

# Vrednotenje napovednih modelov

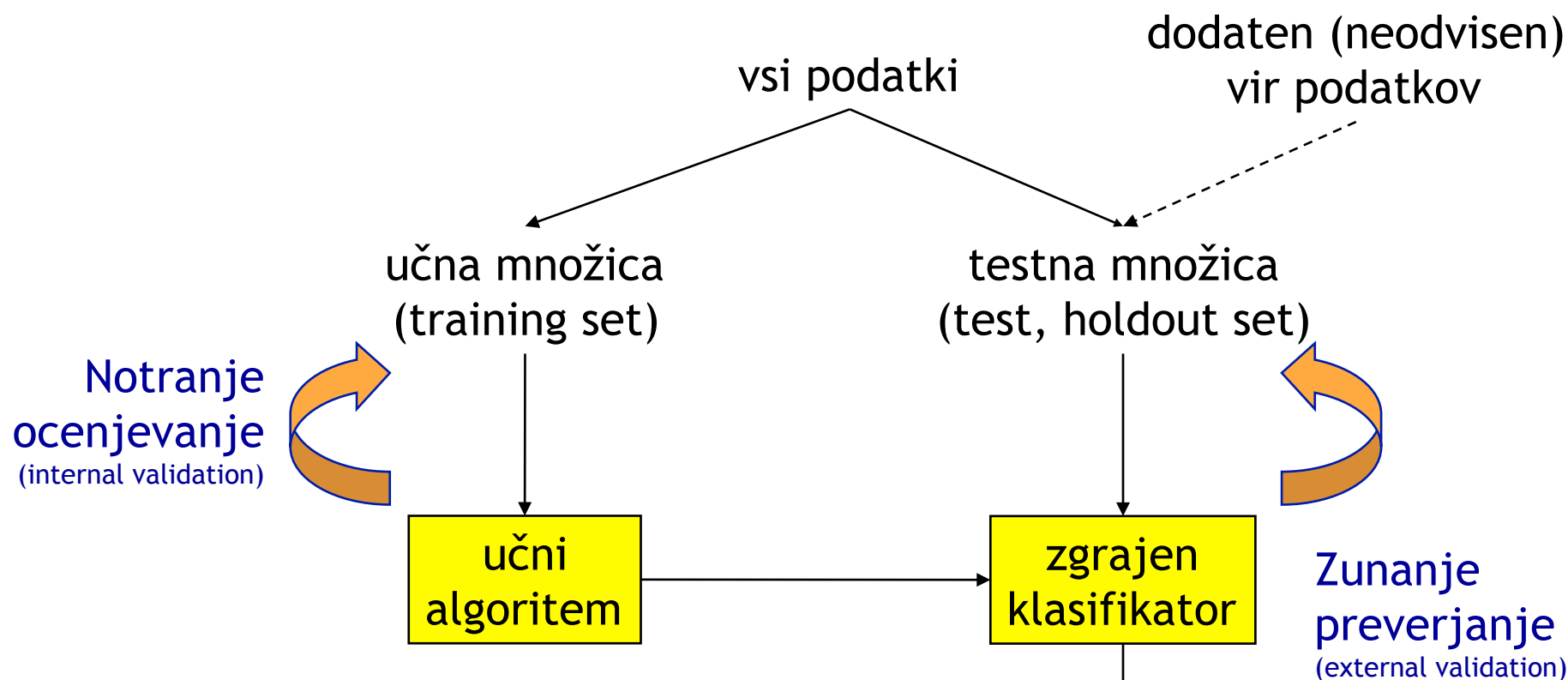
# Vrednotenje modelov

- Kriterija
  - točnost odkritih konceptov (predictive accuracy)
    - točnost (accuracy) = verjetnost pravilne napovedi
    - delež (pogostost) napak (error rate =  $1 - \text{accuracy}$ )
  - razumljivost
- Oba kriterija sta pomembna
  - razumljivost težje merimo...
  - ...zato se osredotočamo na točnost
- Tipi točnosti
  - točnost na učnih podatkih
  - točnost na novih podatkih (dost bolj pomembna!)
  - glavno vprašanje: kako oceniti točnost na novih podatkih

# Izzivi

- Pogosta napaka
  - ocenjevanje pričakovane točnosti na učnih podatkih (**resubstitution** accuracy)
- Velikost podatkov (število primerov)
  - testna množica naj bo reprezentativna (za nove podatke)
  - to ni težava, če je podatkov (primerov) veliko
- Manjši nabori: težava
  - potrebujemo veliko primerov za uspešno učenje
  - potrebujemo veliko primerov za zanesljivo ocenjevanje točnosti

# Postopek ocenjevanja točnosti



Ideja: izmerjena točnost na testnih podatkih je (dober) približek točnosti, ki jo bomo dosegli na novih podatkih

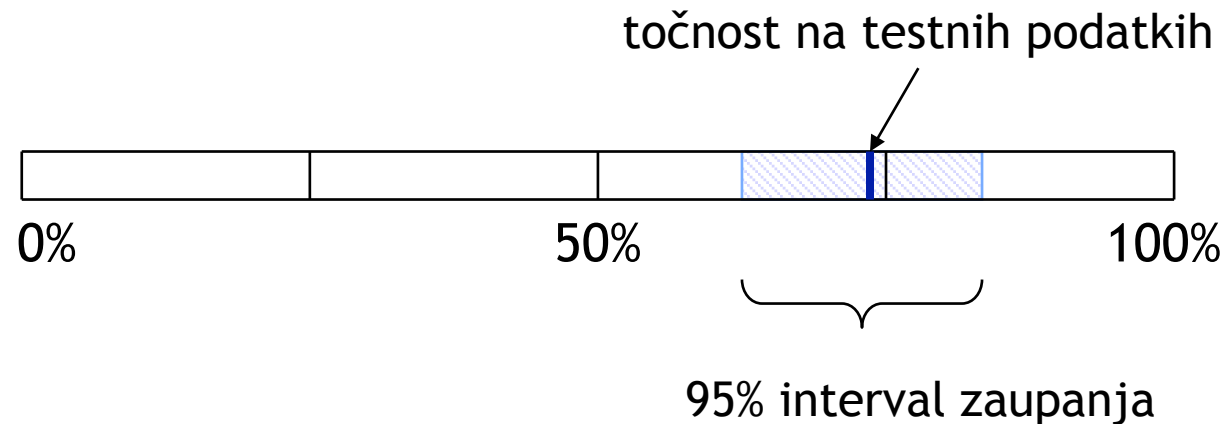
# Ocenjevanje točnosti na testni množici

- Recimo:
  - Klasifikator pravilno napove 73% primerov v testni množici.
  - Pričakujemo lahko 73% točnost na novih primerih:  
Ampak:
    - Kako blizu bo točnost ocenjenim 73%?
    - Koliko zaupamo oceni? (odvisno od velikosti testne množice)

# Intervali zaupanja

(Confidence Intervals)

