



# Logistisk regression

Klassificering



# Agenda

- Typer av "Learning"
- Klassificering
- Logistisk regression - Vad & Hur
- Typer
- Exempel på användning
- Mätvärden
  - True or False
  - Accuracy
  - Confusion Matrix
  - Precision & Recall
  - Cumulative Accuracy Profile & Analysis

# Typer av "Learning"

- **Supervised** (Övervakat)

- Data har "svar" s.k. labelled data.
- Klar input och output
- **Regression** - kontinuerliga värden
- **Klassifikation** - kategorier

- **Unsupervised** (Oövervakat)

- Inga klara "svar"
- Modellen ska hitta mönster/likheter
- **Clustering** - grupper (stora mängder data)

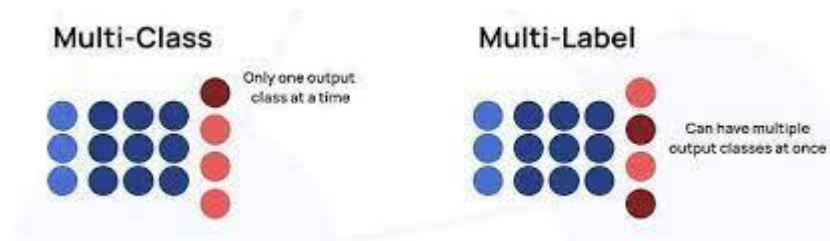
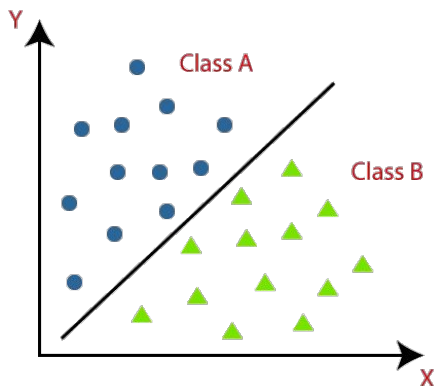


- **Reinforcement** (Förstärkning)

- Belöning och straff (om den gör det "rätta" eller ej)
- Ingen specifik data eller instruktioner
- Miljö som en "agent" interagerar med
  - Bil
  - Spel

# Klassifikation

- Begränsat antal alternativ
- **Binär** - Linjär
  - Endast två val
  - Logistisk regression
  - SVM



- Multi class
  - > 2 möjliga värden
  - ex. siffror
- **Multi label**
  - En samling binära värden
  - Artikel - psykologi och plats
  - Naïve Bayes
- **Multi output**
  - Multi class + multi label
  - TV + drama / Film + drama



# Logistisk regression



# Linjär regression - påminnelse

Oberoende variabel (en eller flera) påverkar en annan (beroende) variabel.

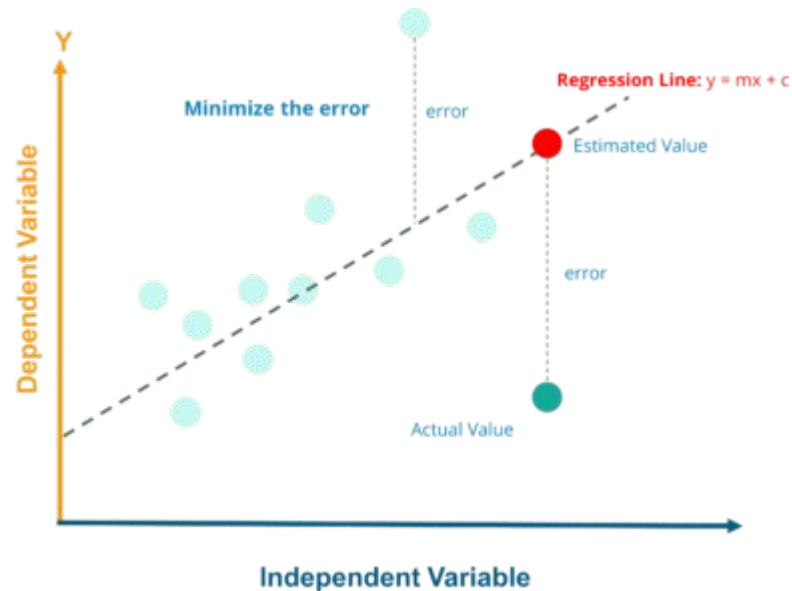
Summerad påverkan av olika variabler.

