ACADEMIA DE STUDII ECONOMICE DIN BUCUREȘTI
Facultatea de Cibernetică, Statistică și Informatică Economică
Clădirea Virgil Madgearu, Calea Dorobanți 15-17, București 010552, 021 311 2066

PROIECT STATISTICĂ NEPARAMETRICĂ

Studiu de caz

Analiza eficienței campaniilor de marketing telefonic prin metode de statistică neparametrică

Studenți:
NEDELCU Alexandru – Daniel grupa 1066 F
NEGULESCU Cristian – Petruț grupa 1066 F

Cadru didactic coordonator:
prof.univ.dr BOBOC Cristina-Rodica

București, 2024

CUPRINS

ACADEMIA DE STUDII ECONOMICE DIN BUCUREȘTI
Facultatea de Cibernetică, Statistică și Informatică Economică
Clădirea Virgil Madgearu, Calea Dorobanți 15-17, București 010552, 021 311 2066

INTRODUCERE

Pentru lucrarea de față am ales această[1] bază de date de pe site-ul Kaggle.com și am utilizat-o pentru a testa ipoteze de statistică neparametrică.

Setul de date este ideal pentru analiza noastră deoarece conține informații detaliate despre campaniile de marketing telefonic desfășurate de o instituție bancară portugheză. Aceste informații ne permit să evaluăm eficacitatea campaniilor și să investigăm diferiți factori care ar putea influența decizia clienților de a contracta un depozit la termen.

Prin utilizarea acestei baze de date, avem posibilitatea de a aplica tehnici de statistică neparametrică pentru a analiza relațiile dintre variabile fără a presupune o distribuție specifică a datelor. Aceasta ne oferă flexibilitatea de a testa ipoteze și de a obține rezultate robuste, indiferent de forma distribuției datelor. Astfel, putem evalua eficiența campaniilor de marketing și putem identifica factorii cheie care contribuie la succesul sau eșecul acestora.

Setul de date are 41188 rânduri (apeluri către clienți) și 21 de coloane (variabile) care descriu anumite aspecte ale apelului.

PREZENTAREA BAZEI DE DATE, VARIABILELOR ȘI TRANSFORMĂRILE ACESTORA

Tabel 1. Bază de date

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	age	Numeric	2	0		None	None	8	Right	Scale	Input
2	job	String	13	0		None	None	13	Left	Nominal	Input
3	marital	String	8	0		None	None	8	Left	Nominal	Input
4	education	String	19	0		{0, unknow...	None	19	Left	Ordinal	Input
5	default	String	7	0		None	None	7	Left	Nominal	Input
6	housing	Numeric	1	0		{0, no}...	None	7	Right	Nominal	Input
7	loan	Numeric	1	0		{0, no}...	None	7	Right	Nominal	Input
8	contact	String	9	0		None	None	9	Left	Nominal	Input
9	month	Date	9	0		None	None	11	Right	Nominal	Input
10	day_of_week	Date	3	0		None	None	8	Right	Nominal	Input
11	duration	Numeric	4	0		None	None	8	Right	Scale	Input
12	campaign	Numeric	2	0		None	None	8	Right	Scale	Input
13	pdays	Numeric	3	0		None	None	8	Right	Scale	Input
14	previous	Numeric	1	0		None	None	8	Right	Nominal	Input
15	poutcome	String	11	0		None	None	11	Left	Nominal	Input
16	emp.var.rate	Numeric	4	1		None	None	8	Right	Scale	Input
17	cons.price.idx	Numeric	6	3		None	None	8	Right	Scale	Input
18	cons.conf.idx	Numeric	5	1		None	None	8	Right	Scale	Input
19	euribor3m	Numeric	5	3		None	None	8	Right	Scale	Input
20	nr.employed	Numeric	6	1		None	None	8	Right	Scale	Input
21	subscribed	Numeric	1	0		{0, no}...	None	15	Right	Ordinal	Input

[1] https://www.kaggle.com/datasets/pankajbhowmik/bank-marketing-campaign-subscriptions