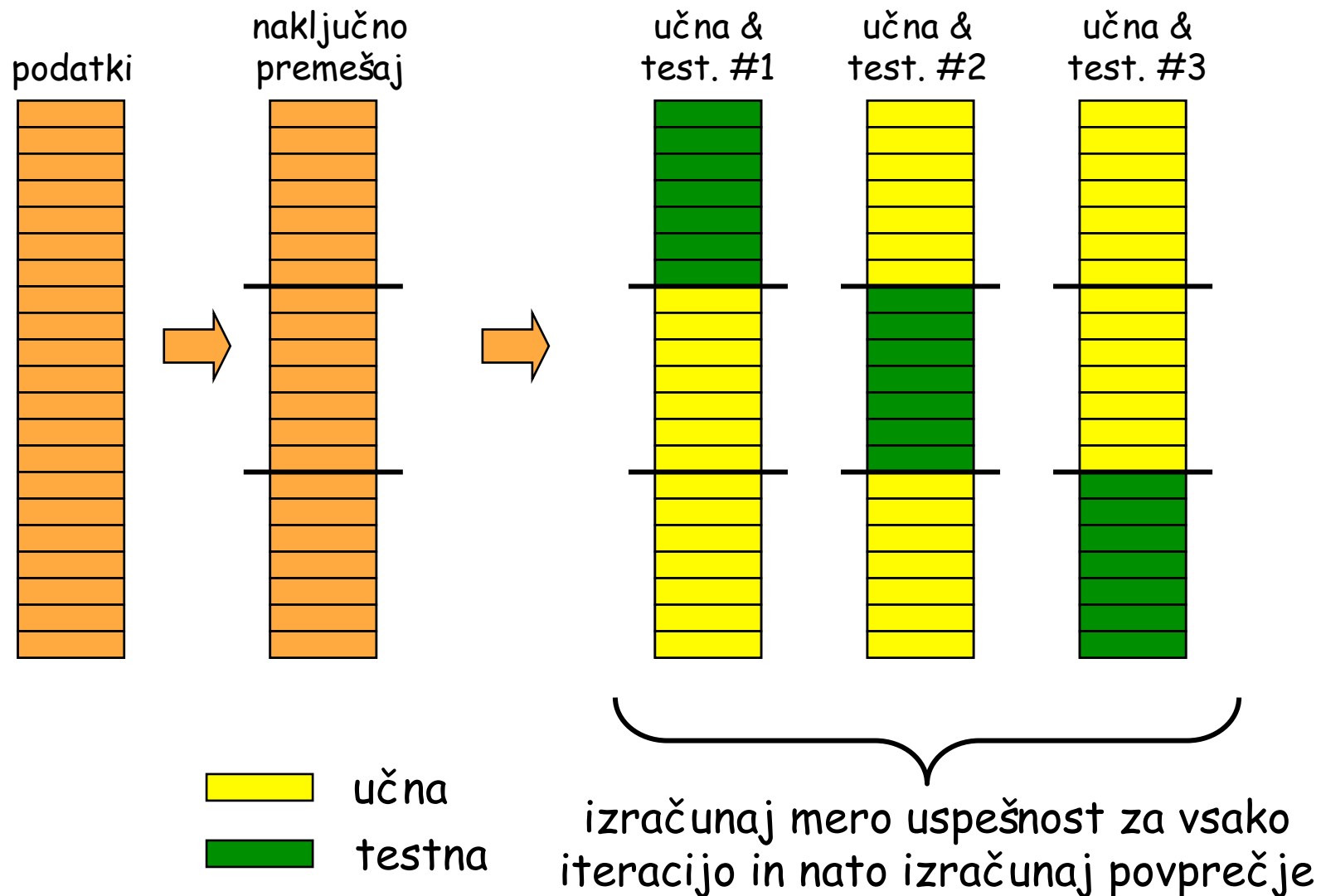
- Uporabimo za določiti intervale zaupanja v ocene točnosti.
- Confidence intervals



# **SHEME vrednotenja (metode vzorčenja)**

# 3-kratno prečno preverjanje

## (3-Fold Cross Validation)





# k-kratno prečno preverjanje

- Podatke razdelimo v k enako velikih podmnožic (zagotoviti je potrebno enako porazdelitev razredov v vsaki)
- For  $i=1$  to  $k$ :
  - množico  $i$  uporabi kot testno, ostalih  $k-1$  množic za učenje
- Izračunaj povprečno točnost.
- Ponovi, recimo, 100-krat.

# Naključno vzorčenje (70/30)

- Podatke naključno razdeli v dve podmnožici:
  - 70% podatkov za učenje
  - 30% podatkov za testiranje
- Model zgradi na učnih podatkih, ovrednoti na testnih podatkih
- Postopek ponovi, recimo, 100-krat in izračunaj povprečno točnost in interval zaupanja.

# Mere kalibracije in razločevanja

(statistics  
calibration  
discrimination)

# Kalibracija in razločevanje

- Kalibracija (umeritev, calibration)
  - kako natančne so verjetnosti, ki jih vrača model
  - klasifikacijska točnost, senзитivnost, specifičnost, ...
- Razločevanje (discrimination)
  - kako dobro model loči med pozitivnimi in negativnimi primeri
  - površina pod krivuljo ROC (area under ROC)

# Testne statistike

## Kontingenčna tabela klasifikacijskih rezultatov (contingency table)

		True Class		Totals
		+	-	
Result from classification model	+	TP	FP	TP+FP
	-	FN	TN	FN+TN
Totals		$P=TP+FN$	$N=FP+TN$	$M$

- **pravilno pozitivni, napačno pozitivni**  
(true positive, false positive)
- **napačno negativni, pravilno negativni**  
(false negative, true negative)

# Klasifikacijska točnost (classification accuracy)

		True Class		Totals
		+	-	
Result from classification model	+	TP	FP	TP+FP
	-	FN	TN	FN+TN
Totals		P=TP+FN	N=FP+TN	M

- $CA = (TP+TN) / (P+N) = (TP+TN) / M$
- Delež pravilno napovedanih primerov.

# Senzitivnost (Sensitivity)

		True Class		
		+	-	Totals
Result from classification model	+	TP	FP	TP+FP
	-	FN	TN	FN+TN
Totals		TP+FN	FP+TN	M

- Sensitivity =  $TP / (TP + FN) = TP / P$
- Delež pravilno napovedanih pozitivnih primerov.
- Medicina (+, -: prisotnost, odsotnost bolezni):
  - verjetnost, da bo model detektiral pacienta z boleznijo

# Specifičnost (specificity)

		True Class		
		+	-	Totals
Result from classification model	+	TP	FP	TP+FP
	-	FN	TN	FN+TN
Totals		TP+FN	FP+TN	M

- $\text{Specificity} = \text{TN} / (\text{FP} + \text{TN}) = \text{TN} / N$
- Delež pravilno napovedanih negativnih primerov.
- Medicina:
  - verjetnost, da bo model pravilno napovedal, da pacient nima bolezni



# Ostale statistike

## Card 2A **DIAGNOSIS**

### Are the results of this diagnostic study valid?

1. Was there an independent, blind comparison with a reference ("gold") standard of diagnosis?
2. Was the diagnostic test evaluated in an appropriate spectrum of patients (like those in whom it would be used in practice)?
3. Was the reference standard applied regardless of the diagnostic test result?
4. Was the test (or cluster of tests) validated in a second, independent group of patients?