Linjärt samband

Anpassa en linje

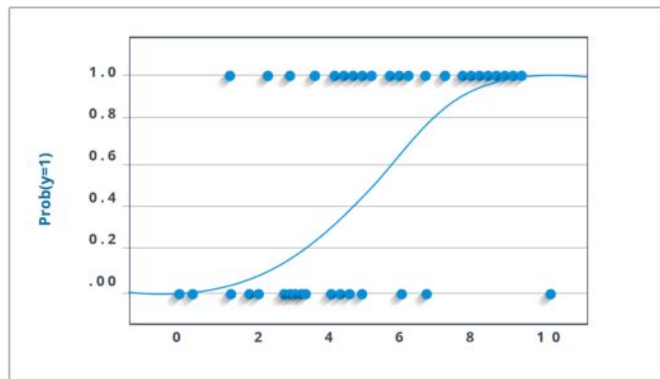
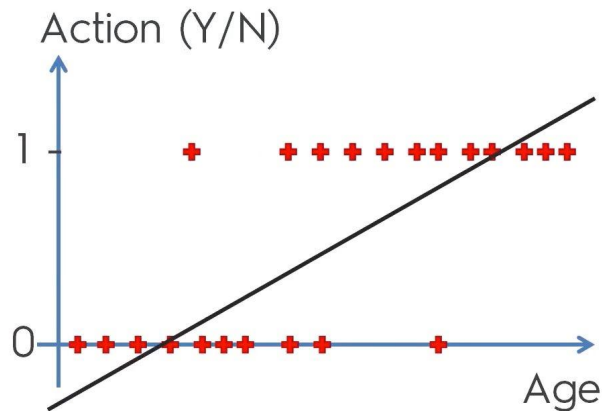
Minimera felmarginalen (error)

Kontinuerliga värden

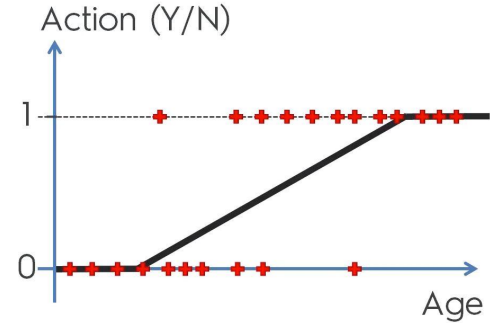
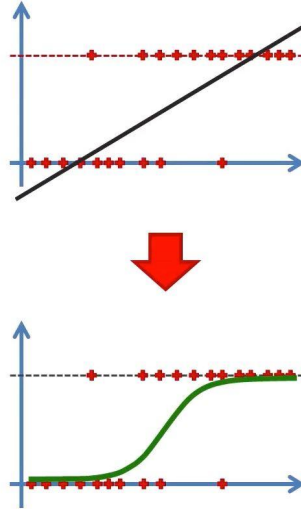
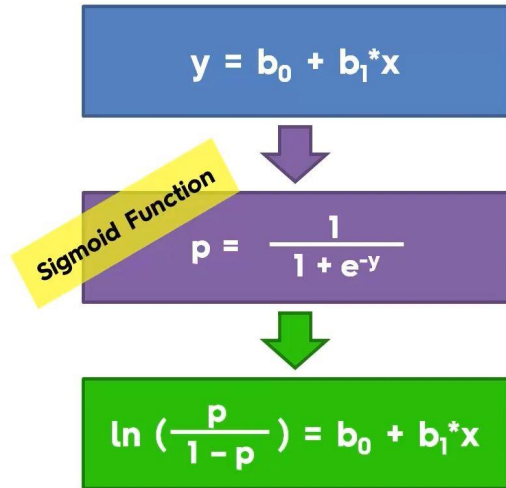


