ROC krivulja - krivulja odnosa specifičnosti i osjetljivosti klasifikatora

Tokić, Lucija

Master's thesis / Diplomski rad

2017

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Science / Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:217:320917

Rights / Prava: In copyright / Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2024-03-29**

Repository / Repozitorij:

Repository of the Faculty of Science - University of Zagreb

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Lucija Tokić

ROC KRIVULJA - KRIVULJA ODNOSA SPECIFIČNOSTI I OSJETLJIVOSTI KLASIFIKATORA

Diplomski rad

Voditelj rada:
prof. dr. sc. Anamarija Jazbec

Zagreb, Rujan 2017.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Zahvaljujem se Anti, mami, tati i Sanji na svoj podršci, razumijevanju i ljubavi koju ste mi pružili. Bez vas ne bi bila tu i to što jesam.

Sadržaj

Sadržaj **iv**

Uvod **2**

1 Procjena binarnih klasifikatora **3**
1.1 Procjena dijagnostičkih testova modelom praga 3
1.2 Mjerenje točnosti dijagnostičkih testova 4
1.3 Metode računanja prediktivnih vrijednosti testa 5

2 Receiver operating characteristic (ROC) krivulja **13**
2.1 Konveksni ROC mnogokut 17
2.2 Površina ispod ROC krivulje 21

3 Primjeri **26**
3.1 Testiranje medicinske opreme 26
3.2 Odabiranje proizvoda u prodajni asortiman 46

Bibliografija **56**

Popis engleskog nazivlja iz SAS-a

Analysis of Maximum Likelihood Estimates - analiza procjene metodom maksimalne vjerodostojnosti
Analysis Variable - analiza varijable
Area - površina
Col Pct - relativna frekvencija stupca
Confidence Limits - pouzdane granice
COUNT - frekvencija
DF - broj stupnjeva slobode
Estimate - procjena (procijenjeni koeficijent)
Frequency - frekvencija
Intercept - slobodni član
Mean - aritmetička sredina
N Obs - broj observacija
Parameter - parametar
Percent - relativna frekvencija
Pr > ChiSq - p-vrijednost χ^2 testa
Row Pct - relativna frekvencija retka
Standard Error - standardna pogreška
Std Dev - standardna devijacija
Variable - varijabla

Uvod

Ocjena prediktivne točnosti je kritička procjena kvalitete prediktivnog modela kao i usporedbe prediktivnih modela, algoritama i tehnologija. U medicini su ROC krivulje postale glavni alat za ovu svrhu te se njihovo korištenje velikom brzinom širi u području financija, znanosti o atmosferi i strojnom učenju. ROC krivulja daje vizualno i opsežno izvješće o točnosti predikcije. One su naširoko primjenjive, neovisno o izvoru predikcija.

Predviđanje određenog ponašanja postao je ključan dio u poslovnom i znanstvenom svijetu. Jednostavan svakodnevni primjer toga je vremenska prognoza. Svi izvori vijesti poput dnevnih novina, radija, televizije i interneta, iznose detaljnu vremensku prognozu. Ona je veoma bitna za mnoge životne aktivnosti kao što su prijevoz, poljoprivreda i građevina. Zbog toga se razvila velika industrija koja se bavi prikupljanjem podataka koji pomažu prilikom predviđanja vremenskih uvjeta i izradi statističkih modela za prognoziranje koristeći te podatke.

Zbog ubrzanog povećanja populacije, jako bitno je postalo predvidjeti potrošnju električne energije, plina, nafte i hrane radi planirane proizvodnje.

U ovom radu bazirat ću se na primjerima iz medicine, stoga ću pojmove tumačiti kroz to područje.

U zadnjem poglavlju iznosim dva primjera napisana u statističkom paketu SAS. U prvom primjeru ću pokazati kako se konstruiraju ROC krivulje. U drugom primjeru pokazujem kako usporedbom ROC krivulja možemo odabrati najbolji logistički model koji opisuje dane podatke.

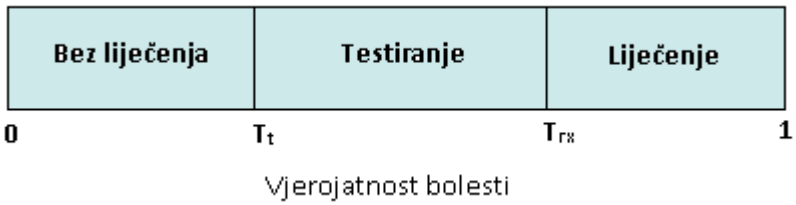
Poglavlje 1

Procjena binarnih klasifikatora

1.1 Procjena dijagnostičkih testova modelom praga

U medicinskoj literaturi riječ "predikcija" ne pojavljuje se često, već se koristi termin dijagnoze. Dijagnoza je predikcija bolesti kod pacijenta koji ima simptome određene bolesti.[2]

Model praga (*eng. Threshold*) je matematički model u kojem se vrijednost praga ili niz vrijednosti pragova koristi za rangiranje vrijednosti gdje se ponašanje predviđeno modelom bitno razlikuje s obzirom u koje područje upada. Model praga je ilustriran na slici 1.1.



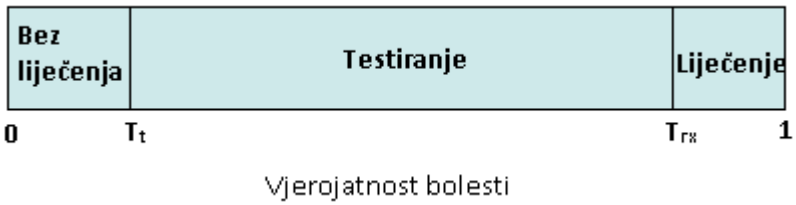
Slika 1.1: Model praga

Pretpostavimo da imamo binarni klasifikator, tj. predikciju bolesti. Izradimo klasifikacijski model $h_\theta(x)$. Ako $h_\theta(x) = 1$, tada predviđamo da je bolest prisutna, ako $h_\theta(x) = 0$, tada predviđamo da je bolest nije prisutna.

Liječnikova procjena da pacijent boluje od bolesti, bez obavljanja dijagnostičkog testa, zove se vjerojatnost bolesti. Kako vjerojatnost poprima vrijednosti između 0 i 1, ilustriramo sve na pravcu između točaka 0 i 1. Točka T_t predstavlja testni prag (*eng. testing threshold*), tj. točku na vjerojatnosnom pravcu u kojoj ne postoji razlika između vrijednosti da se pacijent ne liječi i da se provede dijagnostički test. Slično, prag liječenja T_{rx} (*eng. treatment threshold*) je točka na vjerojatnosnom pravcu u kojoj ne postoji razlika

između vrijednosti da se provede dijagnostički test i da se pacijent liječi bez provođenja testa. O nekoliko faktora ovisi gdje ćemo postaviti točke praga: rizik od dijagnostičkih testova (pretraga), korist od liječenja bolesnog pacijenta, rizik od liječenja bolesnog, ali i zdravog pacijenta te točnost testa. [1]

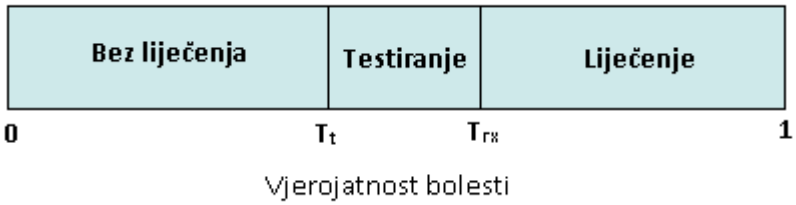
Slika 1.2 prikazuje situaciju u kojoj je test poprilično točan i postoji jako mali rizik za pacijenta.



Slika 1.2: Precizan ili niskorizičan test

U ovom slučaju, liječnik će raditi pretrage kada je vjerojatnost bolesti niska, ali i kada je ona visoka.

Slika 1.3 prikazuje suprotnu situaciju u kojoj test ima malu točnost ili postoji jako velik rizik za pacijenta. Ovakav test se obično ne provodi.



Slika 1.3: Netočan ili visokorizičan test

1.2 Mjerenje točnosti dijagnostičkih testova

Točnost dijagnostičkog testa ima dvije značajke:

1. Test dobro detektira bolest koja se testira, dakle daje pozitivan rezultat kod bolesnih pacijenata. To svojstvo se zove **senzitivnost testa**. Ako test ima visoku senzitivnost, tada ima nisku lažno negativnu stopu (*eng. false-negative rate*), tj. test neće pokazati

lažni negativni rezultat kod mnogo oboljelih pacijenata.
Senzitivnost se definira na nekoliko ekvivalentnih načina:

- vjerojatnost da je test pozitivan kod oboljelih pacijenata
- postotak oboljelih pacijenata kojima je test pozitivan
- stvarno pozitivna stopa (*eng. true-positive rate*)

2. Test dobro prepoznaje pacijente koji nisu oboljeli. To svojstvo se zove **specifičnost testa**. Ako je specifičnost testa visoka, test ima nisku lažno pozitivnu stopu (*eng. false-positive rate*), tj. test neće pokazati lažni pozitivni rezultat u mnogo neoboljelih pacijenata. Specifičnost se također može definirati na nekoliko ekvivalentnih načina:

   - vjerojatnost da je test negativan kod neoboljelih pacijenata
   - postotak neoboljelih pacijenata kojima je test negativan
   - 1 minus lažno pozitivna stopa

Senzitivnost i specifičnost dijagnostičkog testa određeni su podjelom testiranih pacijenata na dvije grupe (binarna klasifikacija): oni pacijenti za koje se zna da boluju (ili imaju dijagnozu) te one za koje se zna da nisu bolesni (nemaju dijagnozu). Senzitivnost se tada dobije kao postotak pacijenata koji su bolesni i kojima je test pozitivan; specifičnost je postotak pacijenata koji nisu oboljeli i kojima je test negativan.[1]

1.3 Metode računanja prediktivnih vrijednosti testa

Senzitivnost i specifičnost, same za sebe, ne mogu se koristiti u određivanju vrijednosti dijagnostičkog testa u pojedinog pacijenta, već zajedno s liječnikovim indeksom sumnje da pacijent boluje (prethodnom vjerojatnošću bolesti) i rezultatima testa se određuje vjerojatnost bolesti. U nastavku iznosim četiri različite metode za računanje prediktivnih vrijednosti.

Metoda 2×2 tablice

57-godišnji čovjek ima konstantnu neugodnu bol u donjem dijelu leđa. Inače je dobrog zdravlja i nema prijašnjih bolova u donjem dijelu leđa. S obzirom na simptome, povijest bolesti i fizički pregled liječnik je ustvrdio da pacijent ima 20-30% šanse da boluje od raka kralježnice. Liječnik mora odlučiti hoće li naručiti pacijenta na ESR test (*eng. erythrocyte sedimentation rate*) za kontrolu brzine sedimentacije eritrocita ili na magnentsku rezonancu (MR) kralježnice. Senzitivnost i specifičnost kada je ESR ≤ 20 mm/h su 78% i 67%, a za MR kralježnice su senzitivnost i specifičnost jednake 95%. Odluka koji test će liječnik izabrati ovisi o tri ključne informacije:

1. vjerojatnosti da pacijent boluje od raka kralježnice prije obavljanja ikakvih dijagnostičkih testova (indeks sumnje)

2. preciznosti ESR testa da prepozna pacijente koji boluju od raka kralježnice od onih kojima je pokazao da boluju (senzitivnost)

3. frekvenciji negativnih ishoda testa u pacijenata koji ne boluju od raka kralježnice (specifičnost).

Liječnik na temelju godina i povijesti bolesti uzme da je indeks sumnje 20%, tj. vjerojatnost da pacijent boluje od raka kralježnice je 20%. Nakon što uvažimo veličine senzitivnosti i specifičnosti ESR testa će se prethodna vjerojatnost od 20% promijeniti. Te novodobivene vjerojatnosti se nazivaju **prediktivna vrijednost pozitivnog testa** ako je rezultat ESR-a pozitivan ili **prediktivna vrijednost negativnog testa** ako je rezultat ESR-a negativan. Zajedno čine nove vjerojatnosti (*eng. posterior probabilities*).

Tablica 1.1 prikazuje kontingencijsku tablicu u kojoj su dane formule za računanje senzitivnosti, specifičnosti, pozitivne i negativne prediktivne vrijednosti testa.

Tablica 1.1: Primjer kontingencijske 2×2 tablice

		Stvarno stanje bolesti		
		Pozitivan (D^+)	**Negativan** (D^-)	Mjere
Dijagnostički test	**Pozitivan** (T^+)	Stvarno pozitivan **TP**	Lažno pozitivan **FP**	Pozitivna prediktivna vrijednost **PPV**: $\frac{TP}{TP+FP}$
	Negativan (T^-)	Lažno negativan **FN**	Stvarno negativan **TN**	Negativna prediktivna vrijednost **PNV**: $\frac{TN}{TN+FN}$
	Mjere	**Senzitivnost** $\frac{TP}{TP+FN}$	**Specifičnost** $\frac{TN}{TN+FP}$	**Točnost** $\frac{TP+TN}{TP+FP+TN+FN}$

Prvi korak u metodi s 2×2 tablicom je određivanje prediktivnih vrijednosti. Umjesto s postotcima, radit ću s cijelim brojevima, tj. 20% šanse da pacijent ima rak kralježnice znači da 200 pacijenata od 1000 koji su kao navedeni pacijent će imati rak kralježnice. U tablici 1.2 broj 200 je u zadnjem retku stupca D^+. Analogno je broj 800 zapisan u zadnjem retku stupca D^-.

Drugi korak je popuniti ćelije tablice koristeći senzitivnost i specifičnost testa. Tablica 1.2 pokazuje da stvarno pozitivna stopa (senzitivnost) odgovara ćeliji T^+D^+ koja je označena s TP. Kako je senzitivnost = 78%, znači da 78% pacijenata od 200, tj. 156 pacijenata će imati pozitivan ESR test, stoga su oni stvarno pozitivni (true-positive). $200 - 156 =$ 44 pacijenta su stvarno negativni (false-negative). Slično zaključujemo da test sa specifičnošću od 67% rezultira s 536 stvarno negativnih (true-negative) pacijenata, na rak

kralježnice, od njih 800 i 800 – 536 = 264 lažno pozitivna (false-positive) pacijenta na rak kralježnice.

Treći je korak dopuniti rubove tablice. Iz prvog retka se vidi da 156 + 264 = 420 ljudi poput ovog pacijenta bi imalo pozitivan ESR test. Slično, 580 pacijenata bi imalo negativan ESR test.

Tablica 1.2: Kontingencijska 2×2 tablica pacijenata koji su bili izloženi ESR testu

ESR test	**Bolest**		
	DA (D^+)	NE (D^-)	Ukupno
Pozitivan (T^+)	**(TP)** 156	**(FP)** 264	420
Negativan (T^-)	**(FN)** 44	**(TN)** 536	580
Ukupno	200	800	1000

U četvrtom koraku računamo prediktivne vrijednosti. Od 420 ljudi s pozitivnim testom njih 156 zaista boluje od raka kralježnice, što je $156/420 = 37\%$. Slično, 536 od 580 pacijenata s negativnim testom, tj. 92% ne boluje od raka kralježnice. Postotak od 37% naziva se prediktivna vrijednost pozitivnog testa, s oznakom PV^+, i označava postotak pacijenata s pozitivnim rezultatom testa koji boluju od raka kralježnice (ili vjerojatnost raka kralježnice uz dani pozitivan test ESR-a). Postotak od 92% je prediktivna vrijednost negativnog testa, s oznakom PV^-, i označava postotak pacijenata s negativnim rezultatom testa koji ne boluju od raka kralježnice (ili vjerojatnost da pacijent ne boluje od raka kralježnice uz dani negativan test ESR-a). Ostale dvije vjerojatnosti se također izračunaju iz tablice te nemaju specifična imena: $264/420 = 0.63$ je vjerojatnost da pacijent ne boluje od raka, iako je test pozitivan; i $44/580 = 0.08$ je vjerojatnost da pacijent boluje od raka, iako je test negativan. Vidi tablicu 1.3.