### Are the valid results of this diagnostic study important?

		Target disorder (iron deficiency anaemia)		Totals
		Present	Absent	
Diagnostic test result (serum ferritin)	Positive ( $< 65$ mmol/L)	731 a	270 b	1001 a+b
	Negative ( $\geq 65$ mmol/L)	78 c	1500 d	1578 c+d
Totals		809 a+c	1770 b+d	2579 a+b+c+d

Sensitivity =  $a/(a + c) = 731/809 = 90\%$ .

Specificity =  $d/(b + d) = 1500/1770 = 85\%$ .

Likelihood ratio for a positive test result =

$$LR+ = \text{sens}/(1 - \text{spec}) = 90\%/15\% = 6.$$

Likelihood ratio for a negative test result =

$$LR- = (1 - \text{sens})/\text{spec} = 10\%/85\% = 0.12.$$

Positive predictive value =  $a/(a + b) = 731/1001 = 73\%$ .

Negative predictive value =  $d/(c + d) = 1500/1578 = 95\%$ .

Pre-test probability (prevalence) =  $(a + c)/(a + b + c + d)$   
 $= 809/2579 = 32\%$ .

Pre-test-odds = prevalence/(1 - prevalence) =  $31\%/69\% = 0.45$ .

Post-test odds = Pre-test odds  $\times$  likelihood ratio.

Post-test probability = post-test odds/(post-test odds + 1).

### Can we apply this valid, important evidence about a diagnostic test in caring for our patient?

1. Is the diagnostic test available, affordable, accurate, and precise in our setting?
2. Can we generate a clinically sensible estimate of our patient's pre-test probability (from personal experience, prevalence statistics, practice databases, or primary studies)?
  - Are the study patients similar to our own?
  - Is it unlikely that the disease possibilities or probabilities have changed since this evidence was gathered?
3. Will the resulting post-test probabilities affect our management and help our patient?
  - Could it move us across a test-treatment threshold?
  - Would our patient be a willing partner in carrying it out?

From Sackett, Straus, Richardson, Rosenberg, Haynes: Evidence-Based Medicine; How to Practice and Teach EBM. Harcourt, Edinburgh, 2000

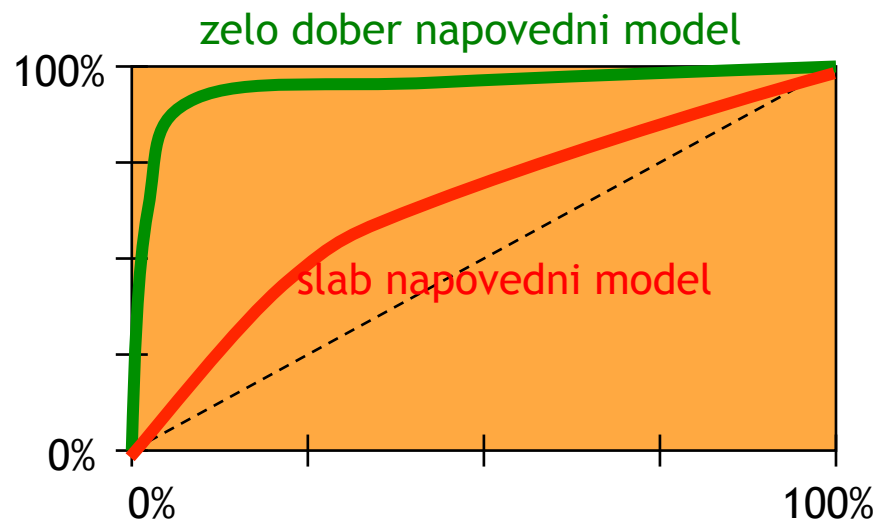
DL Sackett et al.: Evidence-Based Medicine, Churchill-Livingstone, 2000.

# Krivulja ROC (ROC Curves)

- ROC = Receiver Operating Characteristics
- Že od 70-ih uporabljena v medicinski prognostiki
- Uporabljena tudi v strojnem učenju.

$$TPrate = \frac{TP}{P}$$

sensitivity  
[TP rate]



1-specificity  
[FP rate]