# Logistisk regression - Vad

- En klassificeringsalgoritm.
- "Omvandlar svar från den linjära regressionsmodellen till klasser." [\[Ref.\]](#)
- Diskreta värden - specifikt
- Räknar ut sannolikheten för varje värde

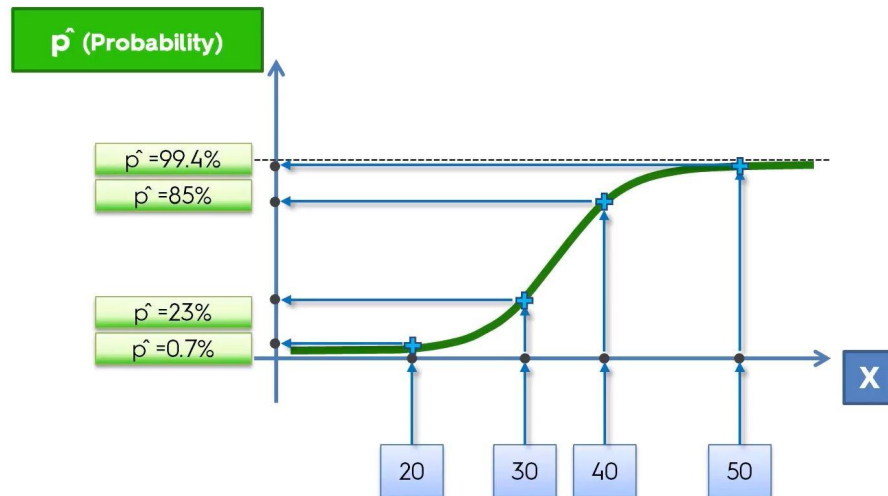
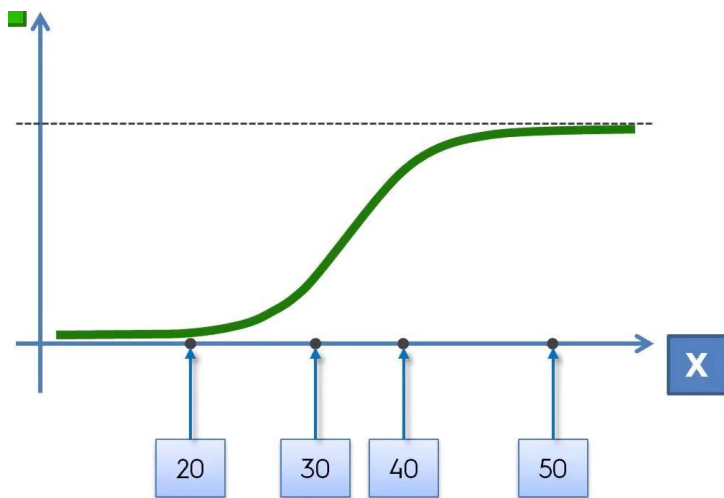