Tablica 1.3: Prediktivne vrijednosti ESR testa

Prediktivna vrijednost pozitivnog ESR testa	$PV^+ = TP/(TP + FP) = 156/420 = 0.371$
Prediktivna vrijednost negativnog ESR testa	$PV^- = TN/(TN + FN) = 536/580 = 0.924$

Dakle, ESR test je umjereno senzitivan i specifičan za otkrivanje raka kralježnice kada je indeks sumnje nizak. Povećava vjerojatnost otkrivanja raka sa 20% na 37% kada je test pozitivan, i povećava vjerojatnost otkrivanja nepostojanja raka sa 67% na 92% kada je test negativan. Općenito, testovi s visokom senzitivnošću su korisni u isključivanju bolesti kod pacijenata s negativnim testom.

Sada pretpostavimo da je test ESR kod spomenutog pacijenta pozitivan. Iz tablice 1.2 znamo da je vjerojatnost raka 37%. Sada se provodi idući dijagnostički test, MR kra-

lježnice. Tad 37% postaje nova prethodna vjerojatnost što znači da 370 ljudi od 1000 je izgledno da imaju rak, a 630 njih da nema. Ovi brojevi su zbroj 1. i 2.stupca u tablici 1.4. Kako MR kralježnice ima senzitivnost i specifičnost od 95%, slijedi da je $0.95\times370 = 351.5$ stvarno pozitivnih pacijenata i $0.95\times630 = 598.5$ stvarno negativnih pacijenata. Iz toga slijedi da je 18.5 njih lažno negativnih i 31.5 lažno pozitivnih. Nakon zbrajanja redaka, prediktivna vrijednost pozitivnog MR-a kralježnice je $351.5/383 = 91.8\%$, a prediktivna vrijednost negativnog MR-a kralježnice je $598.5/617 = 97\%$ (vidi tablicu 1.5).

Tablica 1.4: Kontingencijska 2×2 tablica pacijenata koji su bili izloženi MR testu

MR test	**Bolest**		
	DA (D^+)	NE (D^-)	Ukupno
Pozitivan (T^+)	**(TP)** 351.5	**(FP)** 31.5	383
Negativan (T^-)	**(FN)** 18.5	**(TN)** 598.5	617
Ukupno	370	630	1000

Tablica 1.5: Prediktivne vrijednosti MR testa

Prediktivna vrijednost pozitivnog MR testa	$PV^+ = TP/(TP+FP) = 351.5/383 = 0.918$
Prediktivna vrijednost negativnog MR testa	$PV^- = TN/(TN+FN) = 598.5/617 = 0.970$

Primjer je uzet iz knjige "Basic and Clinical Biostatistics" autora B.Dawson i R.G.Trapp i na njega ću koristiti u objašnjavanju daljnih metoda.[1]

Omjer vjerodostojnosti

Omjer vjerodostojnosti koristi senzitivnost i specifičnost testa pri određivanju mijenja li rezultat testa vjerojatnost bolesti. Pozitivan i negativan test imaju zasebne omjere vjerodostojnosti. Kada je rezultat testa pozitivan, omjer vjerodostojnosti se računa kao omjer senzitivnosti i lažno pozitivne stope:

$$LR^+ = \frac{\text{senzitivnost}}{\text{lažno-pozitivna stopa}} = \frac{\text{senzitivnost}}{\text{1-specifičnost}}$$

Da bismo dobili poslijetestnu šansu bolesti, omjer vjerodostojnosti se pomnoži s prijetestnom šansom bolesti:

$$\text{Prijetestna šansa} \times \text{Omjer vjerodostojnosti} = \text{Poslijetestna šansa}$$

Iz primjera znamo da je senzitivnost ESR testa za dijagnosticiranje raka kralježnice 78%, specifičnost je 67% pa slijedi da je lažno pozitivna stopa 100% − 67% = 33%. Omjer vjerodostojnosti pozitivnog testa je

$$LR^{+} = \frac{0.78}{0.33} = 2.36$$

Da bismo koristili omjer vjerodostojnosti, moramo prethodnu vjerojatnost pretvoriti u prethodnu šansu. Šansa je omjer broja pozitivnih ishoda pokusa i broja negativnih ishoda pokusa. Prethodna vjerojatnost za rak kralježnice iznosi 0.20, a šanse se dobiju tako da se vjerojatnost podijeli s vrijednošću 1-vjerojatnost te dobivamo:

$$\text{Prijetestna šansa} = \frac{\text{Prijetestna vjerojatnost}}{\text{1-Prijetestna vjerojatnost}} = \frac{0.20}{0.80} = 0.25$$

Sada imamo

$$\text{Poslijetestna šansa} = 0.25 \times 2.36 = 0.59$$

Dobivenu poslijetestnu šansu možemo pretvoriti nazad u poslijetestnu vjerojatnost:

$$\text{Poslijetestna vjerojatnost} = \frac{\text{Poslijetestna šansa}}{\text{1+poslijetestna šansa}} = \frac{0.59}{1 + 0.59} = 0.37$$

Poslijetestna vjerojatnost je prediktivna vrijednost pozitivnog testa i rezultat je isti kao što smo dobili i ranije.

Negativan omjer vjerodostojnosti LR^{-} može se koristiti da se izračuna šansa bolesti čak i ako je rezultat testa negativan. Dobije se kao omjer lažno negativne stope i stvarno negativne stope:

$$LR^{-} = \frac{\text{lažno-negativna stopa}}{\text{stvarno-negativna stopa}}$$

Pretpostavimo da je ESR test negativan. Zanima nas kolika je vjerojatnost raka kralježnice.

$$LR^{-} = \frac{1 - 0.78}{0.67} = 0.328$$

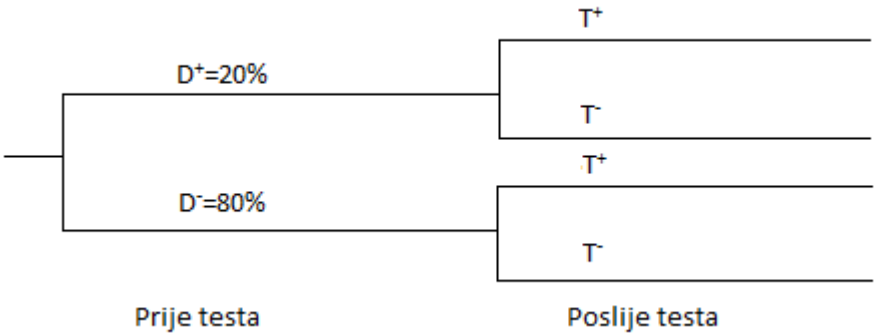
Množeći prijetestnu šansu bolesti s LR^{-} dobivamo $0.328 \times 0.25 = 0.082$ poslijetestnu šansu bolesti s negativnim testom. Opet možemo pretvoriti šansu u vjerojatnost

$$\text{Poslijetestna vjerojatnost} = \frac{0.082}{1 + 0.082} = 0.076 = 7.6\%$$

To znači da osoba, kao pacijent iz primjera, s indeksom sumnje od 20% ima poslijetestnu vjerojatnost za rak kralježnice od 7.6% čak i kada je test negativan. Ovaj rezultat je konzistentan s prediktivnom vrijednošću negativnog testa od 92.4% pa se LR^{-} može interpretirati kao $1 - PV^{-}$.[1]

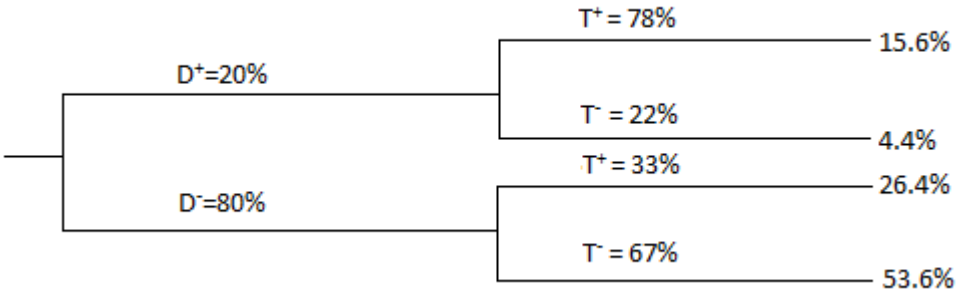
Metoda stablovnog odlučivanja

Kroz primjer raka kralježnice prikazat ću metodu stablovnog odlučivanja za dobivanje prediktivne vrijednosti bolesti iz početne vjerojatnosti bolesti od 20%. Slika 1.4 prikazuje da se prije obavljanja pretraga pacijent može naći u jednoj od dvije situacije: biti bolestan ili ne biti bolestan. Ta dva stanja su prikazana na stablu, s oznakom D^+ ako bolest postoji ili D^- ako bolest nije prisutna. Prethodne vjerojatnosti su prikazane na svakoj grani stabla, 20% na D^+ grani i 100% − 20% = 80% na D^- grani. Rezultat testa može biti pozitivan ili negativan, bez obzira na pacijentovo pravo stanje. Ako je test pozitivan granu označimo s T^+, negativan test označimo s T^-. U sljedećem koraku stablu dodajemo senzitivnost i



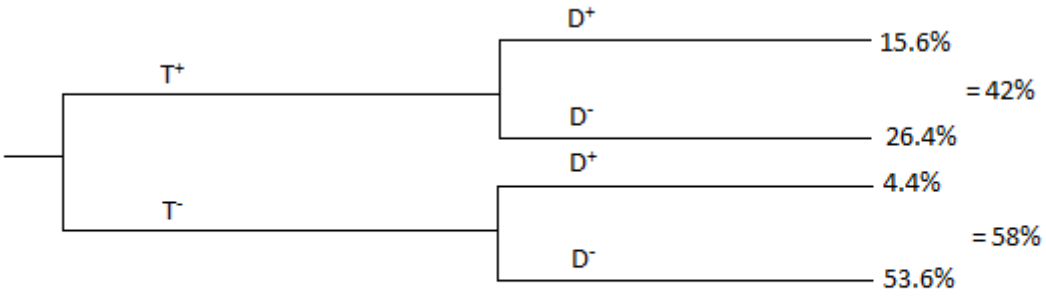
Slika 1.4: Stablo odlučivanja s granama rezultata testa

specifičnost. Gledajući D^+ granu, test ESR je pozitivan u otprilike 78% slučajeva. Slika 1.5 prikazuje senzitivnost testa od 78% napisano na T^+ grani. U 100% − 78% = 22% slučaja je test negativan, što piše na grani T^-. Ovi postotci se množe da se dobiju vjerojatnosti na kraju svake grane. Dakle, 78% ljudi od 20% ljudi s rakom kralježnice, tj. (20%)(78%) = 15.6% je napisano na kraju grane D^+T^+; (20%)(22%) = 4.4% ljudi ima rak kralježnice i negativan test - kraj D^+T^- grane. Sličan izračun se napravi za 80% ljudi koji nemaju rak kralježnice te su postotci prikazani na slici 1.5. Uočimo da je zbroj postotaka na kraju svake grane jednak 100%. Time je stablo odlučivanja ispunjeno. Međutim, stablo se može iskoristiti u pronalaženju prediktivnih vrijednosti testa tako što se stablo obrne. Na slici 1.6 je stablo sa slike 1.5 obrnuto te prve dvije grane predstavljaju rezultate testa na koje se vežu grane o postojanju bolesti D^+ i o izostanku bolesti D^-. Vrijednosti na kraju grana su prepisane sa slike 1.5, tj. 15.6% ljudi koji odgovaraju D^+T^+ su napisani na kraju T^+D^+ grane itd. Zbroj dva rezultata povezana s pozitivnim testom iznosi 42%. Sada se mogu odrediti prediktivne vrijednosti. Ako pacijent ima pozitivan ESR test, obrnuta vjerojatnost pozitivnog testa, tj. prediktivna vrijednost pozitivnog testa se dobije dijeljenjem 15.6% s



Slika 1.5: Stablo odlučivanja s rezultatima testa

42% i dobijemo 37.1%, što je dobiveno u 2×2 metodi. Ako pacijent ima pozitivan ESR test, vjerojatnost da pacijent nije bolestan je $26.4\%/42\% = 62.9\%$. Ako pacijent ima negativan ESR test, vjerojatnost da je pacijent bolestan je $4.4\%/58\% = 7.6\%$. Ako pacijent ima negativan ESR test, prediktivna vrijednost negativnog testa je $53.6\%/58\% = 92.4\%$.[1]



Slika 1.6: Obrnuto stablo

Bayesov teorem

Zadnja metoda koju ću navesti za računanje prediktivnih vrijednosti pozitivnog testa je pomoću formule Bayesova teorema:

$$P(D^+|T^+) = \frac{P(T^+|D^+)P(D^+)}{P(T^+|D^+)P(D^+) + P(T^+|D^-)P(D^-)} \tag{1.1}$$

Dakle, Bayesov teorem govori kolika je vjerojatnost bolesti uz pozitivan rezultat testa. Prva vjerojatnost brojnika $P(T^+|D^+)$ je vjerojatnost da je test pozitivan uz uvjet da je bolest prisutna; druga vjerojatnost $P(D^+)$ označava prethodnu vjerojatnost da je pacijent bolestan (prije testiranja). Nazivnik daje vjerojatnost da je test bolestan bez obzira je li bolest prisutna $P(T^+|D^+)$ ili ne $P(T^+|D^-)$, pomnožene s prethodnom vjerojatnošću ishoda bolesti. Vjerojatnost $P(T^+|D^-)$ je zapravo lažno-negativna stopa, a $P(D^-)$ je $1 - prethodna vjerojatnost$, stoga gornju formulu možemo napisati kao:

$$P(D^+|T^+) = \frac{\text{senzitivnost} \times \text{prethodna vjerojatnost}}{\text{senzitivnost} \times \text{prethodna vjeroj.} + (\text{lažno-pozitivna stopa}) \times (\text{1-prethodna vjeroj.})}$$

Ponovno ću preko istog primjera ilustrirati metodu. Prethodna vjerojatnost raka kralježnice je 0.20, a senzitivnost i specifičnost ESR testa za rak kralježnice su 78% i 67% respektivno. Kada uvrstimo u formulu

$$P(D^+|T^+) = \frac{0.78 \times 0.20}{(0.78 \times 0.20) + (1 - 0.67) \times (1 - 0.20)} = \frac{0.156}{0.156 + 0.264} = 0.37 \tag{1.2}$$

Ovaj rezultat je jednak kao i u metodi stablovnog odlučivanja i 2×2 tablici.