$$FPrate = \frac{FP}{N}$$

# Krivulja ROC

**Qt Classifications**

Options

☒ Show predicted probabilities

yes  
no

☐ Show predicted class

☒ Show true class

Data Attributes

☐ Show all

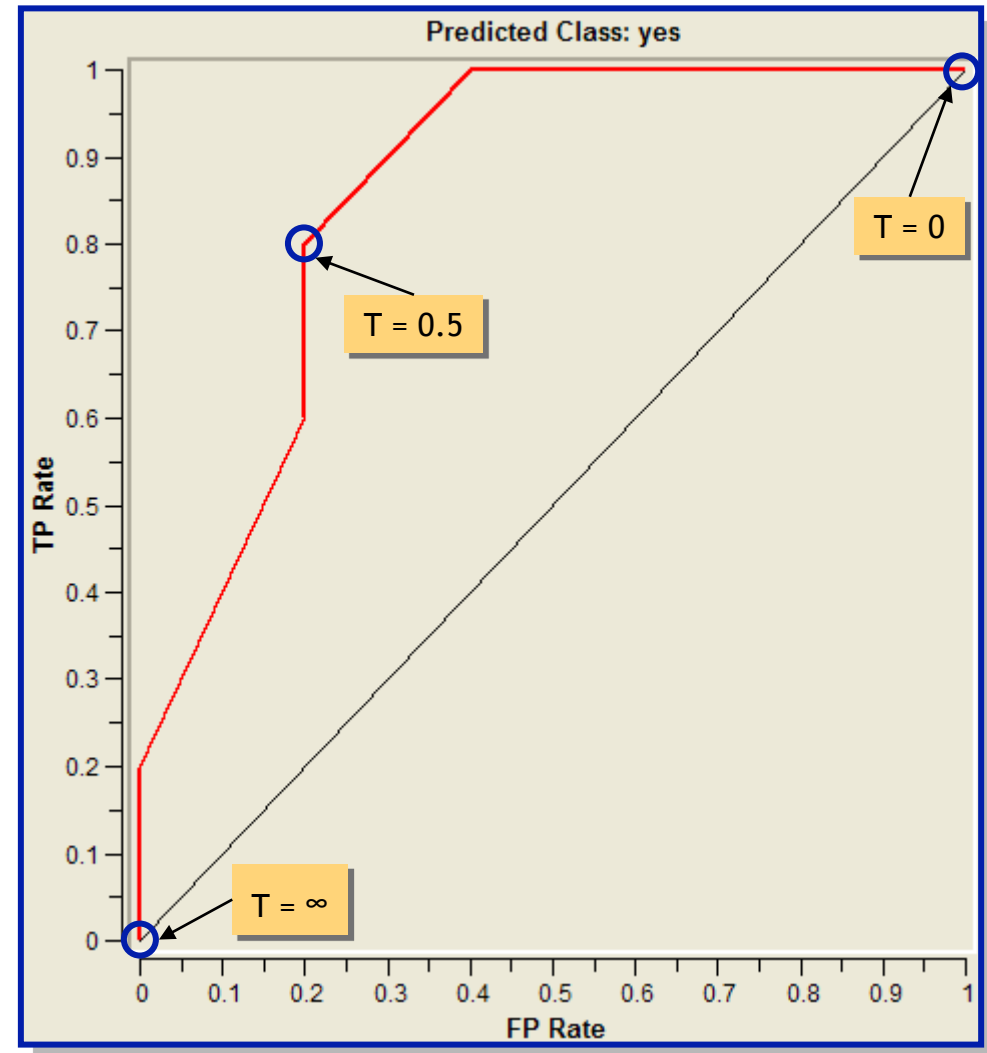
☒ Hide all

	Sail	Naive Bayes
1	yes	0.800000
2	yes	0.888889
3	no	0.333333
4	yes	0.800000
5	no	0.097776
6	no	0.800000
7	yes	0.634229
8	yes	0.333333
9	no	0.097776
10	no	0.097776

Class	P(yes)	$T = \infty$	$T = 0$	$T = 0.5$
yes	0.89	no	yes	yes
yes	0.80	no	yes	yes
yes	0.80	no	yes	yes
no	0.80	no	yes	yes
yes	0.63	no	yes	yes
no	0.33	no	yes	no
yes	0.33	no	yes	no
no	0.10	no	yes	no
no	0.10	no	yes	no
no	0.10	no	yes	no

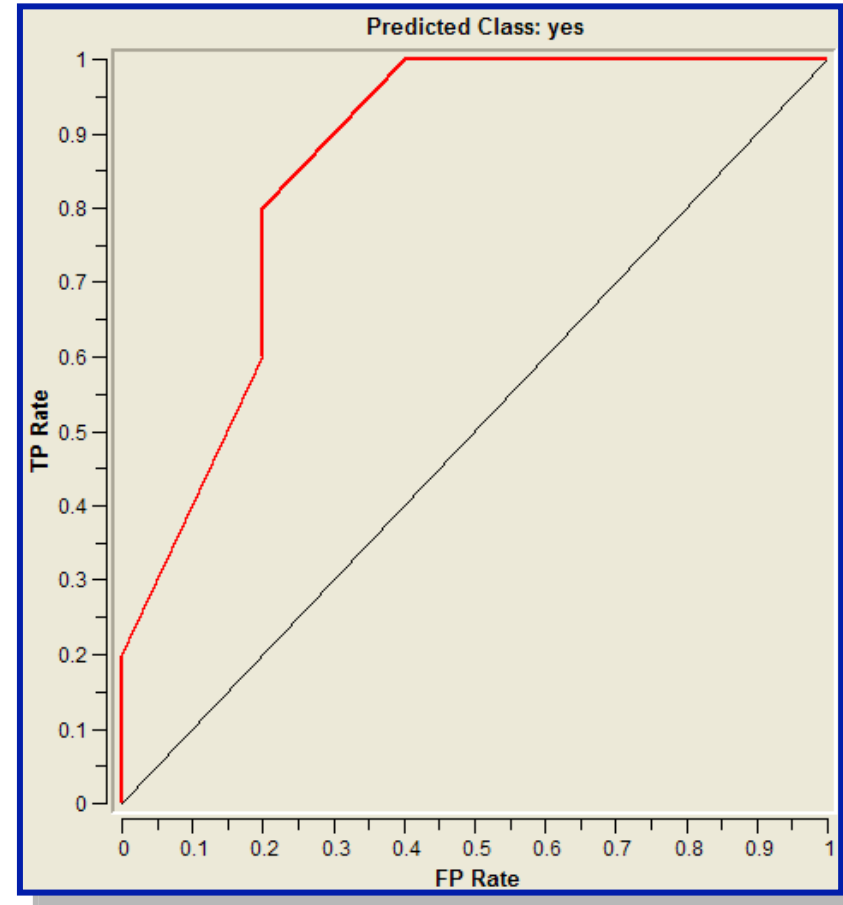
$$TPrate = \frac{TP}{P}$$

$$FPrate = \frac{FP}{N}$$



# Kako narisati krivuljo ROC

1. Narišemo mrežo:
  - horizontalni korak  $1/N$
  - vertikalni korak  $1/P$
2. Napovedi uredimo po padajoči verjetnost (izbranega razreda)
3. Začnemo v točki (0,0)
4. Iz tabele izberemo vrhnje vrstice z največjo verjetnostjo.
5. Če izbrane vrstice vsebujejo p pozitivnih in n negativnih primerov, potem se premaknemo za p mest gor, in za n mest desno.
6. Odstranimo prej izbrane vrstice.
7. Če tabela ni prazna, pojdi v točko 4.

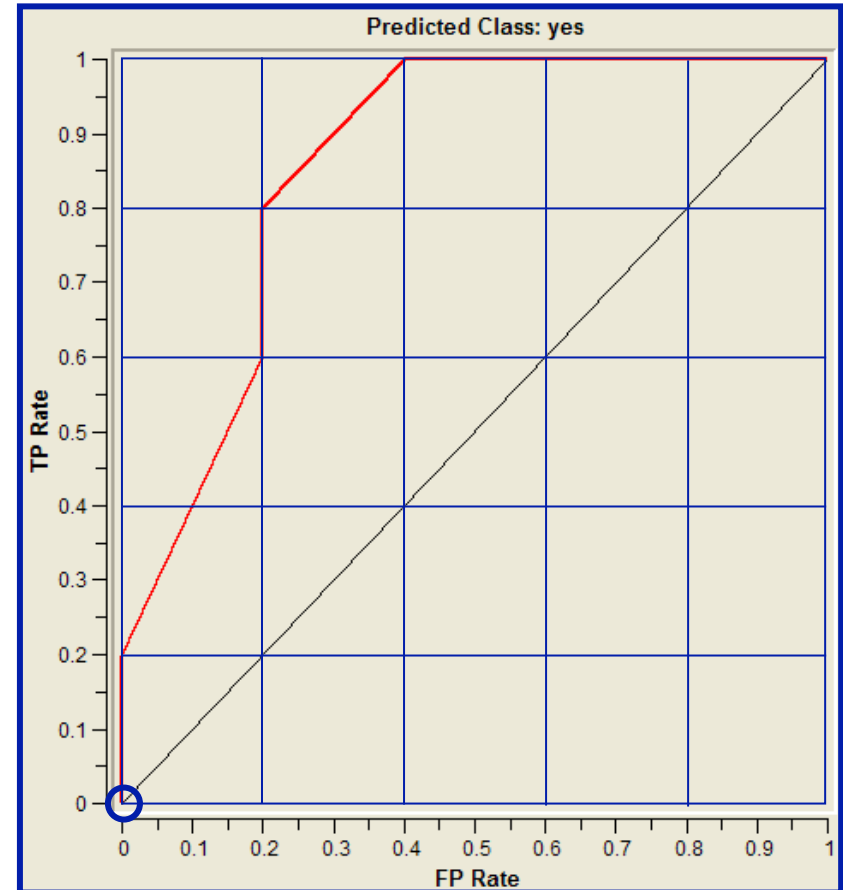


$$TPrate = \frac{TP}{P}$$

$$FPrate = \frac{FP}{N}$$

# Kako narisati krivuljo ROC

Class	P(yes)
yes	0.89
yes	0.80
yes	0.80
no	0.80
yes	0.63
no	0.33
yes	0.33
no	0.10
no	0.10
no	0.10

