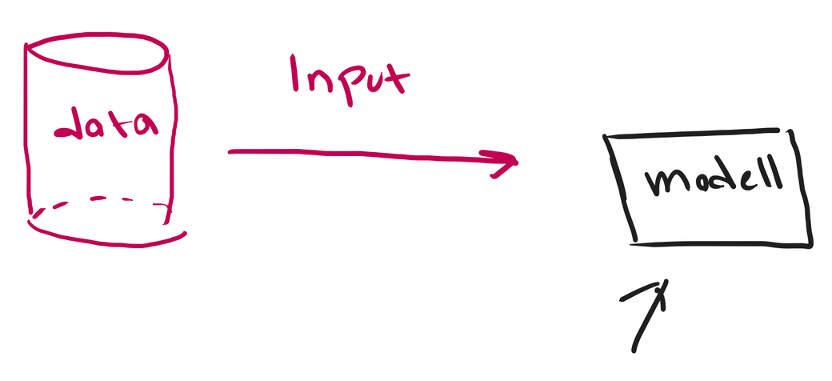
Importance of data quality



målet är att hitta optimala parametrar till vår modell

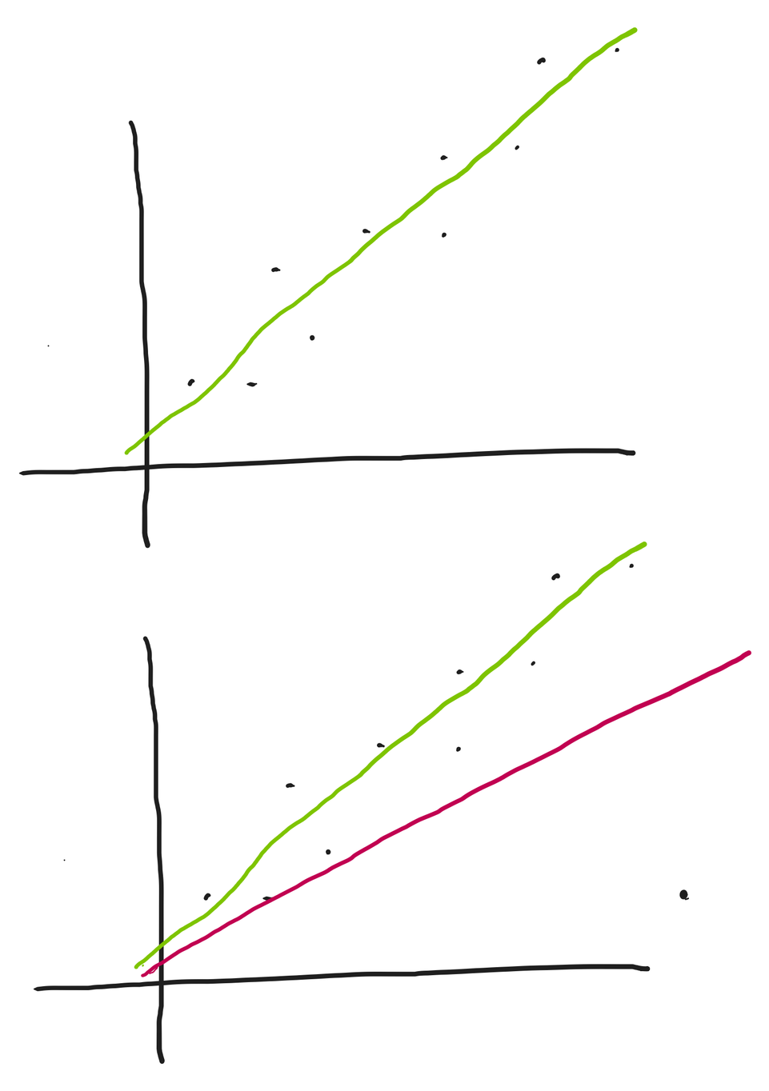
Vad händer om datan är "dålig" på något sätt?

Vad betyder "dålig" data?