Descrierea bazei de date și a variabilelelor utilizate

Variabile care descriu atribute legate direct de client:

a. vârstă

b. job: tipul jobului (ex. 'administrator', 'tehnician', 'șomer', etc.)

c. stare civilă: stare civilă ('căsătorit', 'necăsătorit', 'divorțat', 'necunoscut')

d. educație: nivelul de educație ('elementar.4ani', 'liceu', 'elementar.6ani', 'elementar.9ani', 'curs profesional', 'necunoscut', 'universitate', 'analfabet')

e. restanțe: dacă clientul are credite restante ('nu', 'necunoscut', 'da')

f. locuință: dacă clientul are un credit ipotecar ('nu', 'necunoscut', 'da')

g. împrumut: dacă clientul are un împrumut personal ('nu', 'necunoscut', 'da')

Variabile legate de ultimul contact al campaniei curente:

a. contact: tipul de comunicare ('telefon', 'mobil')

b. lună: luna ultimului contact

c. ziua_săptămânii: ziua ultimului contact

d. durată: durata apelului (în secunde)

Alte variabile legate de campanie (campanii):

a. campanie: numărul de contacte realizate în timpul acestei campanii și pentru acest client

b. zile_de_la_ultimul_contact: numărul de zile trecute de la ultimul contact al clientului dintr-o campanie anterioară

c. contacte_anterioare: numărul de contacte realizate înainte de această campanie și pentru acest client

d. rezultat_campanie_anterioară: rezultatul campaniei de marketing anterioare ('inexistent', 'eșec', 'succes')

Variabile socioeconomice:

a. rata_variației_angajării: indicator trimestrial

b. indicele_prețurilor_de_consum: indicator lunar

c. indicele_confidenței_consumatorilor: indicator lunar

d. euribor3m: rata euribor pe 3 luni - indicator zilnic

e. număr_angajați: indicator trimestrial

Transformarea variabilelor

Old --> New:

'basic.4y' --> '1'
'unknown' --> '0'
'basic.6y' --> '2'
'basic.9y' --> '3'
'high.school' --> '4'
'illiterate' --> '5'
'professional.course' --> '6'
'university.degree' --> '7'

Value Labels:

Value	Label
0	unknown
1	basic.4y
2	basic.6y
3	basic.9y
4	high.school
5	illiterate
6	professional.course
7	university.degree

Old --> New:

'married' --> '1'
ELSE --> '0'

Value Labels:

Value	Label
0	unmarried and others
1	married

1. INDICATORI NEPARAMETRICI

1.1 COEFICIENTUL DE CONTINGENȚĂ CHI-SQUARE (variabilă nominală – variabilă ordinală)

Există asocierea între variabila "education" și variabila "job"

Pentru a testa existența asocierii vom folosii coeficientul de contingență Chi-Square.

H0: Nu există o asociere semnificativă între variabilele categorice "education" și "job"

H1 Există o asociere semnificativă între variabilele categorice "education" și "job"