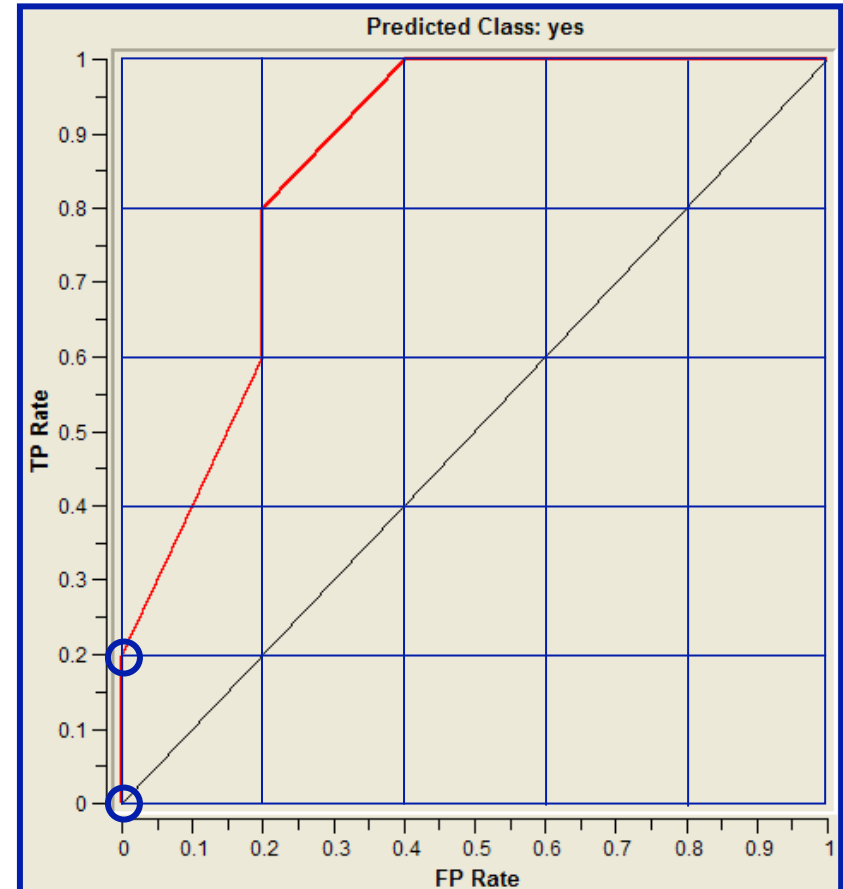


$$TPrate = \frac{TP}{P}$$

$$FPrate = \frac{FP}{N}$$

# Kako narisati krivuljo ROC

Class	P(yes)
yes	0.89
yes	0.80
yes	0.80
no	0.80
yes	0.63
no	0.33
yes	0.33
no	0.10
no	0.10
no	0.10

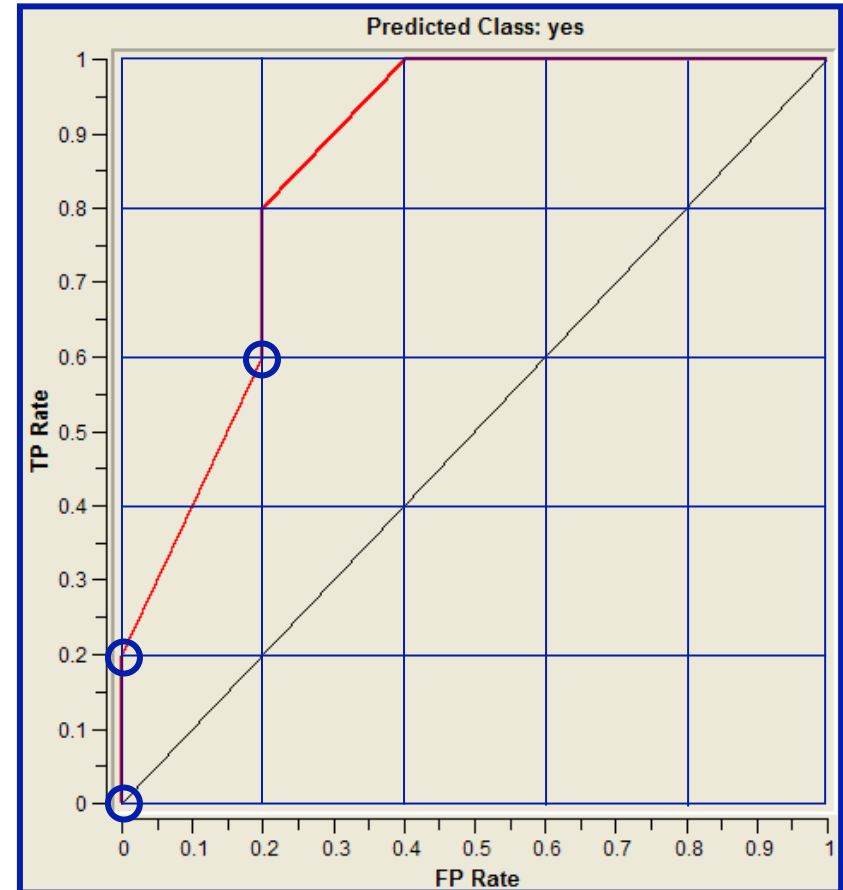


$$TPrate = \frac{TP}{P}$$

$$FPrate = \frac{FP}{N}$$

# Kako narisati krivuljo ROC

Class	P(yes)
yes	0.89
yes	0.80
yes	0.80
no	0.80
yes	0.63
no	0.33
yes	0.33
no	0.10
no	0.10
no	0.10