- Många nulls
- felaktig data
- fel format
- Otydlig data
- Smutsig data
- Biased / fördomsfull data
- Orepresentativ data
- Irrelevant data

exempel — Hur ett enda pyttelitet felaktigt värde i träningsdatan kan försämra prestandan av den tränade modellen

Anta att vi har en feature X_1 och en target y



$$\frac{1}{n}\sum_i (y^i - \hat{y}^i)^2$$

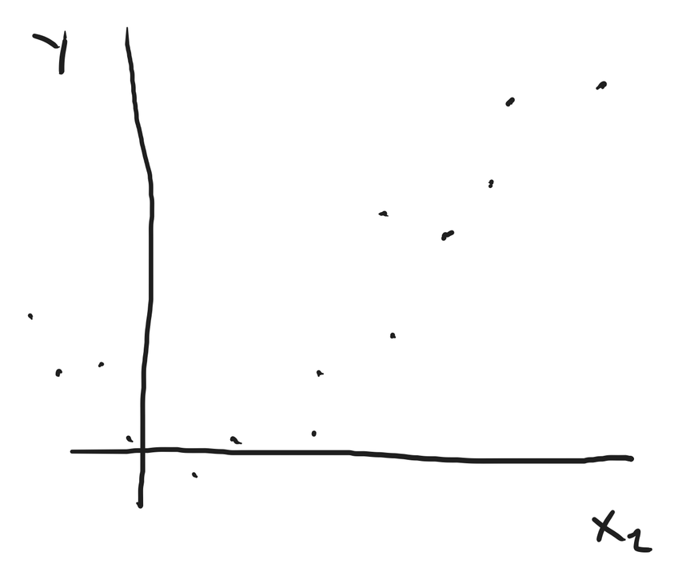
OK, men vilken data vill man samla och träna sin modell på?

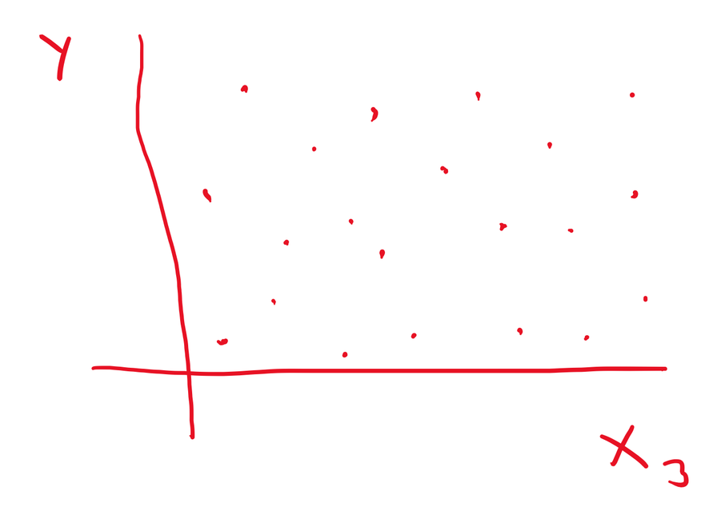
Svar: Relevant data!

Du vill ha features som har en direkt påverkan på din önskade target

Det är viktigt att ha features som är relevanta. Ingenting stoppar dig från att använda även irrelevanta features, men detta kommer sannolikt att ha en negativ påverkar på modellens prestanda eftersom att dessa features kommer att störa modellen.

Många gånger är det svårt att veta vad som är relevant data, och då behöver man samverka med domänexperter.





Features

$x_1, x_2, x_3, x_4, x_5, y$

Du väljer helt och hållet själv vilka features du vill använda för att träna en modell på.

Hur du väljer dina features ska inledningsvis guidas av din intuition för problemet (är features relevanta?) och domänexpertis. Du vill endast ha relevanta features för din träning!