Slična formula se može izvesti za računanje prediktivne vjerojatnosti negativnog testa:

$$P(D^-|T^-) = \frac{P(T^-|D^-)P(D^-)}{P(T^-|D^-)P(D^-) + P(T^-|D^+)P(D^+)} \tag{1.3}$$

Kada uvrstimo u formulu

$$P(D^-|T^-) = \frac{0.67 \times 0.80}{(0.67 \times 0.80) + (1 - 0.78) \times 0.20} = \frac{0.536}{0.58} = 0.924 \tag{1.4}$$

Ovaj rezultat je jednak kao i u metodi stablovnog odlučivanja i 2×2 tablici. [1]

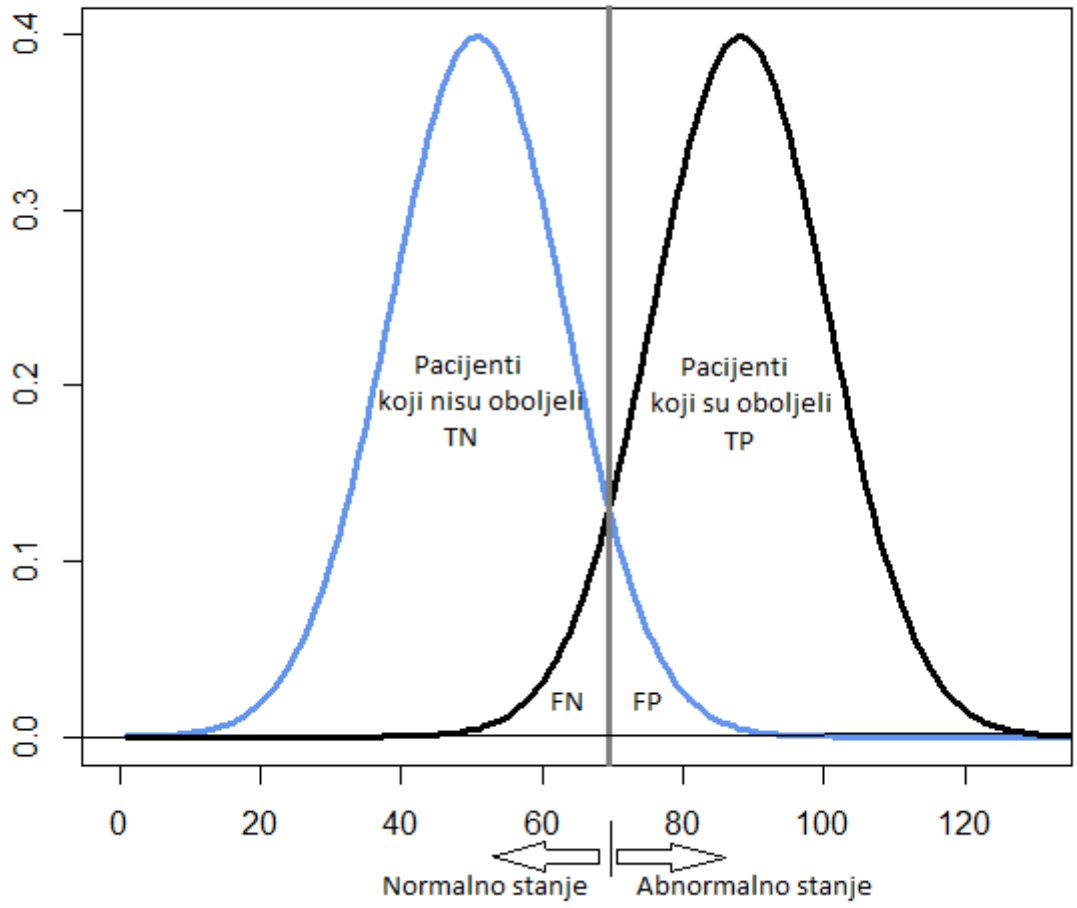
Poglavlje 2

Receiver operating characteristic (ROC) krivulja

Opisane metode računaju vjerojatnost bolesti koristeći prethodnu vjerojatnost bolesti (*eng. prior probability*) i rezultat dijagnostičkog testa koji može biti samo pozitivan ili negativan. Međutim, mnogi testovi za rezultate imaju numeričke vrijednosti. U tom slučaju, senzitivnost i specifičnost ovise o tome gdje će se postaviti granična vrijednost testa (*eng. cutoff*), tj. prag između pozitivnog i negativnog rezultata testa. Takva situacija se može ilustrirati s dvije gaussove distribucije s vrijednostima dijagnostičkog testa: jedna je distribucija za ljude koji su oboljeli, a druga za ljude koji nisu oboljeli (zdravi). Slika 2.1 prikazuje dvije hipotetske distribucije koje odgovaraju ovoj situaciji u kojoj je prosječna vrijednost testa kod oboljelih pacijenata jednaka 90, a prosječna vrijednost onih koji nisu oboljeli je 50. Ako je granična vrijednost testa smještena na vrijednosti 70, oko 10% zdravih ljudi su netočno svrstani među one koji su oboljeli (lažno-pozitivni) zato što im je vrijednost testa veća od 70. Otprilike 10% oboljelih ljudi je netočno svrstano među zdrave jer im je vrijednost testa manja od 70. Drugim riječima, ovaj test ima senzitivnost od 90% te specifičnost od 90%.

Recimo da liječnik želi test koji ima veću senzitivnost, tj. liječnik će radije imati više lažno pozitivnih pacijenata nego propustiti ljude koji su zaista bolesni. Slika 2.2 pokazuje što se događa kada je senzitivnost povećana pomicanjem granične vrijednosti testa na 60. Senzitivnost je povećana, ali tada se specifičnost smanji.

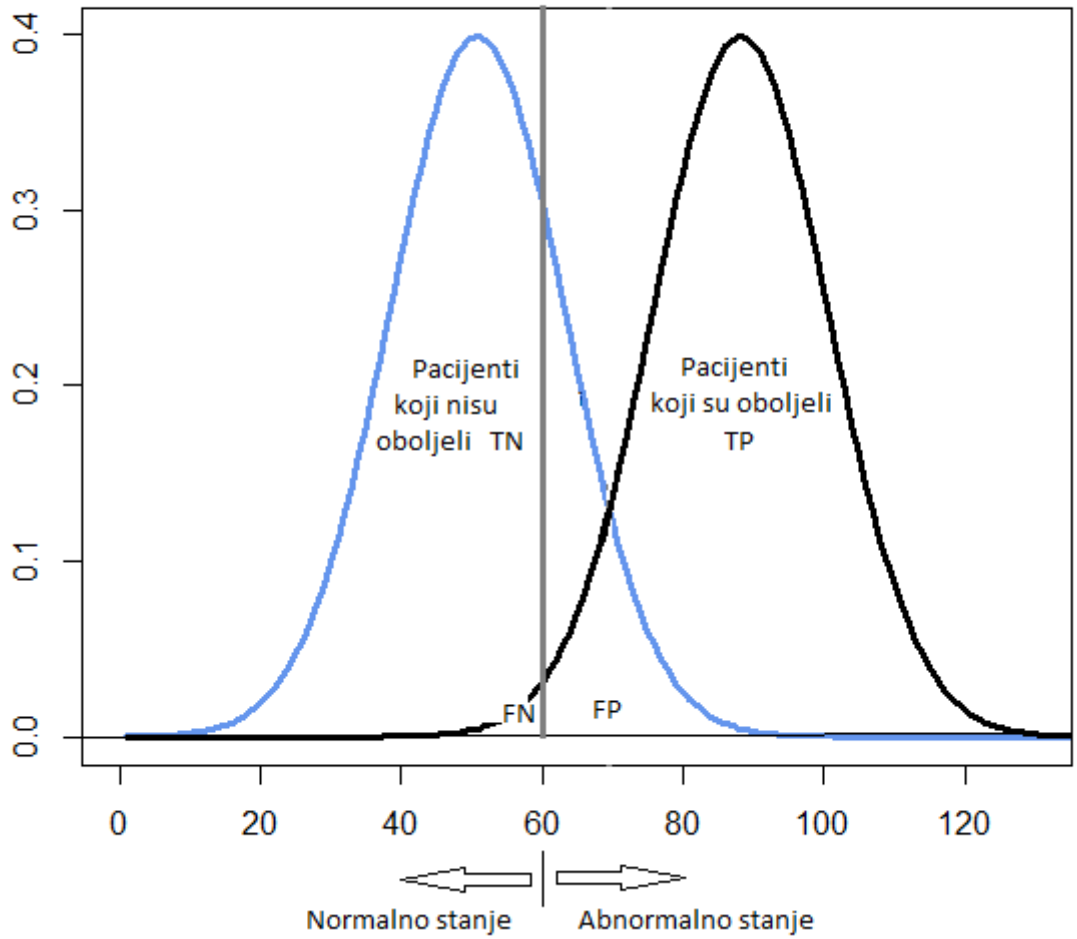
Efektivniji način prikazivanja odnosa senzitivnosti i specifičnosti u ovisnosti o mijenjanju granična vrijednost testa vrijednosti kod testova s numeričkim rezultatima je preko krivulja osjetljivosti (*eng. receiver operating characteristic, ROC curves*), u nastavku ROC krivulja. ROC krivulja je graf na kojem je prikazana senzitivnost (stvarno pozitivna stopa) u odnosu na lažno pozitivnu stopu (1-specifičnost). ROC krivulje se mogu crtati i tako da se prikazuje senzitivnost u odnosu na specifičnost, tada su one zrcalno okrenute u odnosu



Slika 2.1: Dvije hipotetske distribucije s graničnom vrijednosti testa u točki 70. Stvarno pozitivni (TP), lažno negativni (FN), stvarno negativni (TN), lažno pozitivni (FP). Senzitivnost = 90%, specifičnost = 90%.

na standardne. Uvedimo par oznaka:

- Senzitivnost, $tpr = \frac{TP}{TP+FP}$

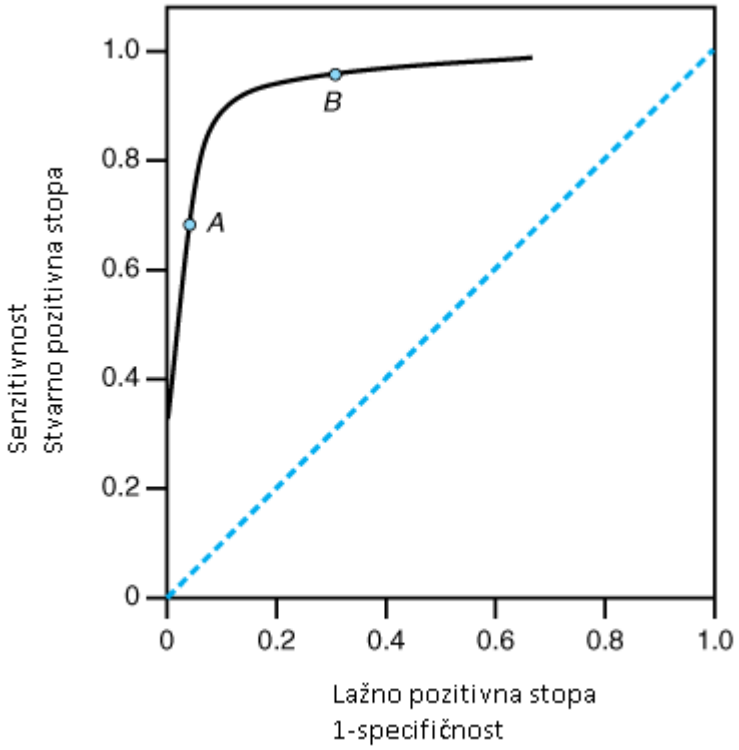


Slika 2.2: Dvije hipotetske distribucije s graničnom vrijednosti testa u točki 60. Stvarno pozitivni (TP), lažno negativni (FN), stvarno negativni (TN), lažno pozitivni (FP). Veća senzitivnost, niža specifičnost.

- Lažno pozitivna stopa, fpr = 1 − specifičnost = $\frac{FP}{TP+FP}$

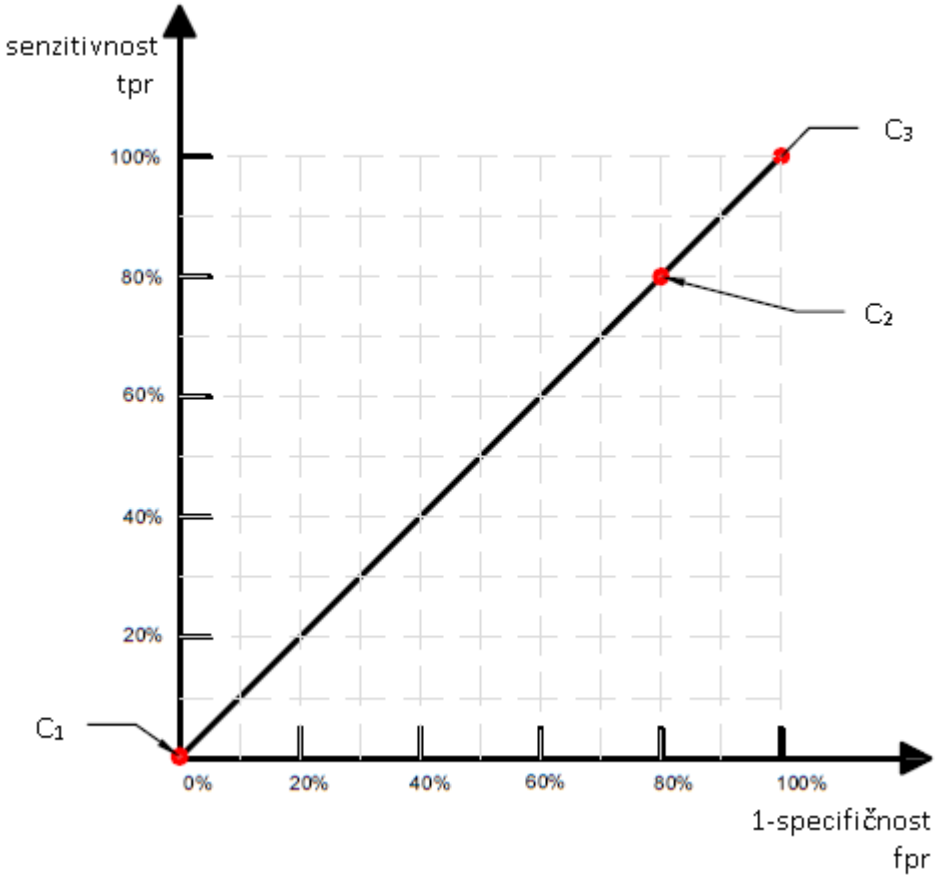
Što je ROC krivulja bliža gornjem lijevom kutu, to je test točniji jer je tada stvarno

pozitivna stopa jednaka 1, a lažno pozitivna stopa je 0. Što je kriterij za pozitivan test stroži, točka na krivulji koja odgovara senzitivnosti i specifičnosti (točka A) se pomiče prema dolje i lijevo (niža senzitivnost, viša specifičnost). Što je kriterij za pozitivan test blaži, točka na krivulji koja odgovara senzitivnosti i specifičnosti (točka B) se pomiče prema gore i desno (viša senzitivnost, niža specifičnost). Vidi sliku 2.3. ROC krivulja je korisna grafička metoda za uspoređivanje dva ili više dijagnostičih testova te za odabiranje granična vrijednost testa vrijednosti za određeni test.



Slika 2.3: ROC krivulja ([1],str.315)

Linija na slici 2.4, koju nazivamo bazična linija (*eng. baseline*), odgovara testu koji je pozitivan ili negativan sasvim slučajno. Na njoj su prikazana 3 slučajna klasifikatora: klasifikator C_1 predviđa nulu, tj. nepostojanje bolesti za svaki rezultat testa, drugi C_2 predviđa jedinicu, tj.postojanje bolesti sa senzitivnosti od 80% i specifičnosti od 20% (tj. fpr=80%), C_3 predviđa bolest u 100% slučajeva. U praksi nikada ne koristimo klasifikator koji je ispod bazične linije jer invertiranjem takvog klasifikatora uvijek možemo dobiti bolji.[6]



Slika 2.4: Bazična linija

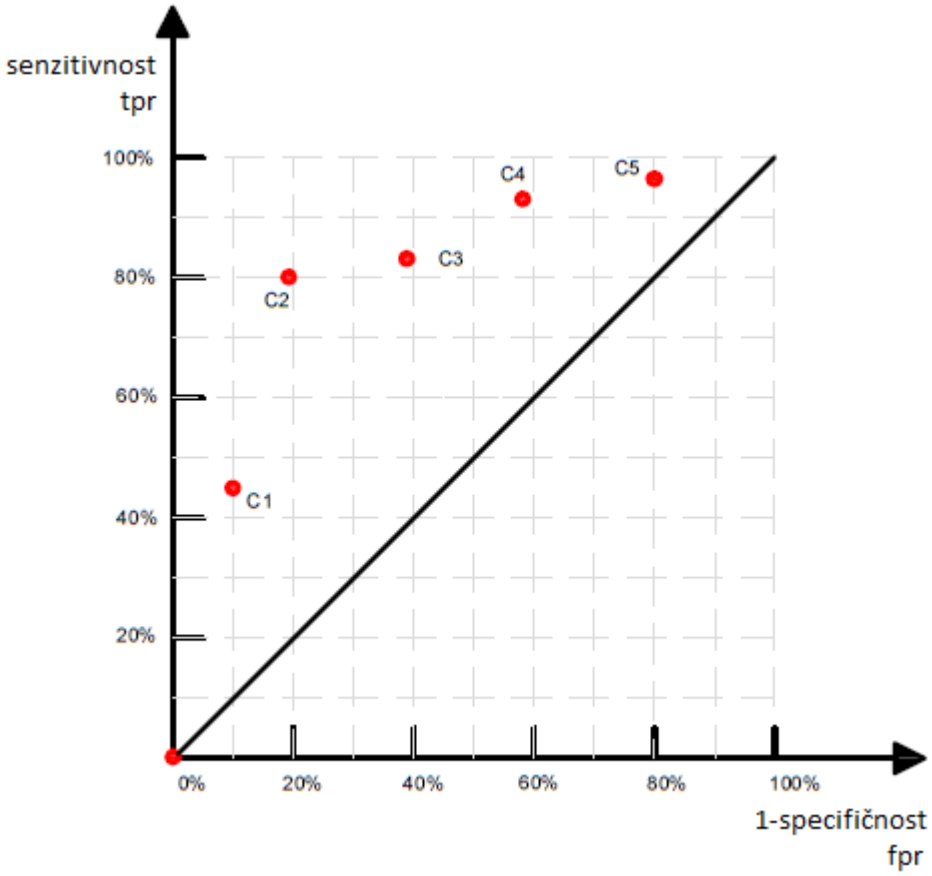
2.1 Konveksni ROC mnogokut

Definirajmo:

Definicija 2.1.1. Dominiranost: **a** dominira nad **b** $\Leftrightarrow \forall i f_i(a) \geq f_i(b)$ i $\exists j f_j(a) > f_j(b)$.