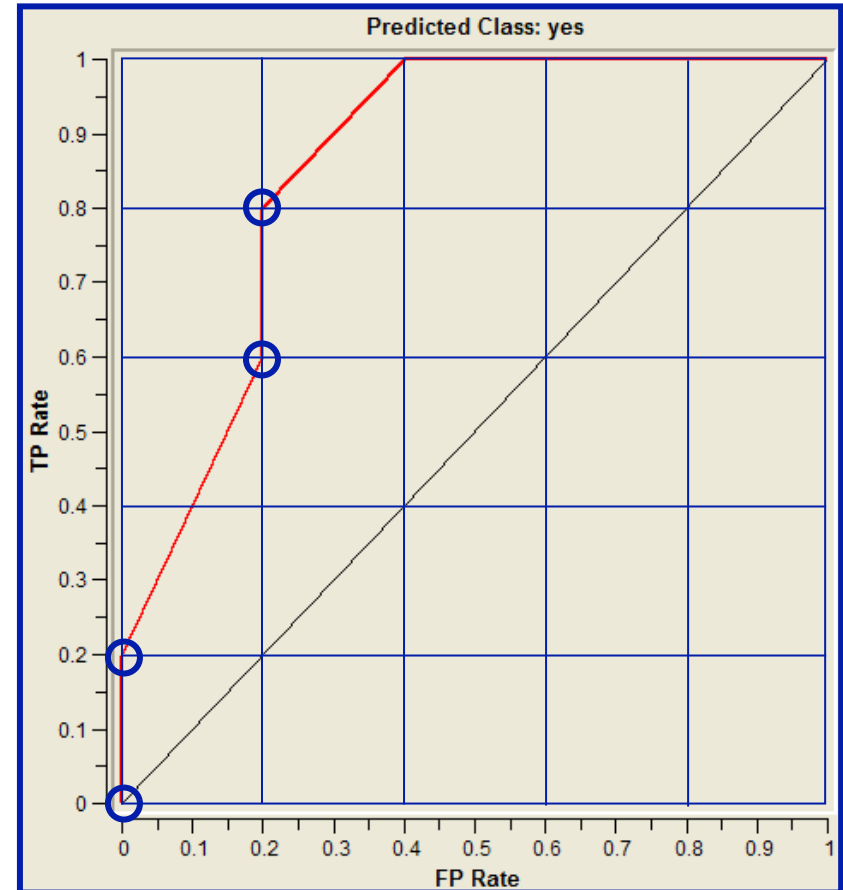


$$TPrate = \frac{TP}{P}$$

$$FPrate = \frac{FP}{N}$$

# Kako narisati krivuljo ROC

Class	P(yes)
yes	0.89
yes	0.80
yes	0.80
no	0.80
yes	0.63
no	0.33
yes	0.33
no	0.10
no	0.10
no	0.10



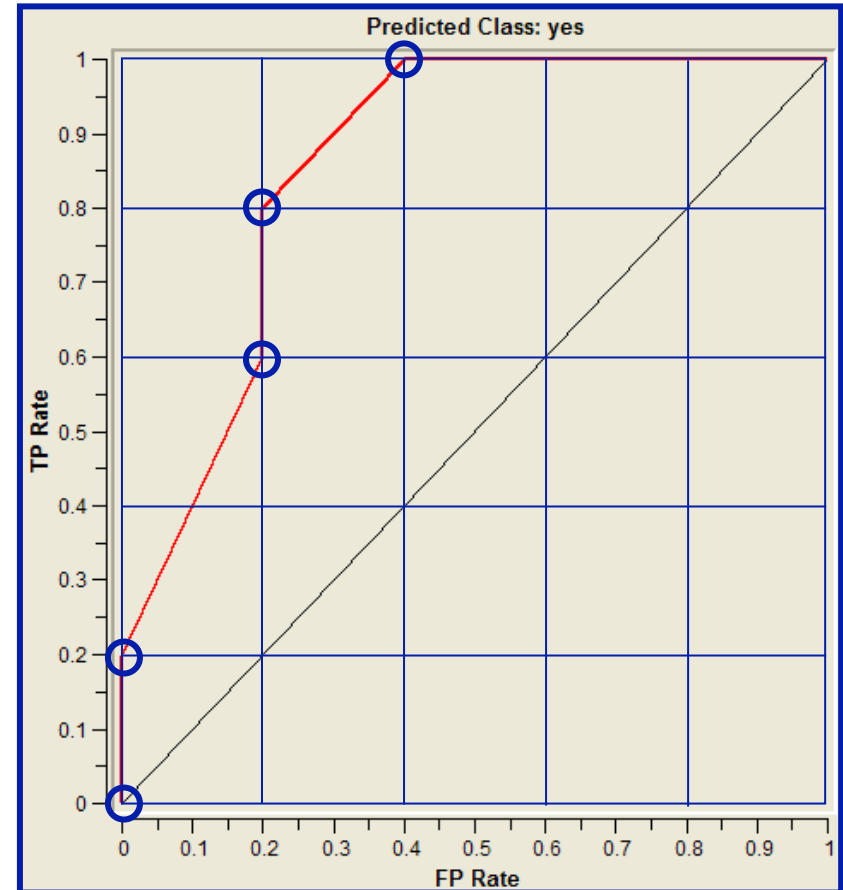
$$TPrate = \frac{TP}{P}$$

$$FPrate = \frac{FP}{N}$$



# Kako narisati krivuljo ROC

Class	P(yes)
yes	0.89
yes	0.80
yes	0.80
no	0.80
yes	0.63
no	0.33
yes	0.33
no	0.10
no	0.10
no	0.10