Det är inget som heller stoppar dig från att skapa dina egna features!

$$x_6 = (x_2)^2$$
$$x_7 = x_1 \cdot x_2$$

Ok, anta nu att vi vill träna en linjär modell med x1, x2, och x7 som features. Hur ser det ut?

$$y = w_7 \cdot x_7 + w_2 \cdot x_2 + w_1 \cdot x_1 + w_0$$

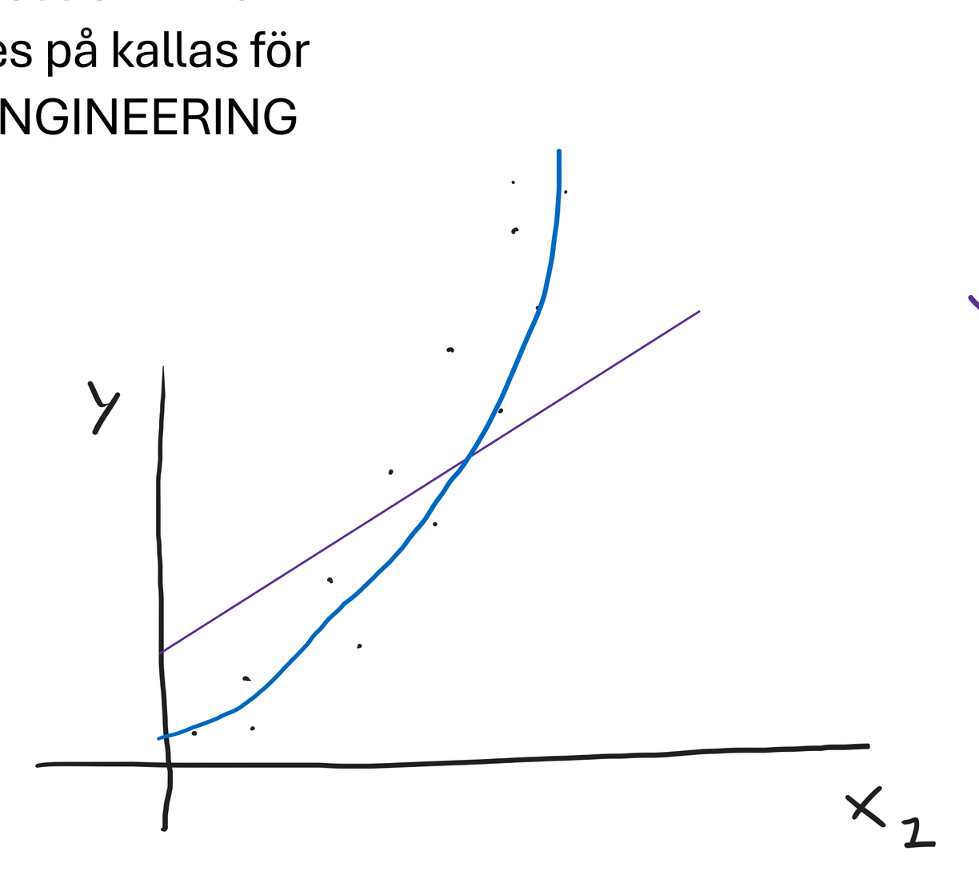
Du har all frihet att skapa dina features hur du vill. Observera dock att detta inte är någon lekstuga utan skapandet av dina features ska ske enligt någon form av logik/domänexpertis.

$$x_8 = x_3 x_4 \sqrt{x_5}$$

Denna process att hitta BRA features på kallas för FEATURE ENGINEERING

exempel

x_2, y



$$y = w_1 \cdot x_1 + w_0$$
$$x_2 = (x_2)^2$$
$$y_2 = w_2 \cdot x_2 + w_1 \cdot x_2 + w_0$$

Vi har nu bland annat sätt hur förödande dålig data är i ML.

För att träna bra modeller behöver vi således också bra data. I många fall behöver vi också mycket (kvantiet) data. Men, i en hel del fall klarar vi oss faktiskt undan med en mindre mängd (kvantitet) data. Så länge det är riktigt bra data!

Kvalitet > kvantitet