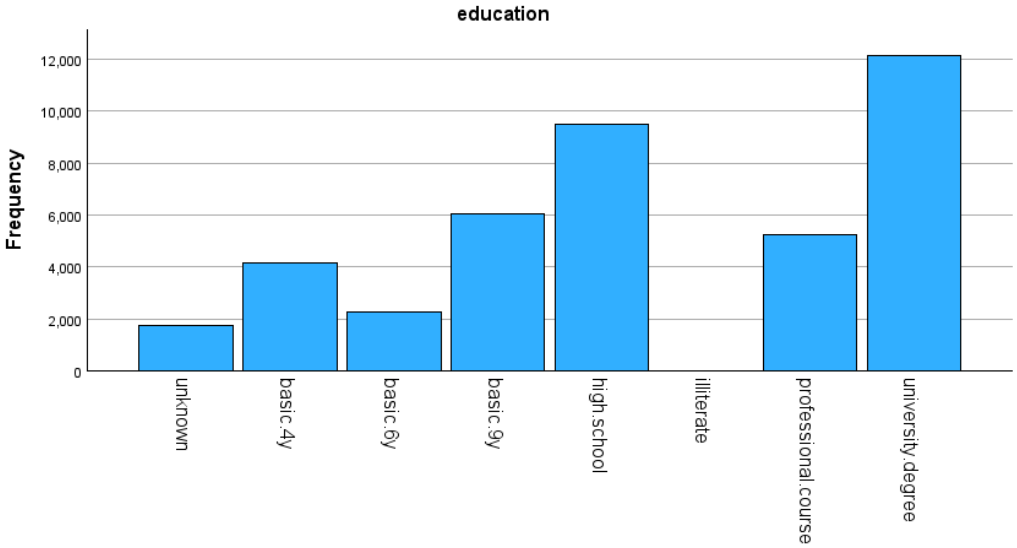

Fig.1 Distribuția variabilei education
Sursa: prelucrare autori SPSS

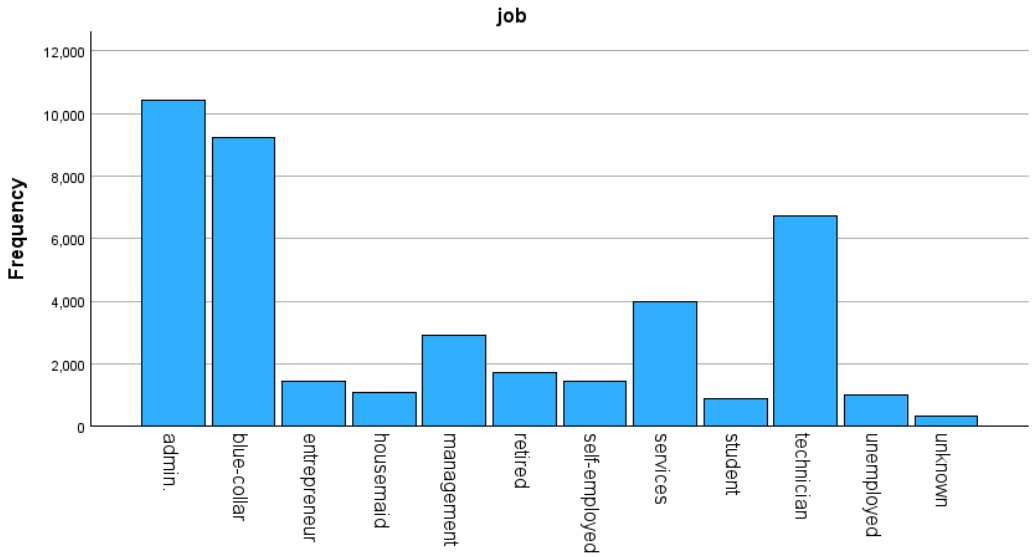

Fig. 2 Distribuția variabilei job
Sursa: prelucrare autori SPSS

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	37338.135[a]	77	<.001
Likelihood Ratio	33462.389	77	<.001
N of Valid Cases	41188		

a. 12 cells (12.5%) have expected count less than 5. The minimum expected count is .14.

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Contingency Coefficient	.690	<.001
N of Valid Cases		41188	

Interpretarea rezultatelor: Chi-Square Test:

- *Valoarea Chi-Square = 15440.126*
- *Grade de libertate (df): 11*
- *Valoarea p: < 0.001*

Deoarece valoarea p este mai mică decât 0.05, respingem ipoteza nulă și concluzionăm că există o asociere semnificativă între "job" și "educație".

Coeficientul de contingență:

Valoarea coeficientului de contingență de 0.690 sugerează o asociere puternică între "job" și "educație".

Decizia: respingem H0, Rezultatele indică faptul că există o asociere semnificativă între cele două variabile.

1.2 COEFICIENTUL DE CORELAȚIE PUNCT BISERIAL (variabilă dihotonică – variabilă cantitativă – distribuită normal)

Există legătură între variabila "marital" şi variabila "cons.conf.idx"

H0 Nu există nicio corelație între statutul marital și indicele confidenței consumatorilor.
H1 Există o corelație semnificativă între statutul marital și indicele confidenței consumatorilor.