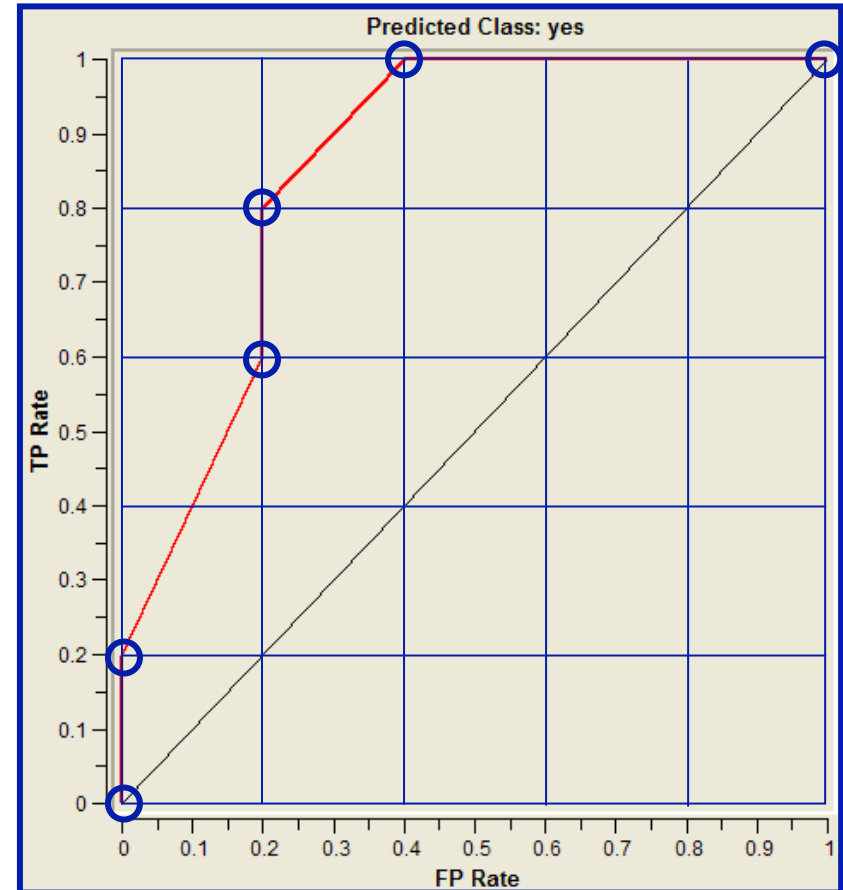


$$TPrate = \frac{TP}{P}$$

$$FPrate = \frac{FP}{N}$$

# Kako narisati krivuljo ROC

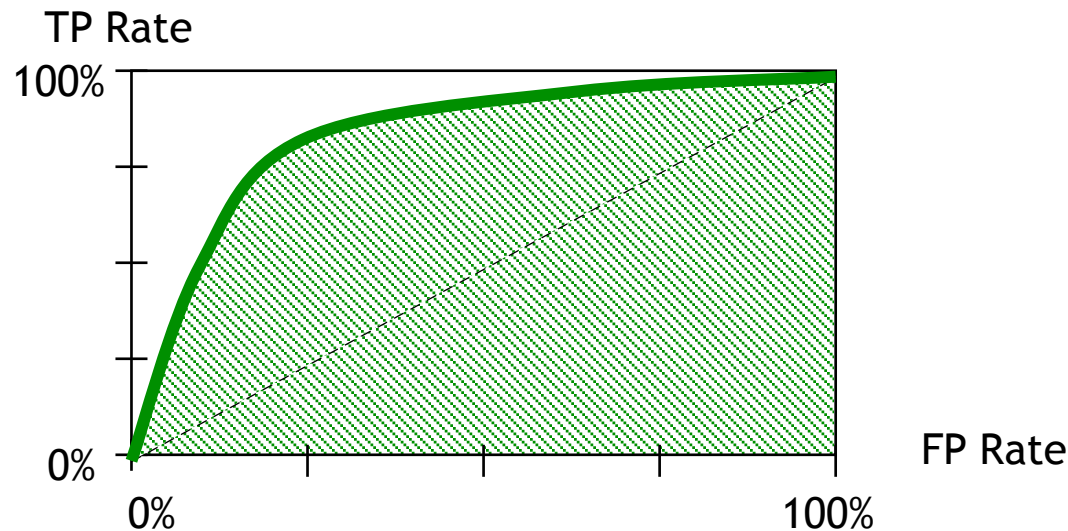
Class	P(yes)
yes	0.89
yes	0.80
yes	0.80
no	0.80
yes	0.63
no	0.33
yes	0.33
no	0.10
no	0.10
no	0.10



$$TPrate = \frac{TP}{P}$$

$$FPrate = \frac{FP}{N}$$

# Površina pod krivuljo ROC (area under ROC)

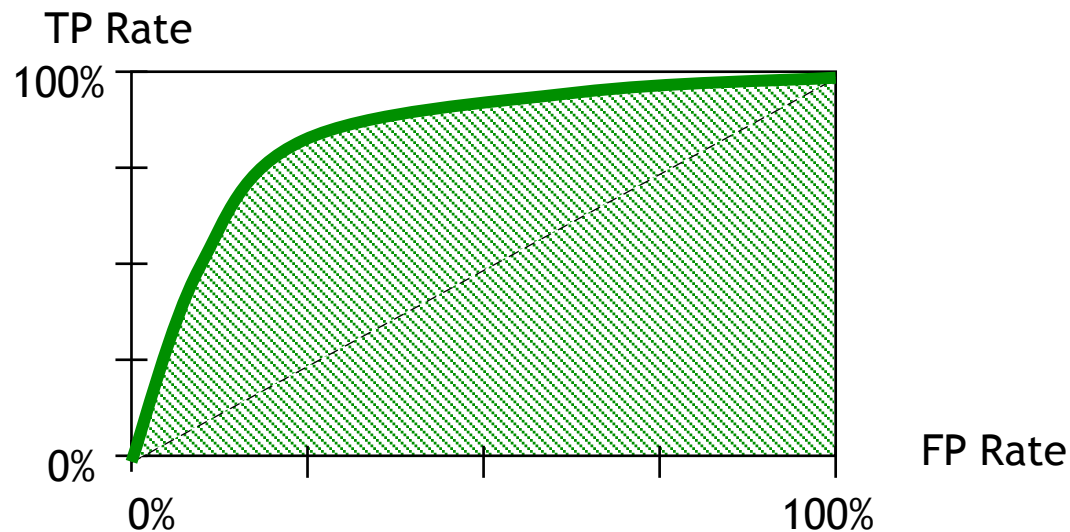


$$A_{ROC} = \int_0^1 \frac{TP}{P} d\frac{FP}{N} = \frac{1}{PN} \int_0^N TP dFP$$

Za vsak negativen primer, seštej število pozitivnih primerov z večjo oceno (napovedano verjetnost). Vsoto normaliziraj s produktom števila pozitivnih in negativnih primerov.

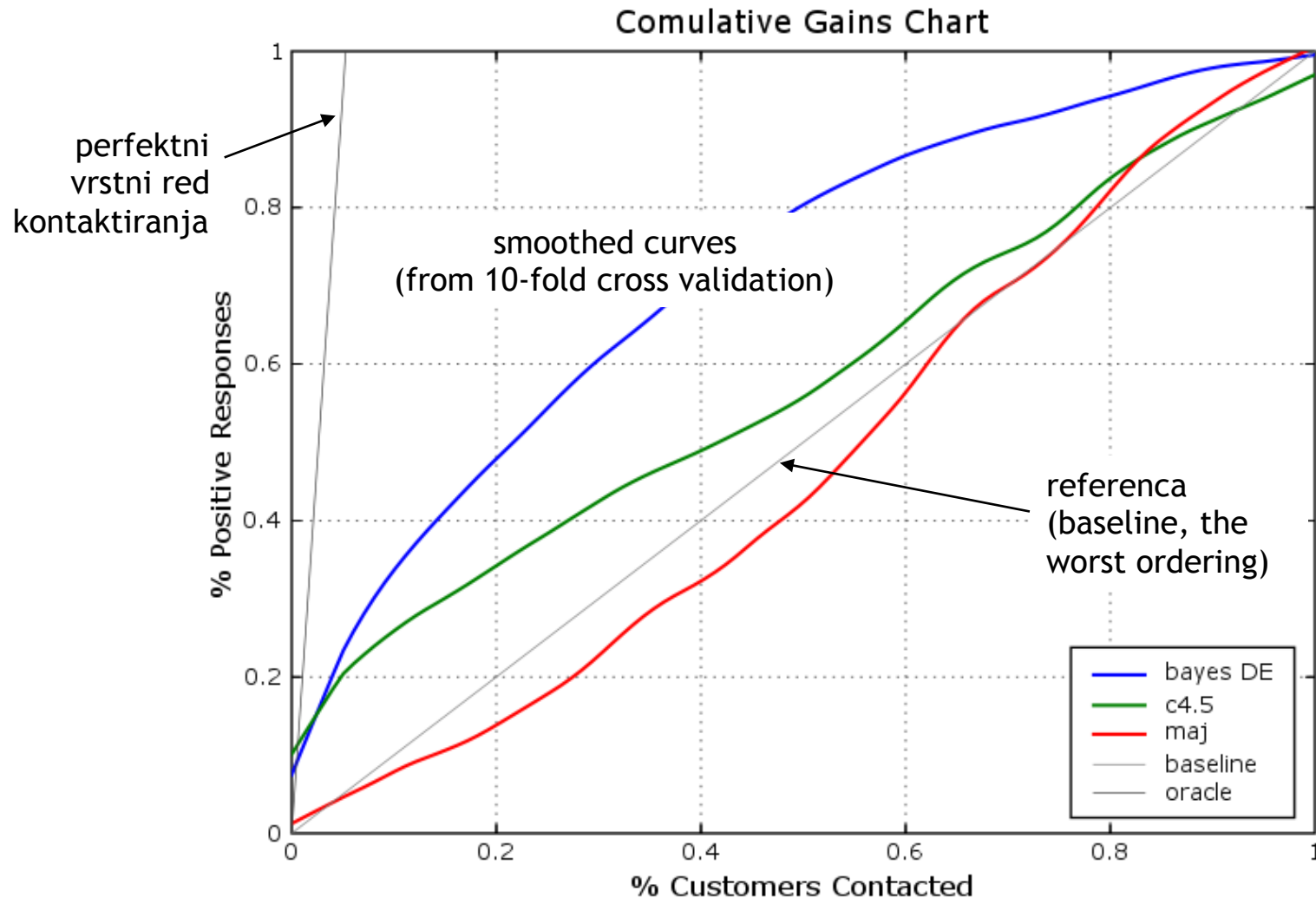
$$A_{ROC} = P [ P^+(\text{pozitivnega primera}) > P^+(\text{negativnega primera}) ]$$