Definicija 2.1.2. Pareto efikasnost je preraspodjela resursa u kojoj je nemoguće napraviti preraspodjelu takvu da se poboljša individualni kriterij bez da se pogorša neki drugi. Pareto granica je skup svih Pareto efikasnih preraspodjela.

Pretpostavimo da imamo 5 klasifikatora C_1, C_2, C_3, C_4, C_5. Za svaki klasifikator izračunamo tpr i fpr te ih nacrtamo na grafu. Po principu dominiranosti ćemo dobiti Pareto granicu



Slika 2.5: Grafički prikaz pet klasifikatora

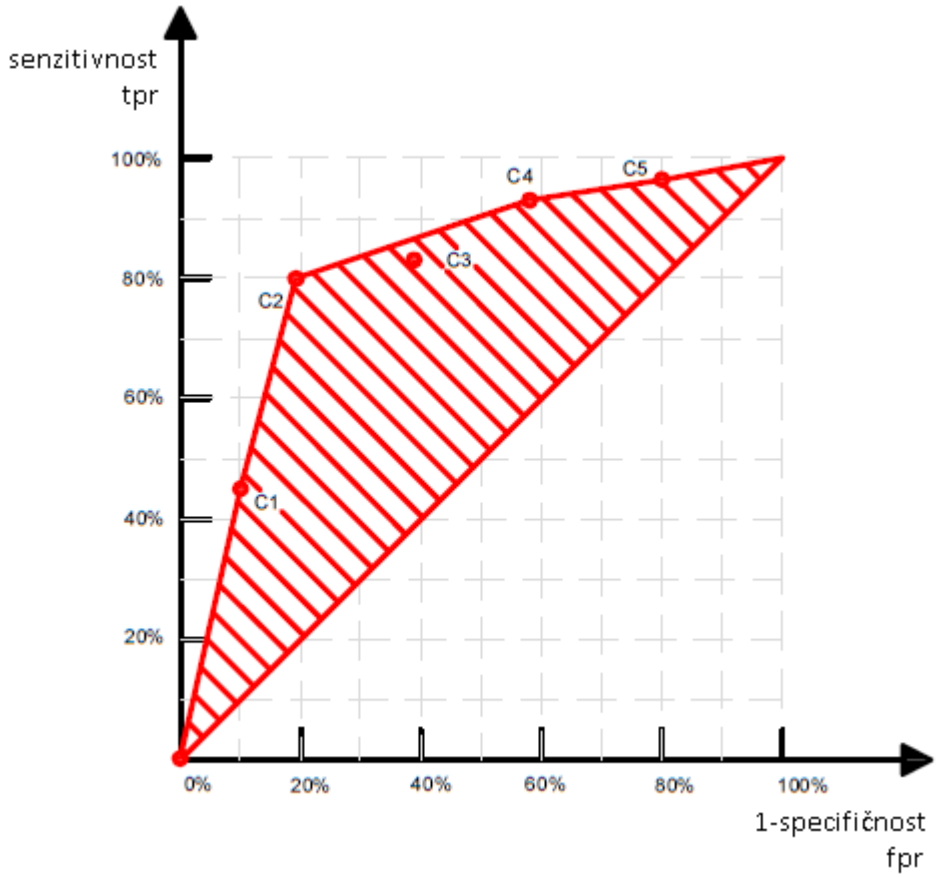
prikazanu na slici 2.6 koja s bazičnom linijom tvori konveksni ROC mnogokut.

Klasifikatori koji se nalaze unutar konveksnog mnogokuta su uvijek lošiji od onih na stranicama mnogokuta. Klasifikator C_3 će uvijek biti lošiji od ostalih.[6]

Linije točnosti

Postoji jednostavna veza između točnosti testa i stvarno pozitivne stope tpr te lažno pozitivne stope fpr. Uvodim nekoliko oznaka:

- N = broj observacija
- NEG = broj negativnih observacija = TN + FP



Slika 2.6: Konveksni ROC mnogokut

- POS = broj pozitivnih observacija = TP + FN
- neg = postotak negativnih observacija = $\frac{NEG}{N}$
- pos = postotak pozitivnih observacija = $\frac{POS}{N}$
- acc = točnost testa (postotak točne predikcije)

$$\begin{aligned} acc &= \frac{TP+TN}{N} \\ &= \frac{TP}{N} + \frac{TN}{N} \\ &= \frac{TP}{POS}\frac{POS}{N} + \frac{NEG-FP}{N} \\ &= \frac{TP}{POS}\frac{POS}{N} + \frac{NEG}{N} - \frac{FP}{NEG}\frac{NEG}{N} \\ &= tpr \cdot pos + neg - fpr \cdot neg \end{aligned} \tag{2.1}$$

Iz toga dobijemo

$$tpr = \frac{acc - neg}{pos} + \frac{neg}{pos} fpr \tag{2.2}$$

Uočimo da smo dobili pravac $y = ax + b$, gdje je $y = tpr, x = fpr, a = \frac{neg}{pos}, b = \frac{acc-neg}{pos}$. Omjer $\frac{neg}{pos}$ je koeficijent smjera našeg pravca. Da bismo izračunali točnost odgovarajućeg testa, pogledamo gdje se siječe linija točnosti (crna linija) i dijagonalna (crvena) linija te očitamo točnost na y osi.[6]

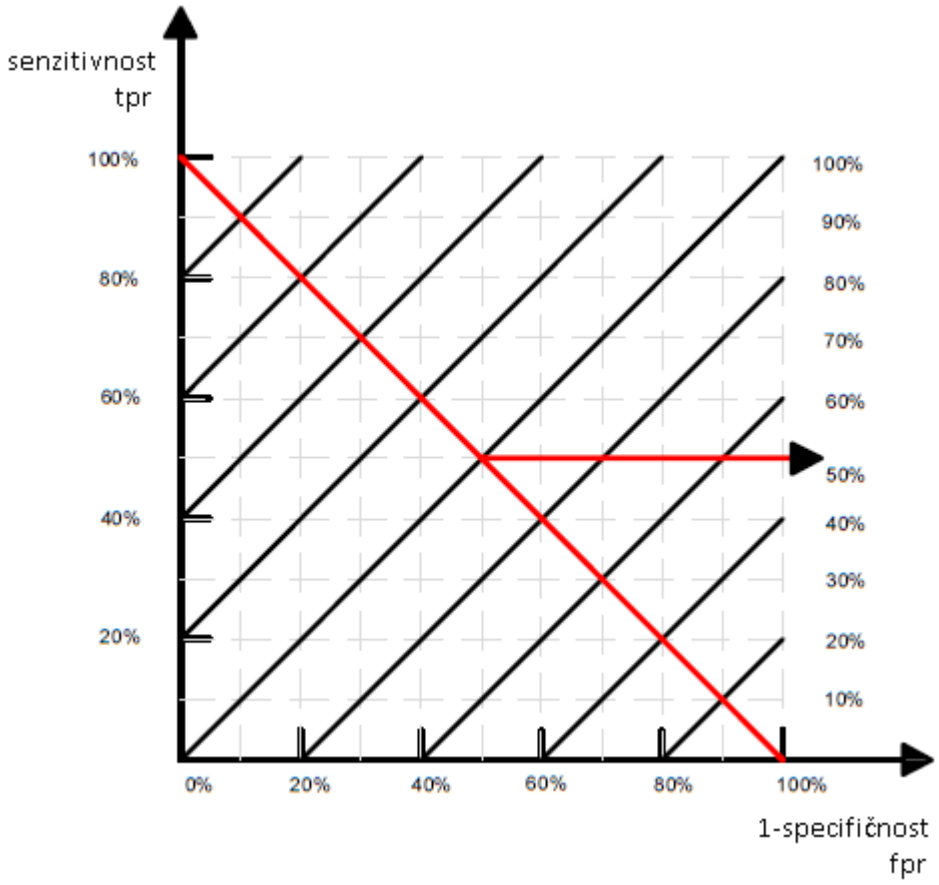
Primjer 2.1.1. *Na slici 2.7 su povučeni paralelni pravci s koeficijentom smjera* $\frac{neg}{pos} = 1$*, dok je na slici 2.8 koeficijent smjera* $\frac{neg}{pos} = 0.5$*.*

Primjer 2.1.2. *Svaka dužina konveksnog mnogokuta sa slike 2.6 je linija točnosti. Klasifikatori koji su na istoj dužini postižu istu točnost testa. Na primjeru ću pokazati kako odabrati optimalni klasifikator. Prvo moramo znati koliko iznosi omjer* $\frac{neg}{pos}$*.*

1. $\frac{neg}{pos} = 1$, stoga je $tpr = \frac{acc-neg}{pos} + 1 \cdot fpr$. Kako ne znamo koliko iznosi $\frac{acc-neg}{pos}$, nacrtamo sve pravce s koeficijentom smjera $\frac{neg}{pos} = 1$ koje prolaze kroz pet točaka klasifikatora. Očitamo točke sjecišta s padajućom dijagonalnom linijom i pripadne vijednosti na y osi. Za klasifikator C_1 je točnost 68%, za C_2 80%, za C_3 71%, za C_4 68%, za C_5 58%. Slika 2.9 pokazuje dobivene točnosti testa za sve klasifikatore te uočavamo da je najbolji klasifikator C_2. Analogno ide postupak za iduća dva koeficijenta smjera.

2. $\frac{neg}{pos} = \frac{1}{4}$; najbolji klasifikator je C_4 i točnost testa je 83%. Vidi sliku 2.10.

3. $\frac{neg}{pos} = \frac{4}{1}$; ponovno je najbolji klasifikator C_2 s točnošću testa od 80%. Vidi sliku 2.11.



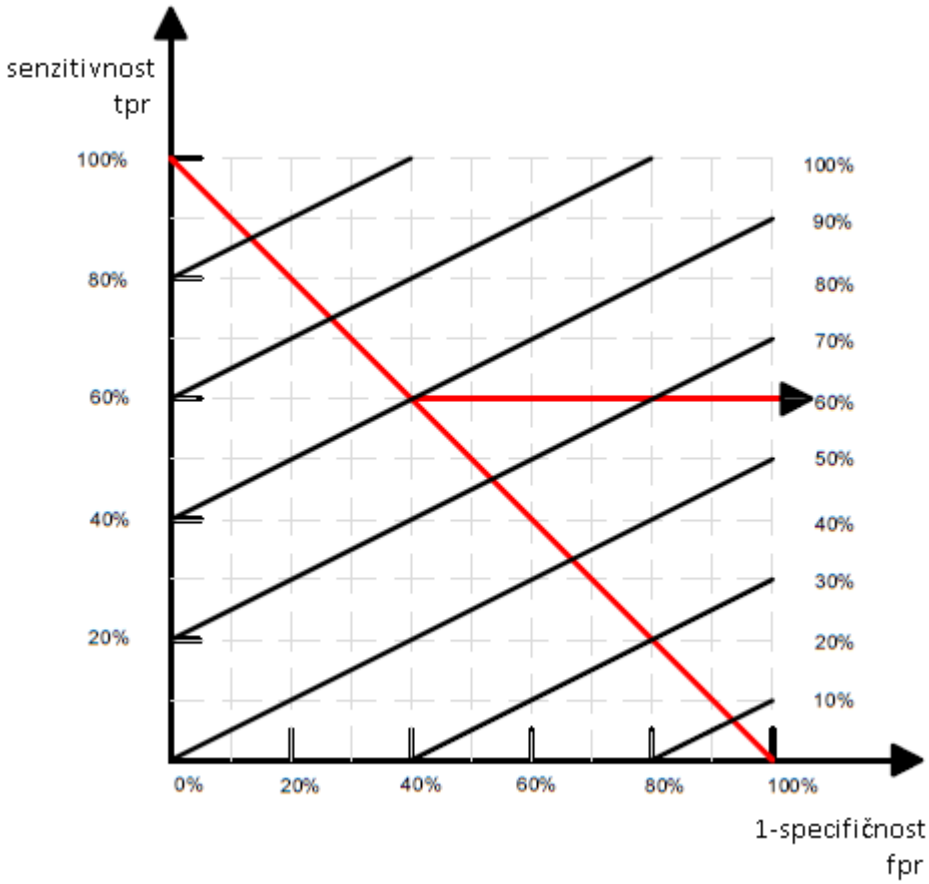
Slika 2.7: Grafički prikaz paralelnih pravaca s koeficijentom smjera $\frac{neg}{pos} = 1$

2.2 Površina ispod ROC krivulje

Površina ispod ROC krivulje (*eng. area under ROC curve, AUC*) je mjera za određivanje učinkovitosti nekog klasifikatora i jednaka je vjerojatnosti da će klasifikator rangirati slučajno odabrani pozitivan primjer višim nego što će rangirati slučajno odabrani negativan primjer, uz pretpostavku da pozitivni primjeri imaju veći rang. Matematičkim riječima zapisujemo:

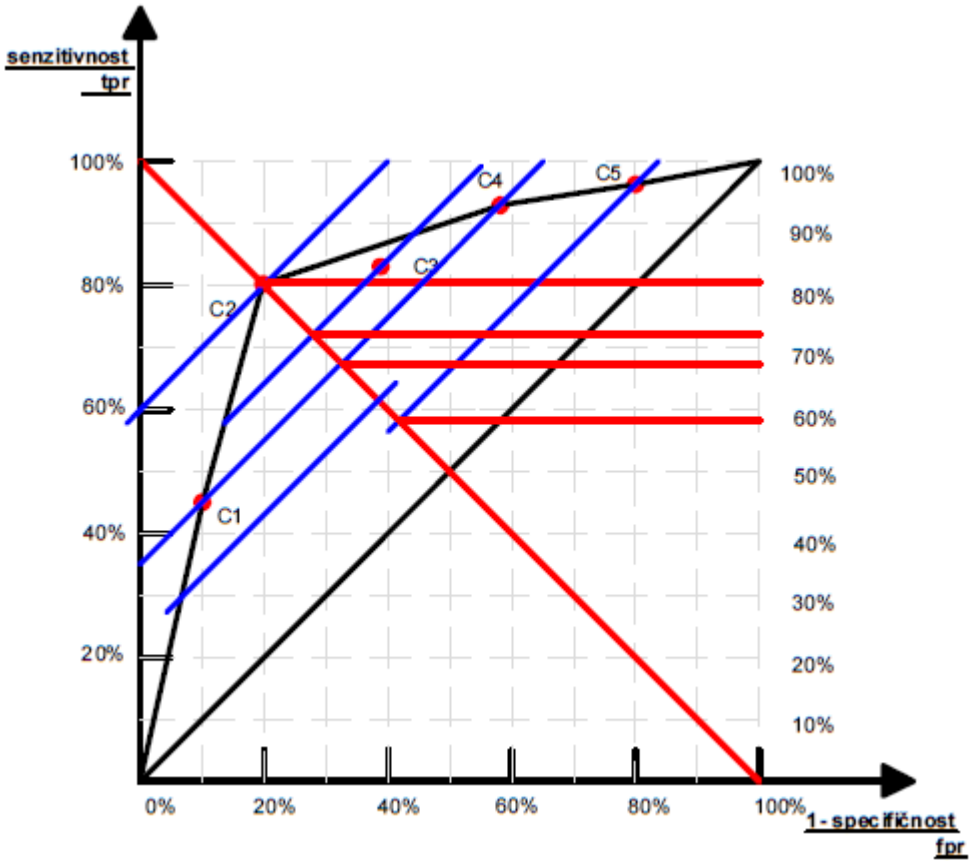
$$AUC(C) = \mathbb{P}[C(x^+) > C(x^-)]$$