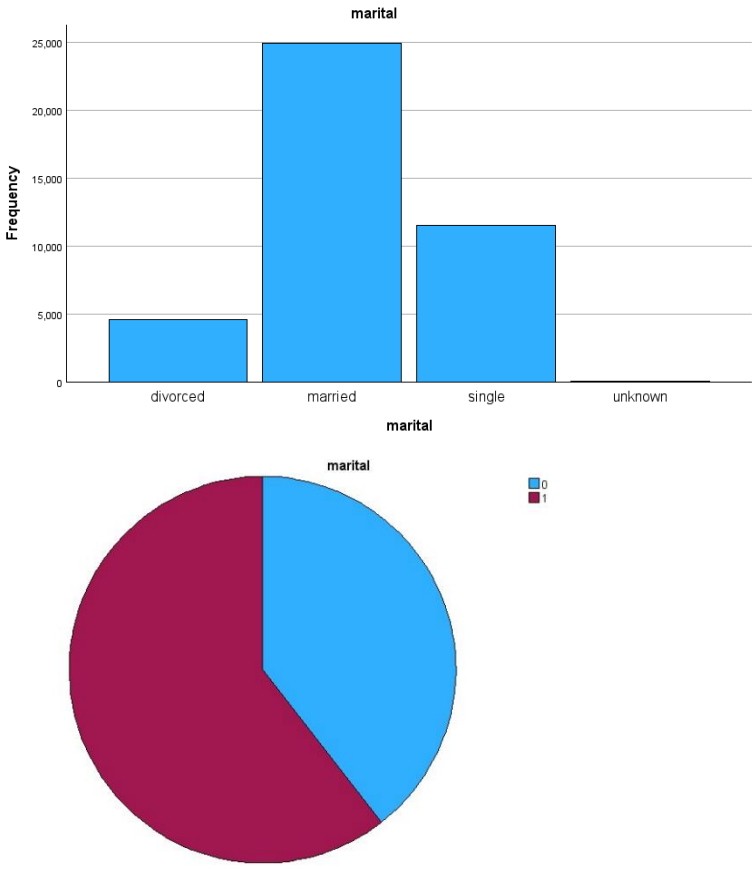

Fig. 3 , 4 Distribuția variabilei marital
Sursa: prelucrare autori SPSS

Correlations

		marital	cons.conf.idx
marital	Pearson Correlation	1	.062**
	Sig. (2-tailed)		<.001
	N	41188	41188
cons.conf.idx	Pearson Correlation	.062**	1
	Sig. (2-tailed)	<.001	
	N	41188	41188

**. Correlation is significant at the 0.01 level (2-tailed).

Tabelul prezintă rezultatele analizei corelației Pearson dintre variabila "marital" (status marital) și variabila "cons.conf.idx" (indice de încredere).

Interpretarea rezultatelor:

Pearson Correlation: Aceasta este valoarea coeficientului de corelație Pearson, care măsoară forța și direcția relației liniare dintre cele două variabile. Pentru marital și cons.conf.idx: Coeficientul de corelație este 0.062, ceea ce indică o corelație pozitivă foarte slabă între statusul marital și indicele de încredere. Sig. (2-tailed): Aceasta este valoarea p asociată coeficientului de corelație. Valoarea p indică probabilitatea ca o corelație de această magnitudine să apară întâmplător într-un eșantion dacă ipoteza nulă (că nu există nicio corelație reală între variabile) este adevărată. Respingem H0.

Pentru marital și cons.conf.idx: Valoarea p este < 0.001, ceea ce înseamnă că această corelație este semnificativă din punct de vedere statistic. Asta sugerează că este foarte puțin probabil ca această corelație să fie rezultatul întâmplării.