# Anpassning - Sigmoid funktionen





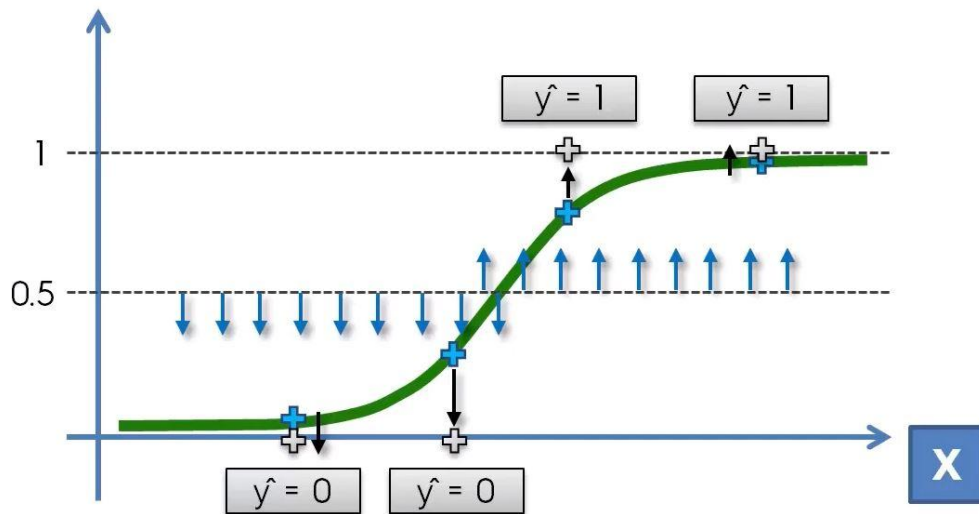
# Sannolikhet - tolkning



RESTORY...

# Logistisk regression - tröskel

Tröskelvärde avgör om vi modellerar ja eller nej



# Logistisk regression - Typer

- **Binär** - endast två val
  - ex. ja eller nej
- **Multinomial** - flera nominella kategorier
  - ex. katt eller hund eller dolfen
- **Ordinal** - ordningskategorier
  - bra, bättre, usch