Površina AUC je jednaka 1 ako se radi o idealnom klasifikatoru koji ispravno prepozna sve bolesne i sve zdrave pacijente, rangira sve pozitivne poslije svih negativnih. Ako je $AUC = 0.5$, tada se radi o slučajnom klasifikatoru koji nasumično određuje koji je pacijent

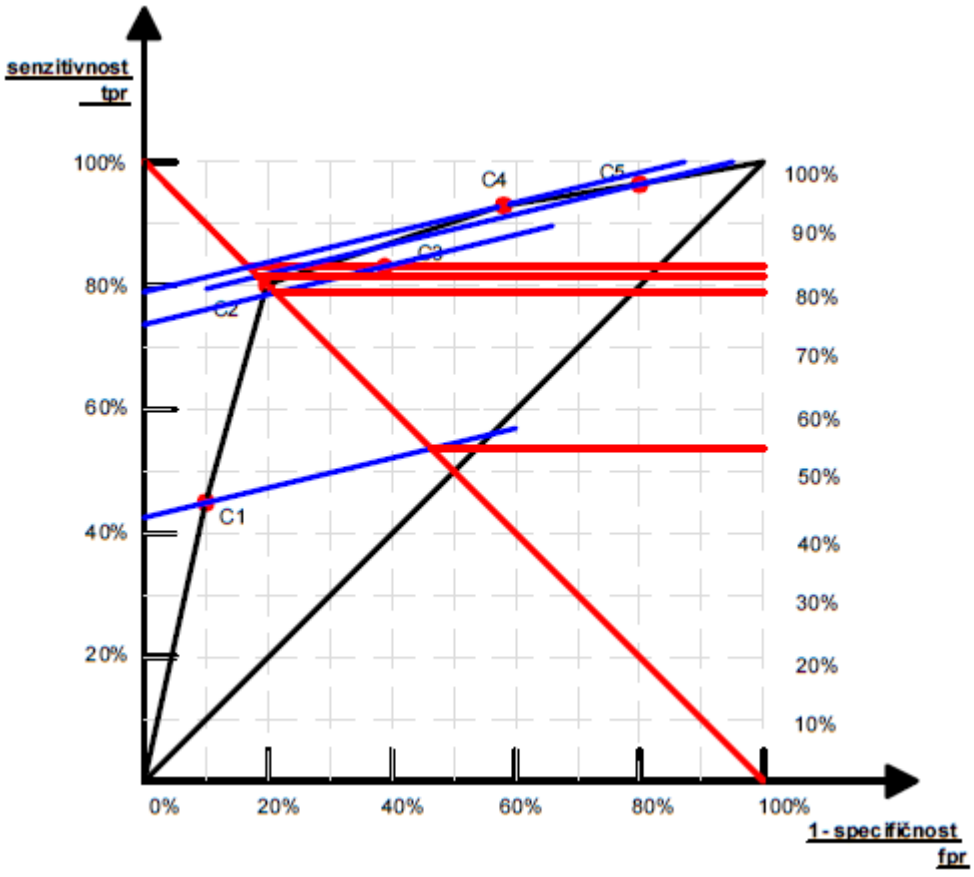


Slika 2.8: Grafički prikaz paralelnih pravaca s koeficijentom smjera $\frac{neg}{pos} = 0.5$

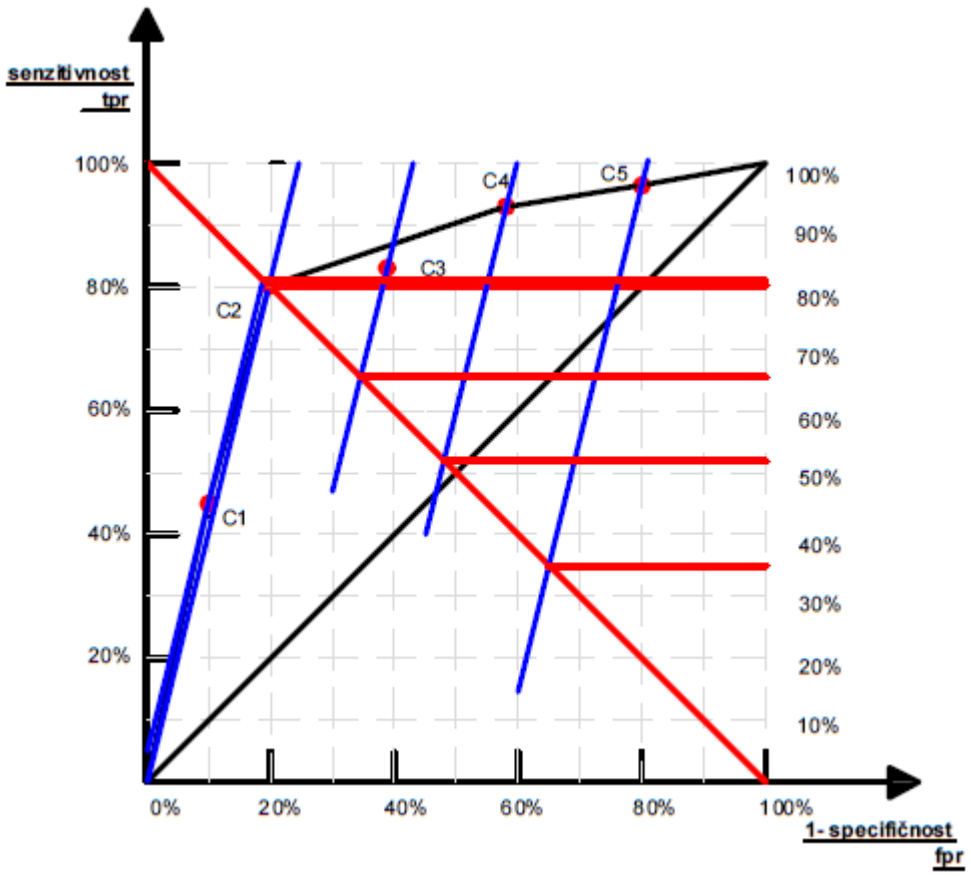
bolestan, koji je zdrav; nasumično rangira pozitivne i negativne. Ako je $AUC = 0$, tada klasifikator ne prepozna ispravno nijednog bolesnog, niti zdravog pacijenta; svi negativni su rangirani više od svih pozitivnih. Dakle $AUC \in [0, 1]$.[6]



Slika 2.9: Grafički prikaz paralelnih pravaca s koeficijentom smjera $\frac{neg}{pos} = 1$



Slika 2.10: Grafički prikaz paralelnih pravaca s koeficijentom smjera $\frac{neg}{pos} = \frac{1}{4}$



Slika 2.11: Grafički prikaz paralelnih pravaca s koeficijentom smjera $\frac{neg}{pos} = \frac{4}{1}$

Poglavlje 3

Primjeri

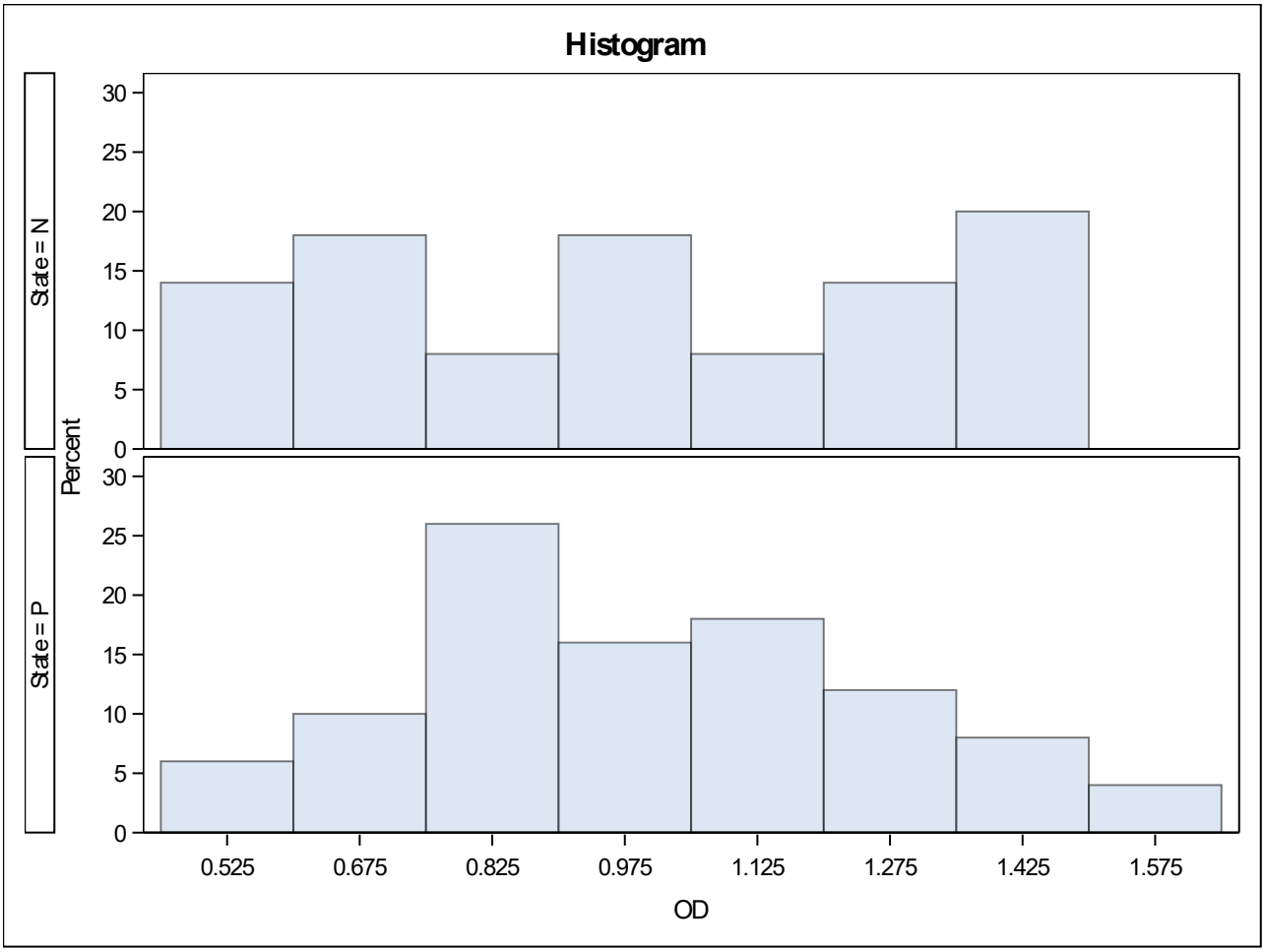
3.1 Testiranje medicinske opreme

Medicinska dijagnostička poduzeća proizvode opremu za snimanje pacijenata za razna medicinska stanja. Ta oprema služi za prepoznavanje bolesti na temelju uzorka testa pojedinog pacijenta. Određeno stanje je prepoznato preko abnormalne optičke gustoće koja je viša ili niža od određene granična vrijednost testa vrijednosti. Prije nego što se medicinska oprema može početi koristiti, potrebno je pokazati da ima dobre mogućnosti prepoznavanja pozitivnih i negativnih uzoraka. Također je nužno odrediti graničnu vrijednost testa (*cutoff value*). Navedeno ću pokazati pomoću ROC krivulje u programu SAS.[7]

Za konstruiranje ROC krivulje medicinske opreme potreban je broj pozitivnih i negativnih uzoraka. Oni su potom testirani preko nove medicinske opreme. Optička gustoća (OD) nove opreme je izmjerena i prikazana u tablici za svih 100 uzoraka, zajedno s njihovim stanjem testa (State) koji može biti pozitivan (P) ili negativan (N). 50 uzoraka je pozitivnih te 50 negativnih. Za dva različita stanja testa napravila sam deskriptivnu statistiku PROC MEANS procedurom. PROC UNIVARIATE procedurom prikazujem histograme, za negativan pa pozitivan test, respektivno. Iz histograma nije jasno koju vrijednost uzeti za graničnu vrijednost testa. Slijedi output iz SAS-a.

Tablica 3.1: Deskriptivna statistika optičke gustoće (OD). (Ispis iz SAS-a)

Analysis Variable : OD						
State	N Obs	Mean	Minimum	Median	Maximum	Std Dev
N	50	0.9852800	0.5030000	0.9915000	1.4770000	0.3210281
P	50	1.0102400	0.5650000	0.9695000	1.5320000	0.2546490



Slika 3.1: Histogram relativnih frekvencija optičke gustoće (OD) za pozitivno (P) i negativno (N) stanje. (Ispis iz SAS-a)

Potom sam odabrala ninc=20 različitih graničnih vrijednosti testa, između lowlim=0.50 i uplim=1.50, te za svaku izračunala TP, TN, FP i FN vrijednosti. ROC krivulja je prikaz stvarno pozitivnih (TP) u odnosu na lažno pozitivne (FP) za svaku graničnu vrijednost testa. Površina ispod ROC krivulje pokazuje mogućnost testa da razlikuje pozitivne od negativnih uzoraka. Krivulja će pomoći u određivanju granične vrijednosti testa. Konačan izbor granične vrijednosti testa ovisi o prirodi testa o kojem se radi. Za neke testove, lažno negativni rezultati mogu imati teže posljedice nego lažno pozitivni, dok za neke testove može biti obrnuto. Kod u SAS-u je sljedeći:

```
proc univariate data=datain;
  class State;
  var OD;
  histogram OD / nrows=2 odstitle="Histogram" ;
  ods select histogram;
run;

proc means data=datain NWAY MEAN MIN MEDIAN MAX STD;
        var=OD;
run;

\%macro roc(datain, lowlim, uplim, ninc=20);
options mtrace mprint;
data roc;
        set &datain;
        lowlim=&lowlim; uplim=&uplim; ninc=&ninc;
         do i=1 to ninc+1;
           cutoff=lowlim+(i-1)*((uplim-lowlim)/ninc);
           if OD > cutoff then test="R"; else test="N";
           output;
         end;
        drop i;
run;

proc print;
run;

proc sort; by cutoff;
run;
```

```
proc freq; by cutoff;
 table test*State / out=pcts1 outpct;
 run;

data truepos; set pcts1; if State="P" and test="R";
 tp_rate=pct_col; drop pct_col;
 run;

data falsepos; set pcts1; if State="N" and test="R";
 fp_rate=pct_col; drop pct_col;
 run;

data roc; merge truepos falsepos; by cutoff;
 if tp_rate=. then tp_rate=0.0;
 if fp_rate=. then fp_rate=0.0;
 run;

proc print;
run;

proc sgplot data=roc;
series x = fp_rate y = tp_rate;
scatter x = fp_rate y = tp_rate / group=cutoff;
styleattrs
  datalinepatterns=(dot solid longdash 26);
  lineparm x=0 y=0 slope=1;
run;
\%mend;

\%roc(datain,lowlim=0.50,uplim=1.50,ninc=20);
```

U macro funkciji se iz početnog skupa podataka kreira novi skup. U data koraku, DO petlja označava novi test pozitivnim (oznaka "R") ako je optička gustoća veća od određene granične vrijednosti testa, inače test označi negativnim (oznaka "N"). Unutar DO pelje u novi skup podataka spremam optičku gustoću, graničnu vrijednost testa i rezultat novog testa. Potom sortiram novi skup podataka po graničnoj vrijednosti testa. PROC FREQ procedura je pozvana s BY naredbom te su stvarno pozitivne i lažno pozitivne stope izlazni podaci za svaku graničnu vrijednost testa. Kreirani su zasebni skupovi podataka za stvarno pozitivne i lažno pozitivne stope za svaku graničnu vrijednost testa i potom spojene po graničnoj vrijednosti testa u novi skup podataka. S procedurom PROC SGPLOT crtam

ROC krivulju tako što prikazujem stvarno pozitivnu stopu u odnosu na lažno pozitivnu stopu za svaku graničnu vrijednost testa. ROC krivulja prati liniju koja označava slučajan test što ukazuje na to da testirana medicinska oprema ne razlikuje dobro pozitivne od negativnih uzoraka. To potvrđuje i površina ispod ROC krivulje koja iznosi samo 0.5284. Slijedi output iz SAS-a.