Pentru ambele variabile (marital și cons.conf.idx): **numărul de observații este 41,188.**

Coeficientul de corelație de 0.062 indică o corelație foarte slabă între statusul marital și indicele de încredere. Chiar dacă relația este pozitivă (în sensul că, în medie, persoanele căsătorite au un nivel ușor mai mare de încredere comparativ cu cele necăsătorite), această legătură este foarte mică.

Așadar, există o corelație pozitivă foarte slabă (0.062) între statusul marital și indicele de încredere.

Această corelație este semnificativă statistic ($p < 0.001$), ceea ce sugerează că nu este probabil rezultatul întâmplării, dar forța relației este foarte mică.

În termeni practici, chiar dacă relația este semnificativă, coeficientul de corelație mic indică faptul că statusul marital explică foarte puțin din variația în nivelul de încredere.

Notă: coeficientul punct-biserial este echivalent cu coeficientul Pearson atunci când una dintre variabile este dicotomică.

2. TESTE NEPARAMETRICE
2.1. TESTUL MANN-WHITNEY U

Tipuri de variabile: Se folosește pentru a compara două grupuri independente pe o variabilă ordinală sau continuă care nu respectă normalitatea. Varilabilele utilizate „loan” – “cons.price.idx”
H0: Nu există diferențe semnificative în prețurile clienților la serviciile bancare în funcție de istoricul de credit.
H1: Există diferențe semnificative în prețurile clienților la serviciile bancare în funcție de istoricul de credit.

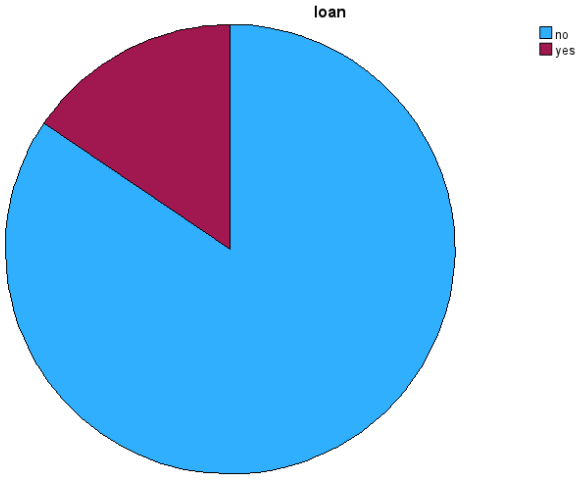

Fig. 5 Proporția variabilei loan
Sursa: prelucrare autori SPSS

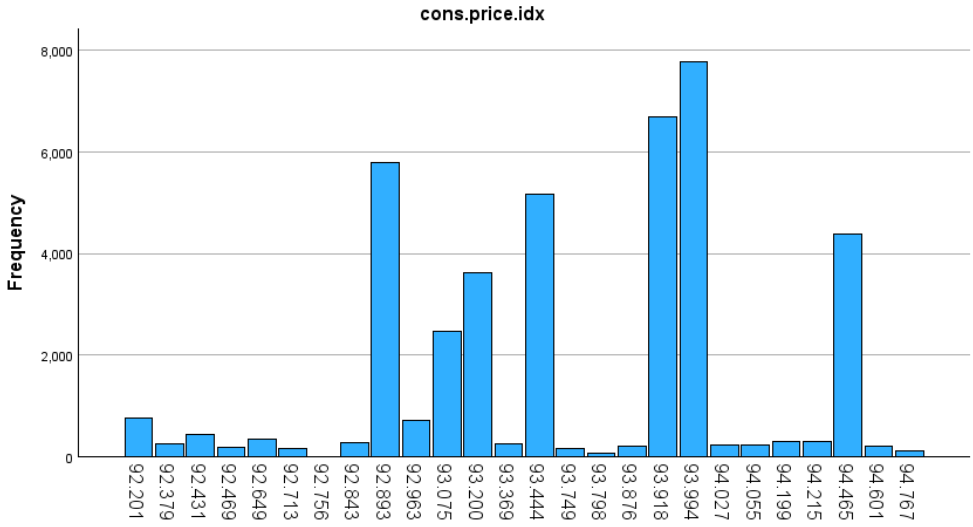

Fig. 6 Distribuția variabilei cons.price.idx
Sursa: prelucrare autori SPSS

Ranks

	loan	N	Mean Rank	Sum of Ranks
cons.price.idx	no	34940	20642.82	721259957.00
	yes	6248	20324.31	126986309.00
	Total	41188		