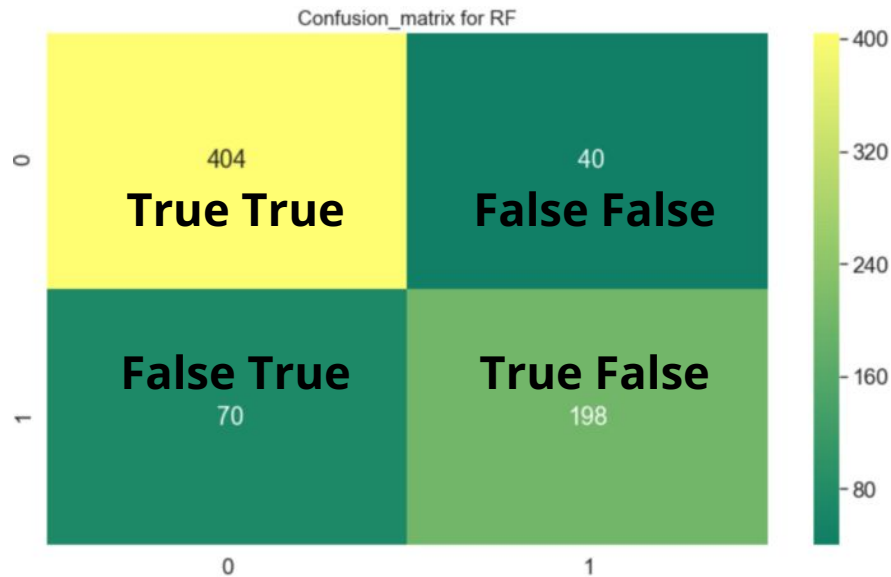
# Logistisk regression - Användning

- Väderlek
  - Regn eller ej
  - Regn, blötsnö, snö.
- Sjukdom

# True or False

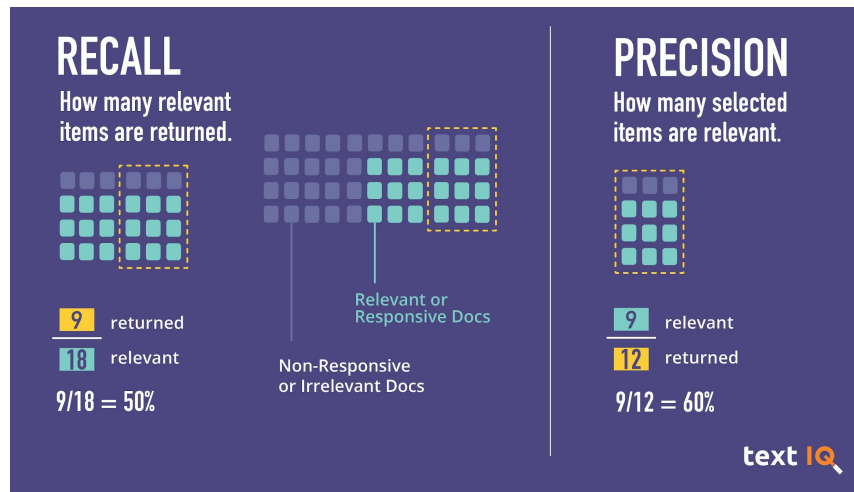
Predicted value vs actual value

	Prediction	Actual
<b>True positive</b>	True	True
<b>False positive</b>	True	False
<b>False negative</b>	False	True
<b>True negative</b>	False	False



# Mätvärden

- **Precision** score
  - Hur exakt modellen är
  - Hur många "svar" av de som förutspåddes är relevanta ("true").
  - Anger hur bra modellen är på att undvika "*false positives*".
  - Correct positive guesses  
Total positive guesses
- **Recall** - andel av de relevanta "svaren" som modellen förutspådde
  - Correct positive cases  
All positive labels



# Mätvärden - precision & recall

Dokumentations  
klassificering

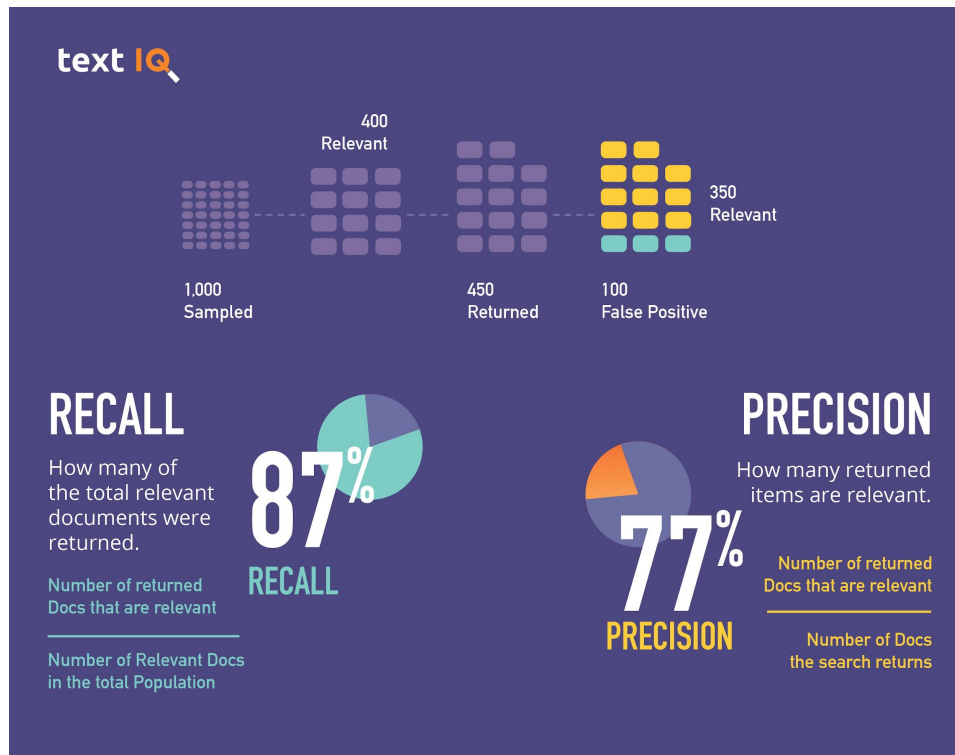


Bild ref. Measure ai

RESTORY...