Tablica 3.2: Tablica frekvencija za graničnu vrijednost 0.5. (Ispis iz SAS-a)

cutoff=0.5

Table of test by State			
test	**State**		
Frequency **Percent** **Row Pct** **Col Pct**	**N**	**P**	**Total**
R	50 50.00 50.00 100.00	50 50.00 50.00 100.00	100 100.00
Total	50 50.00	50 50.00	100 100.00

Tablica 3.3: Tablica frekvencija za graničnu vrijednost 0.55. (Ispis iz SAS-a)

cutoff=0.55

Table of test by State			
test	**State**		
Frequency **Percent** **Row Pct** **Col Pct**	**N**	**P**	**Total**
N	4 4.00 100.00 8.00	0 0.00 0.00 0.00	4 4.00
R	46 46.00 47.92 92.00	50 50.00 52.08 100.00	96 96.00
Total	50 50.00	50 50.00	100 100.00

Tablica 3.4: Tablica frekvencija za graničnu vrijednost 0.6. (Ispis iz SAS-a)

cutoff=0.6

Table of test by State			
test	**State**		
Frequency **Percent** **Row Pct** **Col Pct**	**N**	**P**	**Total**
N	7 7.00 70.00 14.00	3 3.00 30.00 6.00	10 10.00
R	43 43.00 47.78 86.00	47 47.00 52.22 94.00	90 90.00
Total	50 50.00	50 50.00	100 100.00

Tablica 3.5: Tablica frekvencija za graničnu vrijednost 0.65. (Ispis iz SAS-a)

cutoff=0.65

Table of test by State			
test	**State**		
Frequency **Percent** **Row Pct** **Col Pct**	**N**	**P**	**Total**
N	11 11.00 78.57 22.00	3 3.00 21.43 6.00	14 14.00
R	39 39.00 45.35 78.00	47 47.00 54.65 94.00	86 86.00
Total	50 50.00	50 50.00	100 100.00

Tablica 3.6: Tablica frekvencija za graničnu vrijednost 0.7. (Ispis iz SAS-a)

cutoff=0.7

Table of test by State			
test	**State**		
Frequency Percent Row Pct Col Pct	**N**	**P**	**Total**
N	14 14.00 77.78 28.00	4 4.00 22.22 8.00	18 18.00
R	36 36.00 43.90 72.00	46 46.00 56.10 92.00	82 82.00
Total	50 50.00	50 50.00	100 100.00

Tablica 3.7: Tablica frekvencija za graničnu vrijednost 0.75. (Ispis iz SAS-a)

cutoff=0.75

Table of test by State			
test	**State**		
Frequency Percent Row Pct Col Pct	**N**	**P**	**Total**
N	16 16.00 66.67 32.00	8 8.00 33.33 16.00	24 24.00
R	34 34.00 44.74 68.00	42 42.00 55.26 84.00	76 76.00
Total	50 50.00	50 50.00	100 100.00

Tablica 3.8: Tablica frekvencija za graničnu vrijednost 0.8. (Ispis iz SAS-a)

cutoff=0.8

Table of test by State			
test	**State**		
Frequency **Percent** **Row Pct** **Col Pct**	**N**	**P**	**Total**
N	19 19.00 61.29 38.00	12 12.00 38.71 24.00	31 31.00
R	31 31.00 44.93 62.00	38 38.00 55.07 76.00	69 69.00
Total	50 50.00	50 50.00	100 100.00

Tablica 3.9: Tablica frekvencija za graničnu vrijednost 0.85. (Ispis iz SAS-a)

cutoff=0.85

Table of test by State			
test	**State**		
Frequency **Percent** **Row Pct** **Col Pct**	**N**	**P**	**Total**
N	20 20.00 54.05 40.00	17 17.00 45.95 34.00	37 37.00
R	30 30.00 47.62 60.00	33 33.00 52.38 66.00	63 63.00
Total	50 50.00	50 50.00	100 100.00

Tablica 3.10: Tablica frekvencija za graničnu vrijednost 0.9. (Ispis iz SAS-a)

cutoff=0.9

Table of test by State			
test	**State**		
Frequency **Percent** **Row Pct** **Col Pct**	**N**	**P**	**Total**
N	20 20.00 48.78 40.00	21 21.00 51.22 42.00	41 41.00
R	30 30.00 50.85 60.00	29 29.00 49.15 58.00	59 59.00
Total	50 50.00	50 50.00	100 100.00

Tablica 3.11: Tablica frekvencija za graničnu vrijednost 0.95. (Ispis iz SAS-a)

cutoff=0.95

Table of test by State			
test	**State**		
Frequency **Percent** **Row Pct** **Col Pct**	**N**	**P**	**Total**
N	24 24.00 50.00 48.00	24 24.00 50.00 48.00	48 48.00
R	26 26.00 50.00 52.00	26 26.00 50.00 52.00	52 52.00
Total	50 50.00	50 50.00	100 100.00

Tablica 3.12: Tablica frekvencija za graničnu vrijednost 1. (Ispis iz SAS-a)

cutoff=1

Table of test by State			
test	**State**		
Frequency Percent Row Pct Col Pct	**N**	**P**	**Total**
N	25 25.00 49.02 50.00	26 26.00 50.98 52.00	51 51.00
R	25 25.00 51.02 50.00	24 24.00 48.98 48.00	49 49.00
Total	50 50.00	50 50.00	100 100.00

Tablica 3.13: Tablica frekvencija za graničnu vrijednost 1.05. (Ispis iz SAS-a)

cutoff=1.05

Table of test by State			
test	**State**		
Frequency Percent Row Pct Col Pct	**N**	**P**	**Total**
N	29 29.00 50.00 58.00	29 29.00 50.00 58.00	58 58.00
R	21 21.00 50.00 42.00	21 21.00 50.00 42.00	42 42.00
Total	50 50.00	50 50.00	100 100.00

Tablica 3.14: Tablica frekvencija za graničnu vrijednost 1.1. (Ispis iz SAS-a)

cutoff=1.1

Table of test by State			
test	**State**		
Frequency **Percent** **Row Pct** **Col Pct**	**N**	**P**	**Total**
N	31 31.00 51.67 62.00	29 29.00 48.33 58.00	60 60.00
R	19 19.00 47.50 38.00	21 21.00 52.50 42.00	40 40.00
Total	50 50.00	50 50.00	100 100.00

Tablica 3.15: Tablica frekvencija za graničnu vrijednost 1.15. (Ispis iz SAS-a)

cutoff=1.15

Table of test by State			
test	**State**		
Frequency **Percent** **Row Pct** **Col Pct**	**N**	**P**	**Total**
N	33 33.00 50.77 66.00	32 32.00 49.23 64.00	65 65.00
R	17 17.00 48.57 34.00	18 18.00 51.43 36.00	35 35.00
Total	50 50.00	50 50.00	100 100.00

Tablica 3.16: Tablica frekvencija za graničnu vrijednost 1.2. (Ispis iz SAS-a)

cutoff=1.2

Table of test by State			
test	**State**		
Frequency **Percent** **Row Pct** **Col Pct**	**N**	**P**	**Total**
N	33 33.00 46.48 66.00	38 38.00 53.52 76.00	71 71.00
R	17 17.00 58.62 34.00	12 12.00 41.38 24.00	29 29.00
Total	50 50.00	50 50.00	100 100.00

Tablica 3.17: Tablica frekvencija za graničnu vrijednost 1.25. (Ispis iz SAS-a)

cutoff=1.25

Table of test by State			
test	**State**		
Frequency **Percent** **Row Pct** **Col Pct**	**N**	**P**	**Total**
N	36 36.00 46.15 72.00	42 42.00 53.85 84.00	78 78.00
R	14 14.00 63.64 28.00	8 8.00 36.36 16.00	22 22.00
Total	50 50.00	50 50.00	100 100.00

Tablica 3.18: Tablica frekvencija za graničnu vrijednost 1.3. (Ispis iz SAS-a)

cutoff=1.3

Table of test by State			
test	**State**		
Frequency **Percent** **Row Pct** **Col Pct**	**N**	**P**	**Total**
N	39 39.00 47.56 78.00	43 43.00 52.44 86.00	82 82.00
R	11 11.00 61.11 22.00	7 7.00 38.89 14.00	18 18.00
Total	50 50.00	50 50.00	100 100.00

Tablica 3.19: Tablica frekvencija za graničnu vrijednost 1.35. (Ispis iz SAS-a)

cutoff=1.35

Table of test by State			
test	**State**		
Frequency **Percent** **Row Pct** **Col Pct**	**N**	**P**	**Total**
N	40 40.00 47.62 80.00	44 44.00 52.38 88.00	84 84.00
R	10 10.00 62.50 20.00	6 6.00 37.50 12.00	16 16.00
Total	50 50.00	50 50.00	100 100.00

Tablica 3.20: Tablica frekvencija za graničnu vrijednost 1.4. (Ispis iz SAS-a)

cutoff=1.4

Table of test by State			
test	**State**		
Frequency **Percent** **Row Pct** **Col Pct**	**N**	**P**	**Total**
N	42 42.00 47.73 84.00	46 46.00 52.27 92.00	88 88.00
R	8 8.00 66.67 16.00	4 4.00 33.33 8.00	12 12.00
Total	50 50.00	50 50.00	100 100.00

Tablica 3.21: Tablica frekvencija za graničnu vrijednost 1.45. (Ispis iz SAS-a)

cutoff=1.45

Table of test by State			
test	**State**		
Frequency **Percent** **Row Pct** **Col Pct**	**N**	**P**	**Total**
N	46 46.00 48.94 92.00	48 48.00 51.06 96.00	94 94.00
R	4 4.00 66.67 8.00	2 2.00 33.33 4.00	6 6.00
Total	50 50.00	50 50.00	100 100.00

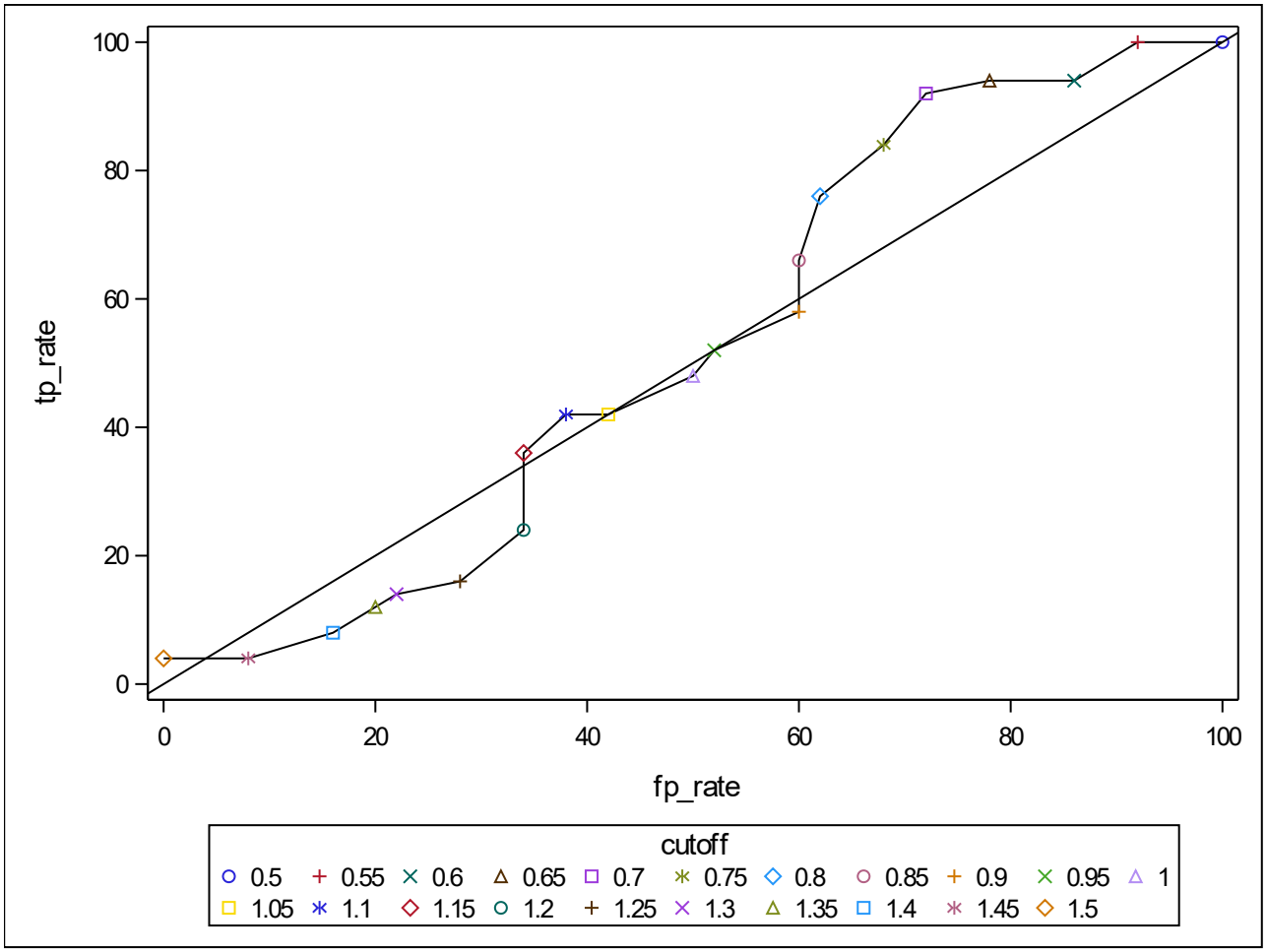
Tablica 3.22: Tablica frekvencija za graničnu vrijednost 1.5. (Ispis iz SAS-a)

cutoff=1.5

Table of test by State			
test	**State**		
Frequency Percent Row Pct Col Pct	**N**	**P**	**Total**
N	50 50.00 51.02 100.00	48 48.00 48.98 96.00	98 98.00
R	0 0.00 0.00 0.00	2 2.00 100.00 4.00	2 2.00
Total	50 50.00	50 50.00	100 100.00

Tablica 3.23: Tablica graničnih vrijednosti testa s pripadajućom senzitivnosti (tp_rate) i 1-specifičnosti (fp_rate). (Ispis iz SAS-a)

Obs	cutoff	test	State	COUNT	PERCENT	PCT_ROW	tp_rate	fp_rate
1	0.50	R	N	50	50	50.000	100	100
2	0.55	R	N	46	46	47.917	100	92
3	0.60	R	N	43	43	47.778	94	86
4	0.65	R	N	39	39	45.349	94	78
5	0.70	R	N	36	36	43.902	92	72
6	0.75	R	N	34	34	44.737	84	68
7	0.80	R	N	31	31	44.928	76	62
8	0.85	R	N	30	30	47.619	66	60
9	0.90	R	N	30	30	50.847	58	60
10	0.95	R	N	26	26	50.000	52	52
11	1.00	R	N	25	25	51.020	48	50
12	1.05	R	N	21	21	50.000	42	42
13	1.10	R	N	19	19	47.500	42	38
14	1.15	R	N	17	17	48.571	36	34
15	1.20	R	N	17	17	58.621	24	34
16	1.25	R	N	14	14	63.636	16	28
17	1.30	R	N	11	11	61.111	14	22
18	1.35	R	N	10	10	62.500	12	20
19	1.40	R	N	8	8	66.667	8	16
20	1.45	R	N	4	4	66.667	4	8
21	1.50	R	P	2	2	100.000	4	0