Test Statistics[a]

	cons.price.idx
Mann-Whitney U	107464433.00
Wilcoxon W	126986309.00
Z	-1.968
Asymp. Sig. (2-tailed)	.049

a. Grouping Variable: loan

Rezultatele testului:

- U-ul lui Mann-Whitney este 107464433.000.
- W-ul lui Wilcoxon este 126986309.000.
- Valoarea Z este -1.968.
- Valoarea asimptotică (2-tailed) este 0.049.

Interpretare:

Valorile negative ale lui Z indică faptul că eșantioanele "yes" și "no" au medii de ranguri diferite.

Valoarea asimptotică (2-tailed) de 0.049 este sub nivelul convențional de semnificație de 0.05, indicând o diferență semnificativă între cele două grupuri în ceea ce privește variabila "cons.price.idx".

În concluzie, respingem H0 la limită, analiza sugerează că există o diferență semnificativă între grupurile "yes" și "no" în ceea ce privește variabila "cons.price.idx".

2.2 TESTUL KOLMOGOROV-SMIRNOV

Pentru a verifica dacă o distribuție a unei variabile se potrivește cu o distribuție teoretică (uniformă).

H0 Distribuția variabilei „month" nu diferă semnificativ de distribuția uniformă.

H1 Distribuția variabilei „month" diferă semnificativ de distribuția uniformă.

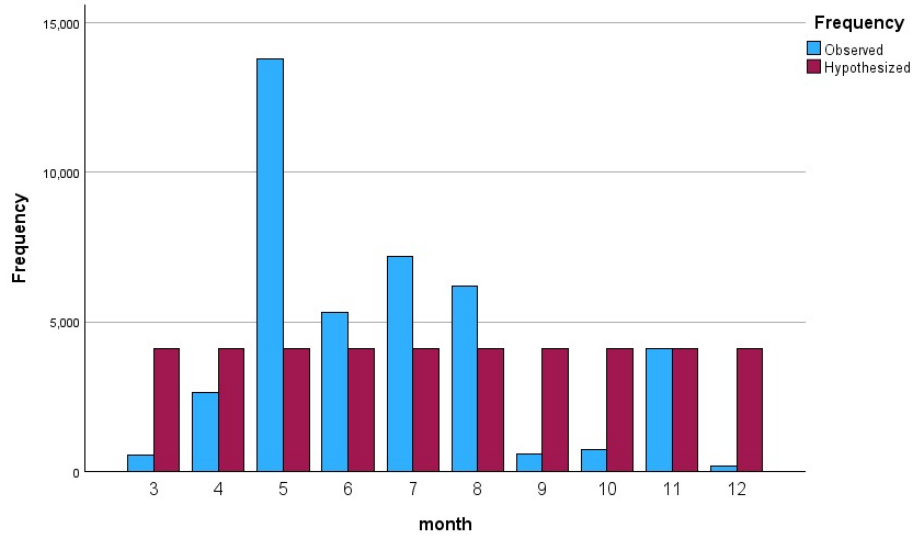

Fig. 7 Distribuția variabilei month și distribuția uniformă
Sursa: prelucrare autori SPSS

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.[a,b]	Decision
1	The categories of month occur with equal probabilities.	One-Sample Chi-Square Test	<.001	Reject the null hypothesis.

a. The significance level is .050.

b. Asymptotic significance is displayed.

One-Sample Chi-Square Test Summary

Total N	41188
Test Statistic	39519.423[a]
Degree Of Freedom	9
Asymptotic Sig.(2-sided test)	<.001

a. There are 0 cells (0%) with expected values less than 5. The minimum expected value is 4118.800.

