Așadar, putem răspunde la întrebările de cercetare formulate inițial:

1. Factorii care s-au dovedit a fi cei mai relevanți în prezicerea comportamentului consumatorului sunt: rata de ieșire, rata de respingere și media valorii paginii.
2. Plasarea unei comenzi de către vizitatorii online poate fi dedusă în următorul mod: cu cât rata de ieșire și rata de respingere sunt mai mari, cu atât probabilitatea de a plasa o comandă este mai mică și cu cât media valorii paginii este mai mare, cu atât probabilitatea ca sesiunea să se încheie cu o comandă este mai mare.
3. Între rata de ieșire și rata de respingere a paginilor există o relație strânsă, întrucât sunt direct proporționale.
4. Probabilitatea de a plasa o comandă în funcție de tipul de utilizator, în ordine descrescătoare este: utilizator de tip necunoscut („other”), utilizator nou și utilizator vechi, între ratele de ieșire nefiind diferențe majore.

#### IV. Concluzii

Având în vedere rezultatele obținute, putem concluziona că rata de ieșire are un impact major asupra intenției de cumpărare a clienților online, deoarece cu cât aceasta are o valoare mai mare, cu atât scad șansele ca sesiunea să se finalizeze cu plasarea unei comenzi. Același lucru se aplică și ratei de respingere, ambele fiind invers proporționale cu probabilitatea de a cumpăra. Pe de altă parte, media valorii paginii joacă, de asemenea, un rol important. În acest sens, cu cât media valorii paginii este mai mare, cu atât este mai probabil ca vizitatorul să efectueze o comandă.