# Mätvärden - F1 score

Maximera både precisionen och “recall”.

Mäter klassificeringsprestandan.

Tar inte i åtanke True negatives eller multiklass scenariön.

$$F_1 = \frac{2 \cdot \textit{precision} \cdot \textit{recall}}{\textit{precision} + \textit{recall}} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$$



# Mätvärden - exempel

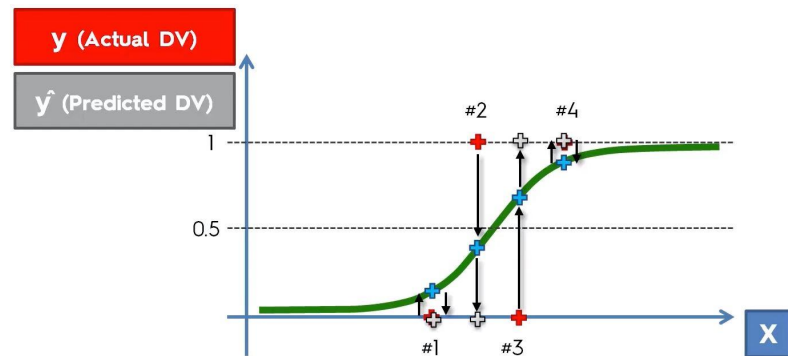
Accuracy (Noggrannhet)

$$(35+50) / 100 = 0.85 \Rightarrow 85 \%$$

		$\hat{y}$ (Predicted DV)	
		0	1
$y$ (Actual DV)	0	35	5
	1	10	50

False Positive (Type I Error) points to the cell (Actual DV=0, Predicted DV=1) with value 5.



False Negative (Type II Error) points to the cell (Actual DV=1, Predicted DV=0) with value 10.



# Mätvärden - Accuracy Paradoxen

Bias i data

Klassificerar alla i den större gruppen

		$\hat{y}$ (Predicted DV)	
		0	1
y (Actual DV)	0	9,850 ← 0 	
	1	150 ← 0 	

## Scenario 1:

Accuracy Rate = Correct / Total  
AR = 9,800/10,000 = 98%

## Scenario 2:

Accuracy Rate = Correct / Total  
AR = 9,850/10,000 = 98.5%

**RESTOR\...**

# Mätvärden - Cumulative Accuracy Profile

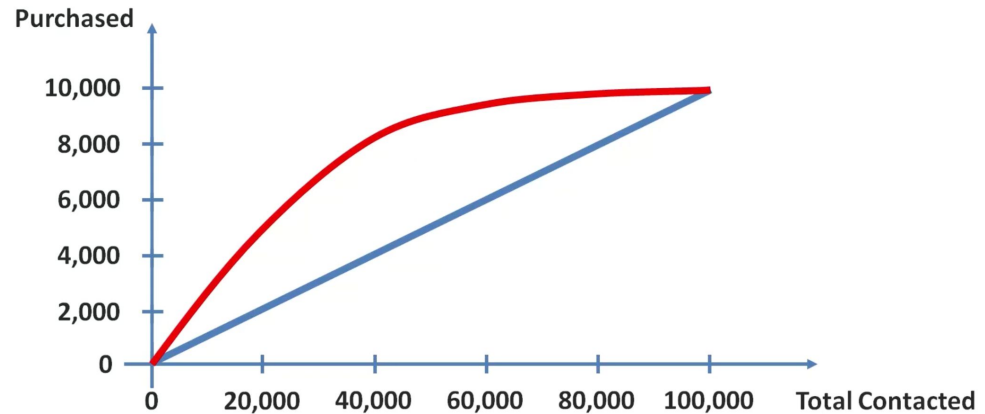
Utskick till 100K random personer.