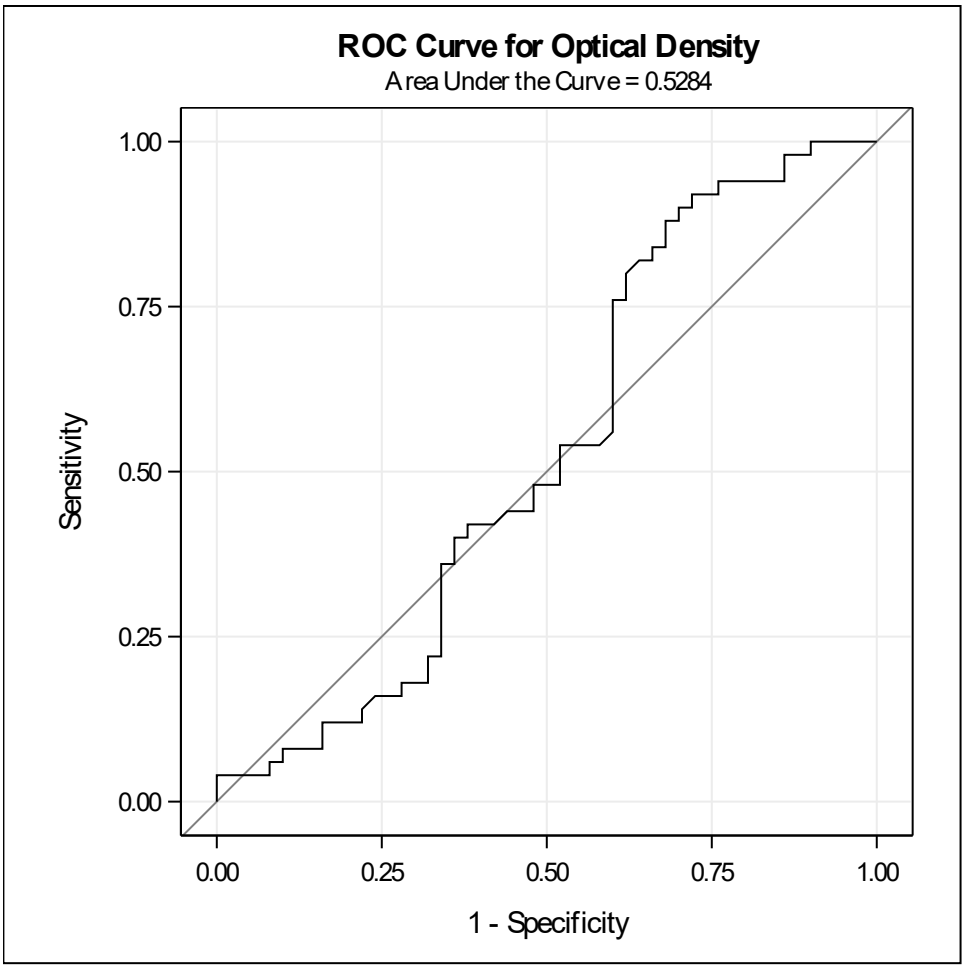


Slika 3.2: Prikaz 1-specifičnosti (fp_rate) i senzitivnosti (tp_rate) za različite granične vrijednosti (cutoff) koristeći macro funkciju. (Ispis iz SAS-a)

U tablicama 3.2 do 3.22 su prikazane tablice frekvencija za svaku graničnu vrijednost testa, potom je u tablici 3.23 za svaku graničnu vrijednost testa navedena stvarno pozitivna stopa i lažno pozitivna stopa te su ti podaci prikazani na ROC krivulji (Slika 3.2). Slika 3.3 prikazuje istu ROC krivulju dobivenu PROC LOGISTIC procedurom te tablicu 3.24 u kojoj je navedena površina ispod ROC krivulje, standardna pogreška i 95% pouzdani interval koji sadrži 0.5, što nam govori da se ROC krivulja za optičku gustoću statistički ne razlikuje od slučajnog testa koji odgovara dijagonali na ROC grafu. Slijedi SAS kod [9]:

```
ods graphics on;
proc logistic data=datain plots=roc;
        model State(event='P') = OD / nofit;
    roc 'Optical Density' OD;
run;
ods graphics off;
```

Primjer je uzet iz članka "Constructing ROC Curves with the SAS System".[5]



Slika 3.3: Prikaz 1-specifičnosti (fp_rate) i senzitivnosti (tp_rate) za različite granične vrijednosti (cutoff) koristeći proc logistic proceduru. (Ispis iz SAS-a)

Tablica 3.24: Tablični prikaz površine ispod ROC krivulje i 95% pouzdanog intervala. (Ispis iz SAS-a)

ROC Association Statistics							
	Mann-Whitney						
ROC Model	Area	Standard Error	95% Wald Confidence Limits		Somers' D	Gamma	Tau-a
Optical Density	0.5284	0.0597	0.4114	0.6454	0.0568	0.0569	0.0287

3.2 Odabiranje proizvoda u prodajni asortiman

Kroz ovaj primjer želim pokazati kako se ROC krivulje koriste ne samo u medicinskom području, već u raznim poljima kao što su telekomunikacije, IT, industrija itd. Pokazat ću kako se ROC krivulje mogu koristiti za usporedbu logističkih modela. U ovom primjeru promatram skup podataka "import" od 8080 zabilježenih svojstava proizvoda za koje trebamo odlučiti koji će se nastaviti prodavati, a koji se trebaju izbaciti iz inventara. Koristeći logističku regresiju ću kreirati 3 modela te usporedbom pripadajućih ROC krivulja ću dobiti koje varijable je najbolje promotriti prilikom određivanja o proizvodima u asortimanu. Podaci su uzeti s web stranice www.kaggle.com [4] i sadrže sljedeće varijable:

- SKU_number: jedinstveni broj za svaki proizvod
- SoldFlag: 1 = proizvod je prodan u zadnjih 6 mjeseci, 0 = nije prodan u zadnjih 6 mjeseci
- MarketingType: dvije vrste plasiranja proizvoda u prodaju (D i S)
- ReleaseNumber: broj izdanja
- New_Release_Flag: 1 = proizvod koji ima novije izdanje,0 = nema
- StrengthFactor: faktor prodaje
- PriceReg: troškovi proizvodnje
- ReleaseYear: godina puštanja u prodaju
- ItemCount: količina proizvoda
- LowUserPrice: najniža cijena za korisnike
- LowNetPrice: najniža online cijena

Logistička regresija je regresijski model čija je zavisna varijabla kategorijska, a nezavisne su kvantitativne ili kategorijske. Logistička funkcija je definirana s

$$p(x) = \frac{1}{1 + e^{-x}} \tag{3.1}$$

gdje je $x \in < -\infty, +\infty >$, $p(x) \in < 0, 1 >$.
Logit funkcija je funkcija inverzna logističkoj, tj.:

$$logit(p(x)) = log\Big[\frac{p(x)}{1 - p(x)}\Big] \tag{3.2}$$

gdje je $p \in < 0, 1 >$, $logit(p) \in < -\infty, +\infty >$. [3]

Podacima prilagođavam 3 različita logistička modela:

1. **Model 1**: $logit(SoldFlag) = \beta_1 SKU_number + \beta_2 ReleaseNumber + \beta_3 New_Release_Flag + \beta_4 StrengthFactor + \beta_5 PriceReg + \beta_6 ReleaseYear + \beta_7 ItemCount + \beta_8 LowUserPrice + \beta_9 LowNetPrice$

2. **Model 2**: $logit(SoldFlag) = \beta_1 New_Release_Flag + \beta_2 PriceReg + \beta_3 ItemCount$

3. **Model 3**: $logit(SoldFlag) = \beta_1 StrengthFactor + \beta_2 PriceReg + \beta_3 LowUserPrice + \beta_4 LowNetPrice$

Deskriptivna statistika dobivena PROC MEANS procedurom je dana u Tablici 3.25.

Za svaki model konstruiram ROC krivulju te ih na istom grafu uspoređujem. ROC krivulja koja je najbliža gornjem lijevom kutu, tj. s najvećom površinom daje model koji najbolje odgovara podacima.

PROC LOGISTIC procedura uzima podatke iz danog skupa podataka "import". Opcija NOFIT u proceduri MODEL služi za to da navedene nezavisne varijable ne budu fitane u model, već da se fitaju modeli navedeni u ROC proceduri. ROCCONTRAST uspoređuje gornje dobivene ROC krivulje. U REFERENCE opciji navodimo koja krivulja će biti referentna te se konstruira kontrastna matrica razlika između svake ROC krivulje i referentne krivulje. ESTIMATE opcija procjenjuje i računa svaku usporedbu, opcija E navodi kontrastne koeficijente. Slijedi SAS kod i ispis.[8] U ispisu su za svaki model navedeni koeficijenti, njihove procjene dobivene metodom maksimalne vjerodostojnosti te ROC krivulja. Iz usporedbe ROC krivulja se vidi da je Model 1 najbolji model.

```
proc means data=import NWAY MEAN MIN MEDIAN MAX STD;
var SoldFlag ReleaseNumber New_Release_Flag StrengthFactor
ReleaseYear PriceReg ItemCount LowUserPrice LowNetPrice;
run;

ods graphics on;
proc logistic data=import plots=roc;
  class MarketingType;
  model SoldFlag(event='1') = SKU_number MarketingType
  ReleaseNumber New_Release_Flag StrengthFactor  PriceReg
  ReleaseYear ItemCount LowUserPrice LowNetPrice / nofit;
  roc 'Model1' SKU_number ReleaseNumber New_Release_Flag
  StrengthFactor PriceReg ReleaseYear ItemCount LowUserPrice
```

```
  LowNetPrice;
  roc 'Model2' New_Release_Flag PriceReg ItemCount;
  roc 'Model3' StrengthFactor PriceReg LowUserPrice LowNetPrice;
  roccontrast reference('Model1') / estimate e;
run;
ods graphics off;
```

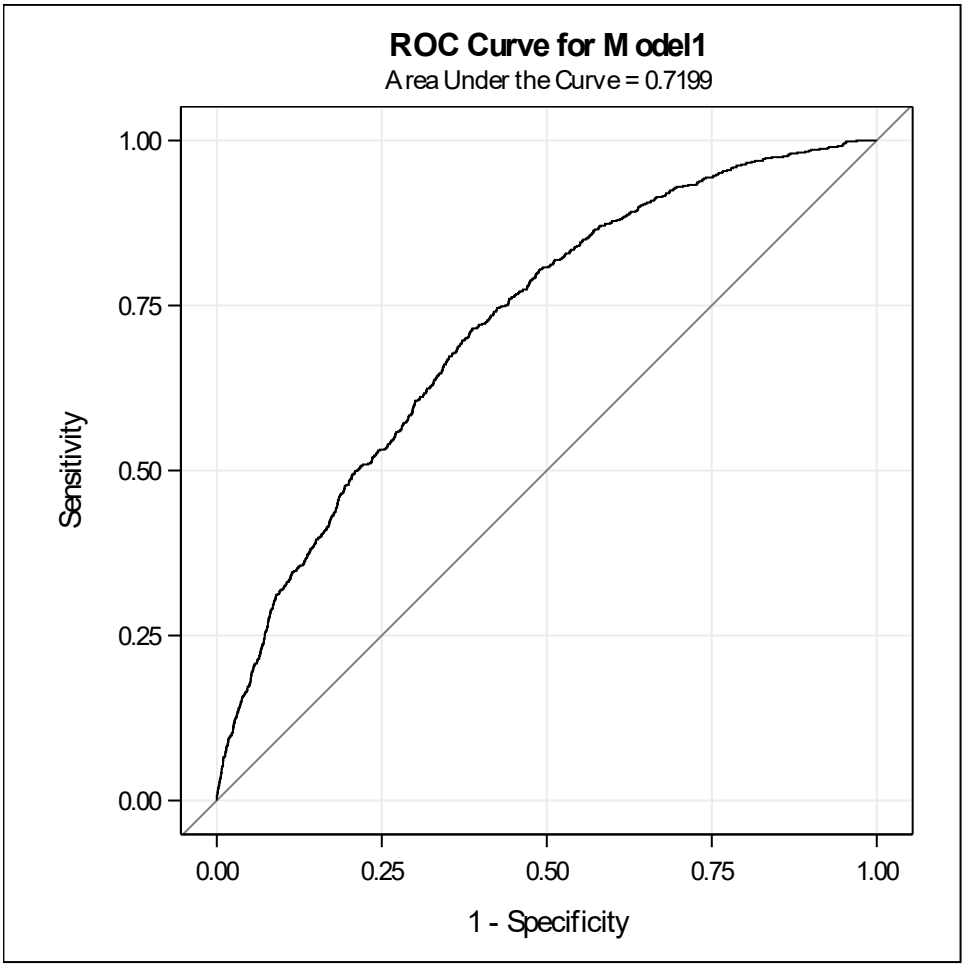
Tablica 3.25: Deskriptivna statistika. (Ispis iz SAS-a)

Variable	Mean	Minimum	Median	Maximum	Std Dev
SoldFlag	0.0882426	0	0	1.0000000	0.2836649
ReleaseNumber	3.8240099	0	3.0000000	61.0000000	3.7214855
New_Release_Flag	0.7465347	0	1.0000000	1.0000000	0.4350219
StrengthFactor	2076031.27	230.0000000	1392331.50	16411446.00	2137116.10
ReleaseYear	2004.09	1954.00	2005.00	2016.00	6.7231370
PriceReg	92.3964950	0	72.0000000	1580.00	75.9888005
ItemCount	35.1070545	0	27.0000000	851.0000000	30.1014878
LowUserPrice	61.8407191	0	46.9900000	7781.00	112.8327558
LowNetPrice	52.5049059	0	35.4900000	19138.79	250.1497131

ROC Model: Model1

Tablica 3.26 : Procjena parametara metodom maksimalne vjerodostojnosti za Model 1. (Ispis iz SAS-a)

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	-21.0304	15.5621	1.8263	0.1766
SKU_number	1	-5.02E-8	6.877E-8	0.5322	0.4657
ReleaseNumber	1	0.0120	0.0111	1.1687	0.2797
New_Release_Flag	1	0.3433	0.1191	8.3007	0.0040
StrengthFactor	1	-5.07E-7	4.461E-8	129.2772	<.0001
PriceReg	1	0.000870	0.000516	2.8432	0.0918
ReleaseYear	1	0.00939	0.00775	1.4676	0.2257
ItemCount	1	0.00592	0.00112	28.0204	<.0001
LowUserPrice	1	0.000214	0.000241	0.7867	0.3751
LowNetPrice	1	-0.00098	0.000803	1.4852	0.2230