Rezultatele testului:

- Valoarea asimptotică (2-tailed) este <0,001.

Interpretare:

Valoarea asimptotică (2-tailed) este mai mică decât 0,01; deci este sub nivelul convențional de semnificație de 0.05, indicând o distribuție neuniformă.

În concluzie, respingem H0 și putem afirma că distribuția variabilei „month" diferă semnificativ de distribuția uniformă.

3. REGRESIA LOGISTICĂ

Am folosit o regresie de tip "binary logistic" în care am stabilit variabila dependentă ca fiind "loan", iar variabilele "education", "emp.var.rate", "cons.price.idx", "cons.conf.idx", "euribor3m" și "nr.employed" cele dependente.

După aplicarea metodei "Forward: Conditional" am obținut că doar variabilele "education" și "cons.conf.idx" sunt semnificative statistic (Sig. < 0,05).

Variables not in the Equation

			Score	df	Sig.
Step 2	Variables	emp.var.rate	.983	1	.322
		cons.price.idx	.538	1	.463
		euribor3m	.765	1	.382
		nr.employed	1.593	1	.207
	Overall Statistics		9.220	4	.056

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 2[a]	education	.017	.006	7.850	1	.005	1.018
	cons.conf.idx	-.009	.003	8.566	1	.003	.991
	Constant	-2.153	.127	288.899	1	<.001	.116

a. Variable(s) entered on step 2: education.

Semnificația parametrilor

Pentru variabila "education", avem coeficientul β1=0,017, deci educația influențează pozitiv împrumutul, cu cât un client are o educație cât mai bogată, cu atât poate face un împrumut mai ușor. Cu fiecare nivel de educație în plus, o persoană poate să își crească șansele de a face un împrumut de 1,018 ori (cu 0,018%).

Variabila "cons.conf.idx" are β2=-0,009, adică va influența negativ împrumutul. Astfel, cu fiecare mărire cu o unitate a indicelui confidenței consumatorilor șansele de a face un împrumut scad de 0,991 ori (cu 0,009%).

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 2	Step	7.875	1	.005
	Block	15.596	2	<.001
	Model	15.596	2	<.001

Validarea modelului

Ipotezele testului sunt :

▪ H0: modelul nu este valid

▪ H1: modelul este valid

Sig=0,005 deci se respinge la limită H0, modelul poate fi validat cu un nivel de semnificație de 0,005.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
2	35046.713[a]	.000	.001

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Calitatea modelului

Toți acești coeficienți sunt mai mici decât 0,3, ceea ce indică un model de o calitate slabă.

Classification Table[a]

			Predicted		
			loan		Percentage Correct
	Observed		no	yes	
Step 2	loan	no	34940	0	100.0
		yes	6248	0	.0
	Overall Percentage				84.8

a. The cut value is .500

Procentul previziunilor corecte

Procentul total de corectitudine de 84.8% poate părea inițial ridicat, dar nu oferă o imagine completă. De fapt, această valoare este înșelătoare din cauza distribuției inegale a datelor și a performanței modelului pentru fiecare categorie.

Pentru clasa loan = no (fără împrumut), modelul a clasificat corect toate cele 34,940 cazuri. Acest lucru înseamnă că modelul este foarte bun la identificarea cazurilor fără împrumut.

Pentru clasa loan = yes (cu împrumut), modelul nu a reușit să identifice corect niciun caz. Toate cele 6,248 cazuri au fost clasificate greșit ca no.

Concluzii

Prin aplicarea tehnicilor de statistică neparametrică pe acest set de date complex, am obținut o înțelegere profundă a factorilor care influențează eficiența campaniilor de marketing telefonic. Rezultatele acestui proiect vor contribui la îmbunătățirea strategiilor de marketing și la optimizarea resurselor bancare pentru obținerea unor rezultate mai bune în viitor.

Bibliografie

Boboc C., Suport de curs și seminar Statistică neparametrică, București, 2024

Site-uri web:
https://guides.lib.uoguelph.ca/c.php?g=525348&p=5286104#s-lg-box-16636558