Datainsamling av värden på de som köpte - profil.

Skicka ut endast till de som matchar profilen.

Ny linje av de som köpte - logistisk regression.

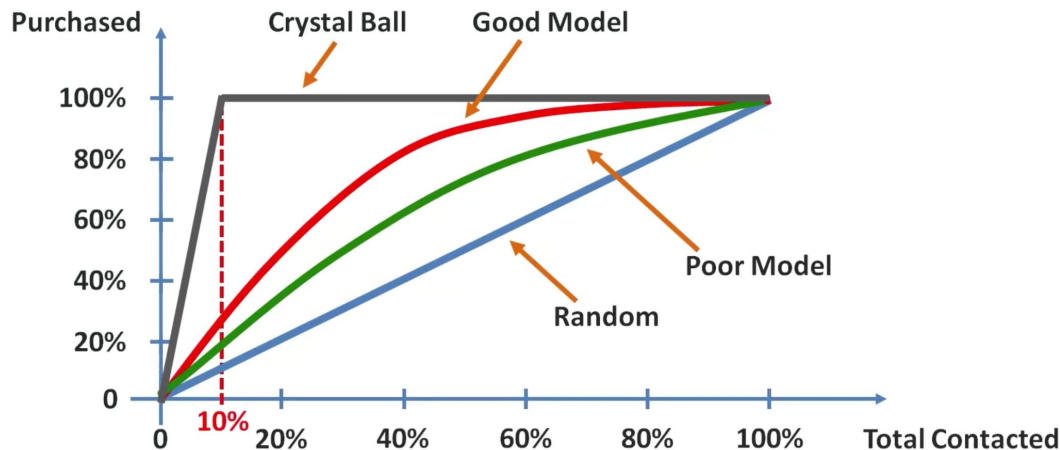
Sannolikhet att de köper produkten.



# Mätvärden - Cumulative Accuracy Profile

Jämför med en annan modell.

Jämför med den *perfekta* modellen.



# Mätvärden - CAP Analysis

Accuracy Ratio (AR) -  
Förhållandet mellan vår  
modell och den "perfekta"  
modellen.

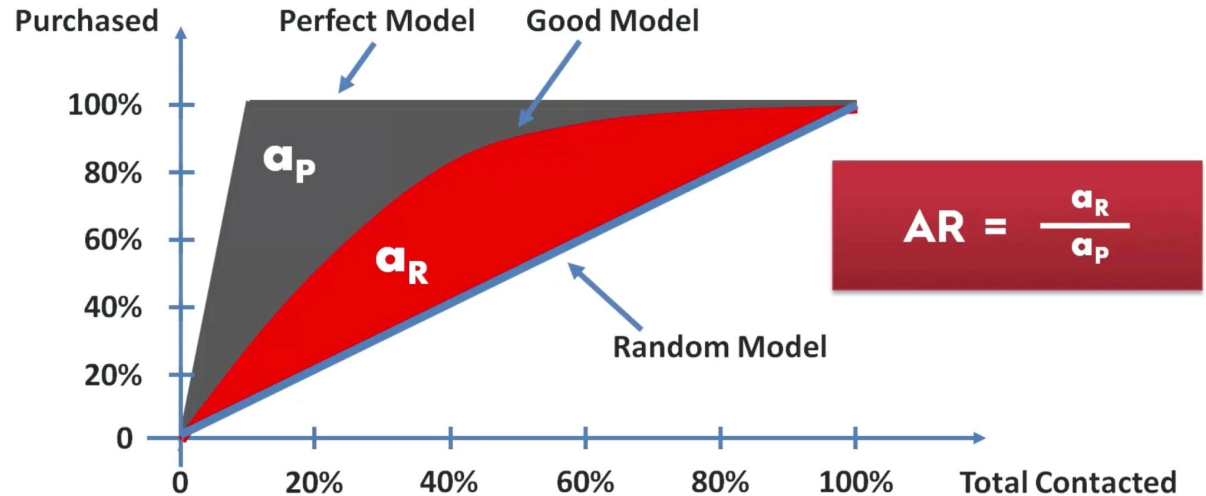
Jobbigt att räkna ut!