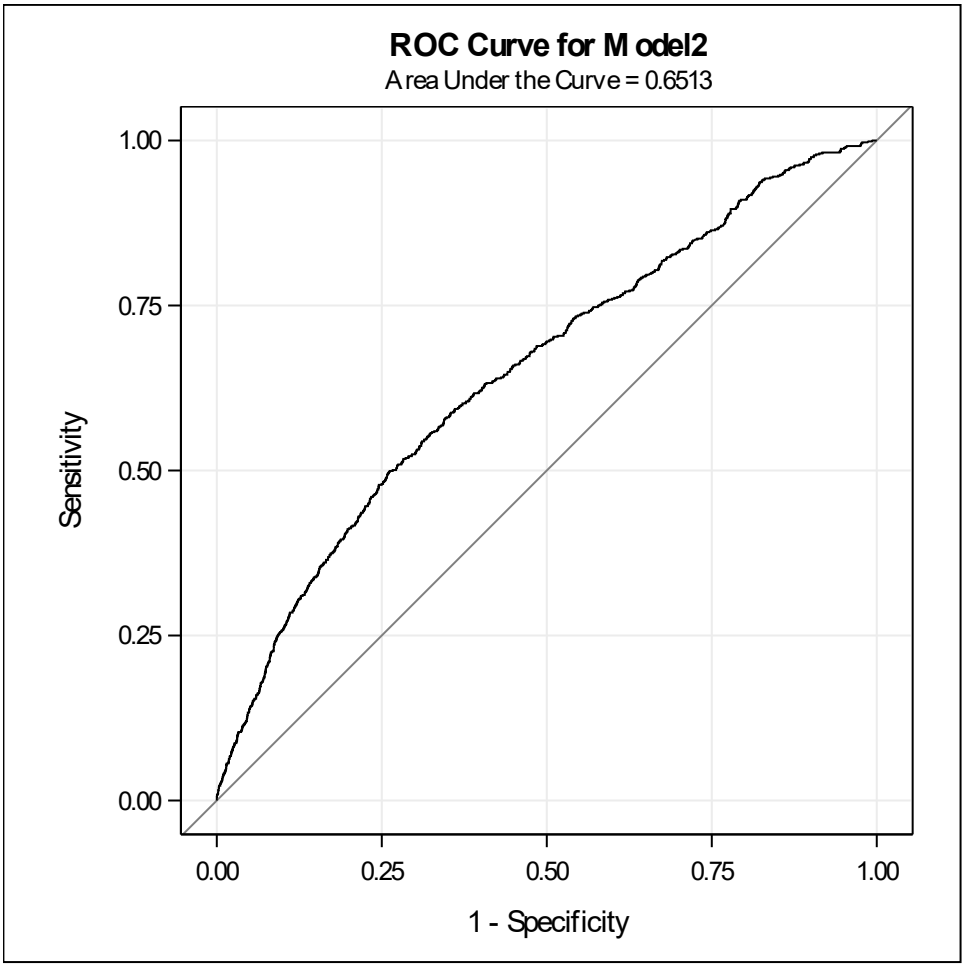


Slika 3.4: ROC krivulja za Model 1. (Ispis iz SAS-a)

ROC Model: Model2

Tablica 3.27 : Procjena parametara metodom maksimalne vjerodostojnosti za Model 2. (Ispis iz SAS-a)

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	-3.2984	0.1122	864.2893	<.0001
New_Release_Flag	1	0.4777	0.1053	20.5971	<.0001
PriceReg	1	0.00102	0.000473	4.6258	0.0315
ItemCount	1	0.0124	0.00108	132.0480	<.0001

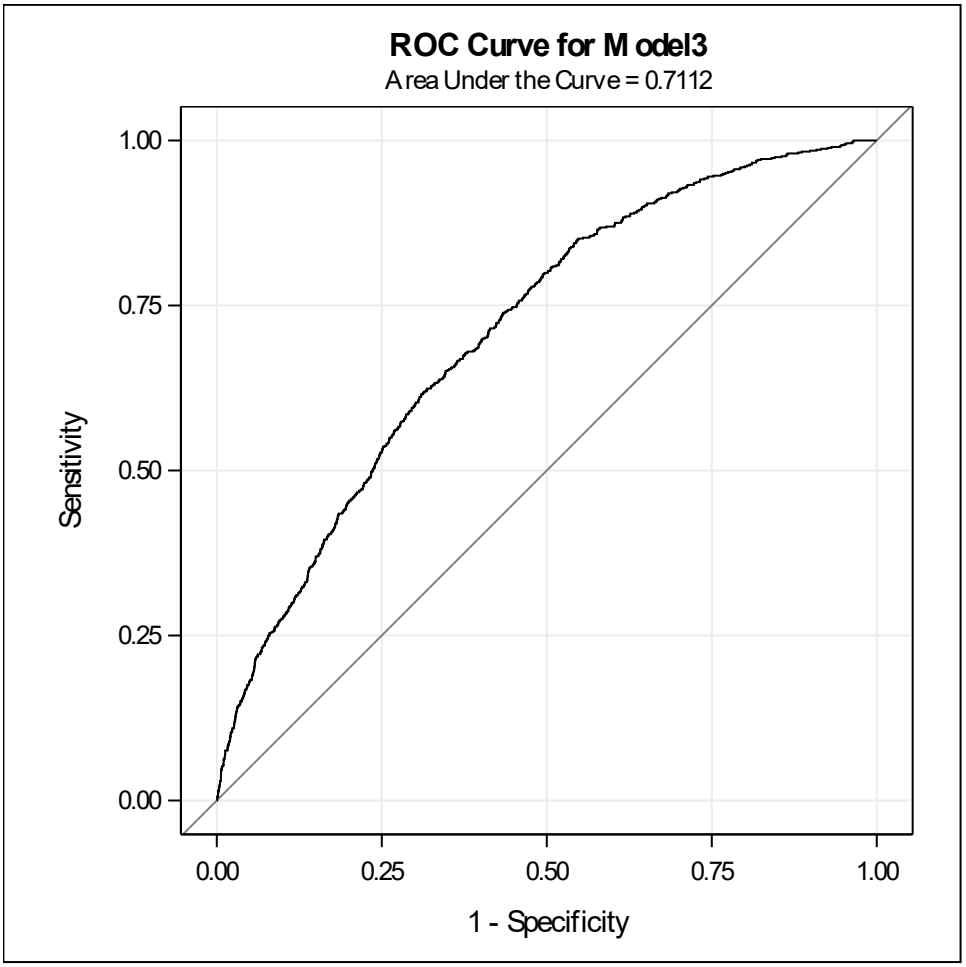


Slika 3.5: ROC krivulja za Model 2. (Ispis iz SAS-a)

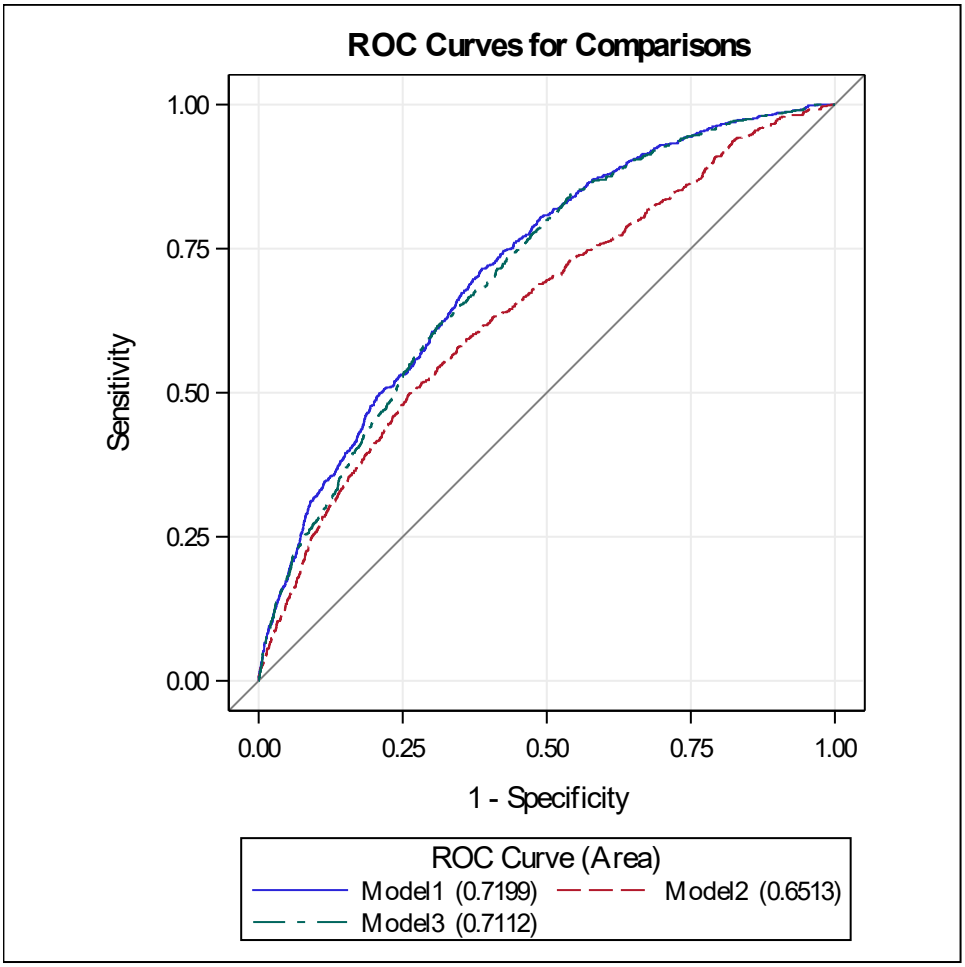
ROC Model: Model3

Tablica 3.28 : Procjena parametara metodom maksimalne vjerodostojnosti za Model 3. (Ispis iz SAS-a)

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	-1.5493	0.0785	389.4418	<.0001
StrengthFactor	1	-5.97E-7	4.171E-8	204.7846	<.0001
PriceReg	1	0.00136	0.000500	7.3997	0.0065
LowUserPrice	1	0.000239	0.000229	1.0947	0.2954
LowNetPrice	1	-0.00191	0.000889	4.5959	0.0320



Slika 3.6: ROC krivulja za Model 3. (Ispis iz SAS-a)



Slika 3.7: Usporedba ROC krivulja Modela 1, 2 i 3. (Ispis iz SAS-a)

Tablica 3.29 : Tablični prikaz površina ispod ROC krivulja i 95% pouzdani intervali za Model 1, 2 i 3. (Ispis iz SAS-a)

ROC Association Statistics							
	Mann-Whitney						
ROC Model	Area	Standard Error	95% Wald Confidence Limits		Somers' D	Gamma	Tau-a
Model1	0.7199	0.00951	0.7013	0.7386	0.4399	0.4399	0.0708
Model2	0.6513	0.0110	0.6297	0.6729	0.3025	0.3026	0.0487
Model3	0.7112	0.00953	0.6926	0.7299	0.4225	0.4225	0.0680

Tablica 3.30 : Kontrastna matrica. (Ispis iz SAS-a)

ROC Contrast Coefficients		
ROC Model	Row1	Row2
Model1	-1	-1
Model2	1	0
Model3	0	1

Tablica 3.31 : Referentni model Model 1. (Ispis iz SAS-a)

ROC Contrast Test Results			
Contrast	DF	Chi-Square	Pr > ChiSq
Reference = Model1	2	136.9198	<.0001

Tablica 3.32 : Tablica značajnosti modela. (Ispis iz SAS-a)

ROC Contrast Estimation and Testing Results by Row						
Contrast	Estimate	Standard Error	95% Wald Confidence Limits		Chi-Square	Pr > ChiSq
Model2 - Model1	-0.0687	0.00853	-0.0854	-0.0519	64.7975	<.0001
Model3 - Model1	-0.00868	0.00392	-0.0164	-0.00101	4.9136	0.0266

U tablici 3.29 navedene su površine ispod ROC krivulje za svaki model, standardna pogreška i 95% pouzdani interval. Nijedan interval ne sadrži 0.5, što znači da se sva tri modela statistički razlikuju od slučajnog modela koji je prikazan dijagonalom na ROC grafu.

Tablica 3.30 pokazuje da je Model 1 referentni model. Iz tablice 3.31 vidimo da se Model 1 razlikuje od barem jednog statistički značajnog modela na razini od 5%.

Tablica 3.32 govori da se krivulje u Modelu 2 i Modelu 1 te krivulje u Modelu 3 i Modelu 1 statistički razlikuju. Dakle, nije svejedno koji model ćemo uzeti. Kako Model 1 ima najveću površinu ispod ROC krivulje te se statistički razlikuje od ostalih modela, zaključujem da je taj model najbolji.

Bibliografija

[1] B. Dawson, R.G. Trapp, *Basic & Clinical Biostatistic*, Lange, 2004.

[2] M. Gonen, *Receiver Operating Characteristic (ROC) Curves*, Memorial SLoan-Kettering Cancer Center, dostupno na www2.sas.com/proceedings/sugi31/210-31.pdf

[3] A.Jazbec, *Odabrane statističke metode u biomedicini*, PMF-MO, nastavni materijali 2016.

[4] *Kaggle*, https://www.kaggle.com/primaryobjects/voicegender

[5] A.G.Stead, K.G.MacDonald, Organon Teknika Corporation, Durham, NC, *Constructing ROC Curves with the SAS System*, www2.sas.com/proceedings/sugi22/POSTERS/PAPER219.PDF

[6] *ROC Analysis*, dostupno na ://mlwiki.org/index.php/ROC_Analysis

[7] *SAS University Edition*, http://support.sas.com/software/products/university-edition/index.html

[8] *SAS/STAT(R) 9.22 User's Guide: Comparing Receiver Operating Characteristic Curves*, dostupno na https://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_logistic_sect066.htm

[9] *SAS/STAT(R) 9.22 User's Guide: ROC analysis for binary response models fit in the GLIMMIX, NLMIXED, GAM or other procedures*, dostupno na http://support.sas.com/kb/41/364.html

Sažetak

Prediktivni, klasifikacijski modeli se koriste u svakom području znanosti te u mnogo grana poslovnog svijeta. U ovom radu su pokazana dva načina mjerenja točnosti predikcija. Prvi način se koristi kada želimo procjeniti neko binarno ponašanje, tj. kada ishod mogu biti samo dvije opcije (primjerice pozitivan i negativan). Tada želimo ocijeniti koliko je dobar binarni klasifikator. Ocjenjujemo prediktivnu točnost testa pomoću mjera senzitivnosti (stvarno pozitivne stope) i specifičnosti (stvarno negativne stope).

Drugi način se koristi kada su predikcije neprekidne (numeričke) te ih po nekom kriteriju možemo svesti na problem procjene binarnog ponašanja. Takve kriterije nazivamo granična vrijednost testa vrijednosti (pragove) koji će služiti kao granice odjeljivanja pozitivnog ishoda predikcije od negativnog. ROC krivulja je grafički prikaz senzitivnosti u odnosu na 1-specifičnost, tj. lažno pozitivnu stopu za svaku granična vrijednost testa vrijednost. Prediktivni test se smatra točnijim što se bliže nalazi gornjem lijevom kutu, a lošijim što se približava dijagonali koja odgovara slučajnom testu. Mjera za određivanje točnosti testa pomoću ROC krivulje je površina ispod ROC krivulje, gdje površina od 1 odgovara idealnom testu, dok površina od 0.5 odgovara slučajnom testu.

U prvom SAS primjeru je pokazano kako se konstruira ROC krivulja, dok je u drugom primjeru pokazano kako se najbolji logistički model može odabrati usporedbom ROC krivulja.

Summary

Predictive classification models are used by every field of science and almost all business enterprise. This work shows two ways of measuring predictive accuracy. First one is used for estimating binary behaviour (ex. positive or negative). The goal is to estimate the accuracy of a binary classificator. Measurements sensitivity and specificity determine predictive accuracy of the test.

Second one is used when predictions are continuous (numerical). Since we know how to analyze predictive accuracy for a binary predictor, we may consider transforming the predicted probability into a dichotomy by using a threshold (cutoff value). ROC curve plots sensitivity on y-axis and 1-specificity on x-axis for each cutoff value. We consider predictive test to be more accurate when the ROC curve is closer to the upper left corner, and less accurate when the ROC curve is closer to diagonal which represents a random test. Measure for evaluating the performance of a classifier is area under the ROC curve (AUC). AUC = 1 is for a perfect classifier and AUC=0.5 is for a random classifier.

First example shows how to construct ROC curve. Second example shows how to choose the best logistic model by comparing ROC curves.

Životopis

Rođena sam 5.11.1993. godine u Splitu. Završavam OŠ "Jesenice", Orij 2008. godine i upisujem V. gimnaziju "Vladimir Nazor" u Splitu. 2012. godine upisujem sveučilišni preddiplomski studij Matematika na Prirodoslovno-matematičkom fakultetu u Zagrebu. Po završetku preddiplomskog studija 2015. godine upisujem diplomski sveučilišni studij Matematička statistika na istom fakultetu.