# Površina pod krivuljo ROC (area under ROC)



- V obsegu od 0.5 to 1.0.
- Ocena je neodvisna od porazdelitve razredov.
- Tipični pragi:
  - 0.5: naključne napovedi (random classifier)
  - pod 0.7: slab klasifikator
  - 0.7 do 0.8: ok
  - 0.8 to 0.9: zelo dobri napovedni modeli

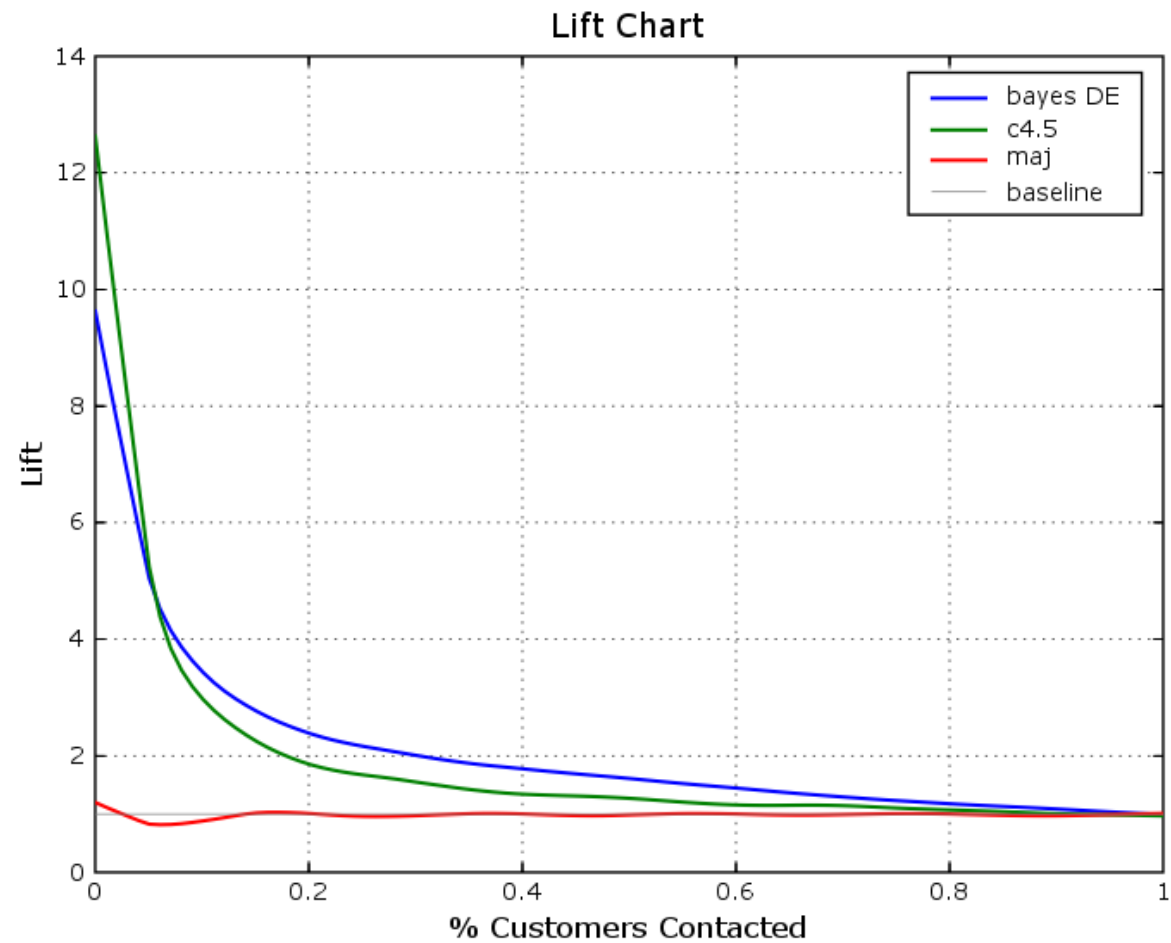
# Cumulative Gains Chart



# Cumulative Gains Chart

- os y: delež pozitivnih odzivov (od vseh možnih pozitivnih odzivov)
- os x: delež kontaktiranih strank (od vseh strank, ki jih imamo)
- Referenca (celoten odziv): če kontaktiramo X% strank, se jih Y% odzove pozitivno

# Dvižna krivulja (lift chart)



# Dvižna krivulja

## (lift chart)

- Uporabimo napovedi modela za določitev vrstnega reda kontaktiranja strank. Za vsak delež kontaktiranih strank (os x), izračunamo delež pozitivnih odgovorov.
- Položaj točk na osi y izračunamo tako, da delež pozitivnih odgovorov, ki smo jih dobili z modelom, delimo s pričakovanim deležem pozitivnih odgovorov, kar bi dobili brez modela (naključni vrstni red kontaktiranja strank).
- Primer: kontaktiramo 10% strank. Če stranke izberemo naključno, dobimo 10% odziv. Če uporabimo napovedni model, je odziv 30%. Torej,  $\text{lift} = 30 / 10 = 3$ .



# Zaključne misli

- Nikoli ne zaupaj točnosti doseženi na učni množici.
- Testno množico pridobi z vzorčenjem.
- Ovrednoti oboje:
  - napovedno sposobnost (točnost)
  - novo znanje, ki je zapisano v modelu
- Zaključek: model je dober, če je uporaben v praksi.