Påminner om R2?

$$SS_{\text{res}} = \text{SUM } (y_i - \hat{y}_i)^2$$

$$SS_{\text{tot}} = \text{SUM } (y_i - y_{\text{avg}})^2$$

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

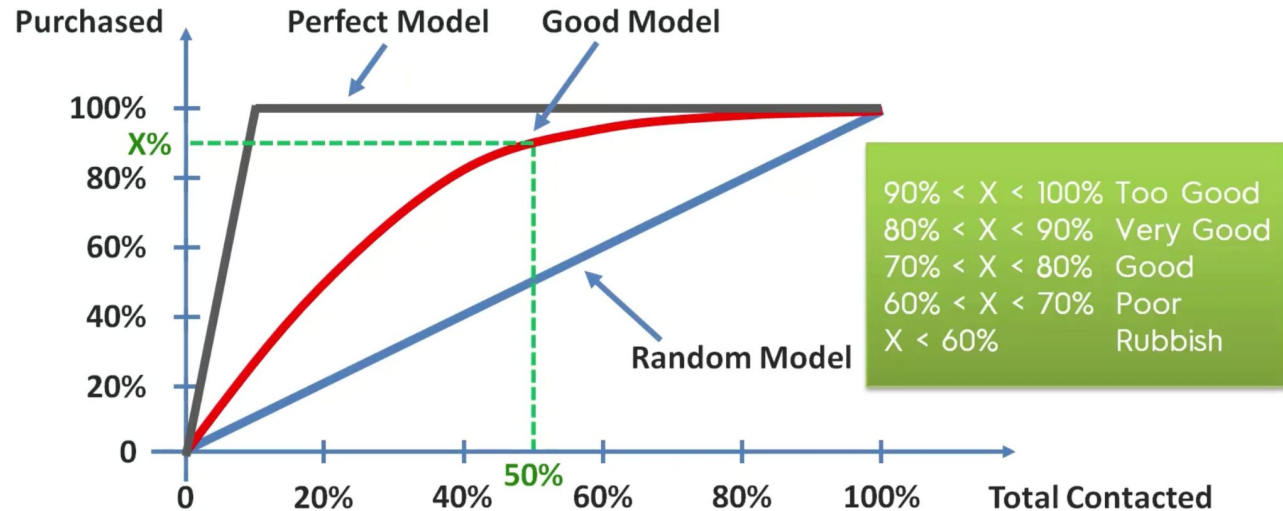


RESTORY...

# Mätvärden - CAP Analysis

Hur bra om vi bara använder oss av 50%?

Decision boundry



RESTORY...

# Mätvärden - Sammanfattning

- **Precision** (noggrannhet) -  $tp/(tp+fp)$
- **Recall** () -  $tp/(tp+fn)$
- **F-score** - Viktat medelvärde av precision och recall
- CAP - Cumulative Accuracy Profile
- ROC kurva - Receiver operating characteristic

CAP = Cumulative Accuracy Profile



ROC = Receiver Operating Characteristic

# Länkar

- [Ultimate guide to regression](#)
- [Classification algorithms in ML](#)
- [Multi class applications in real life](#)
- [Supervised learning](#)
- [Linjär regression - edureka](#)
- [Elements of ai - Linjär regression](#)
- [Measurements recall and precision](#)
- [Wikipedia: precision and recall](#)
- [Logistic regression - IRL](#)
- [ROC](#)
- [CAP](#)
- [CAP - sida 30](#)
- [Precision, Recall & F1 - vid](#)

Övning:

Logistisk regression på [Breast cancer data](#).