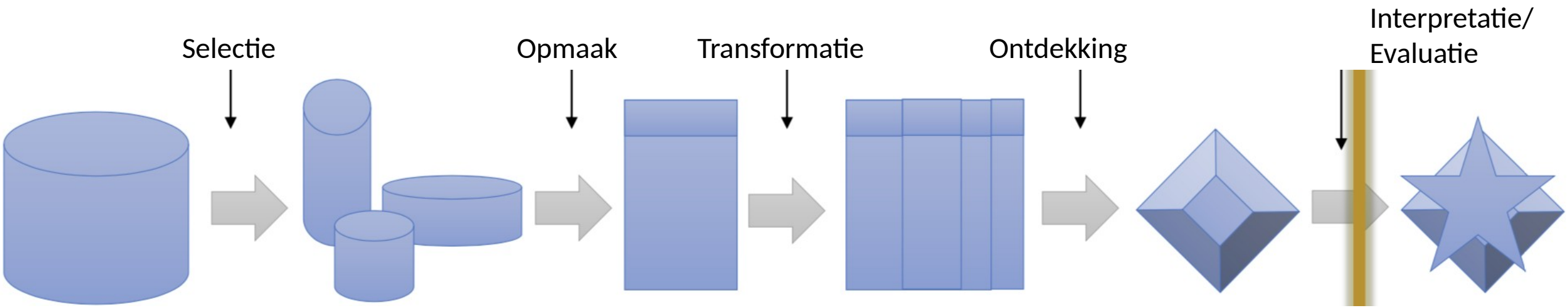
Kennis Management En Business Intelligentie

Model Evaluatie

Hoofdstuk 10
en 11

Deze cursus


Selectie
Opmaak
Transformatie
Ontdekking
Interpretatie/
Evaluatie
Data
Geselecteerde data
Opgemaakte/
verwerkte
data
Getransformeerde
data
Ontdekte
patronen
Kennis/
Inzicht
Gecontroleerd/
ongecontroleerd
Visualisatie/
validatie
Voor-processing: steekproef afname, normaliseren,
Beperkte selectie, ordelijk maken

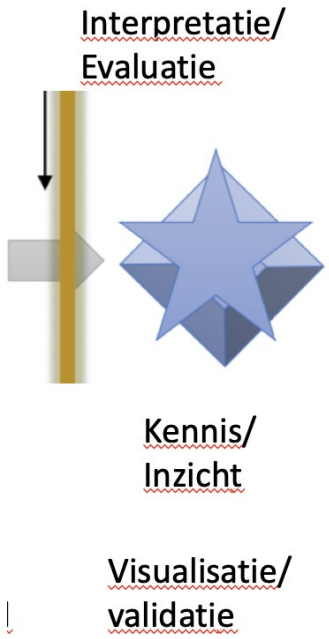
Model validatie en evaluatie

Wat te valideren?

- Model specificatie (bv. Selectie van variabelen, definiëring business vraag)
- Model quantificatie (bv. schatting van coëfficiënten, lay-out van keuze boom)
- Model performantie: de voorspelbaarheid van het model

Soorten validatie

- Apparentie (eigen staal)
- Intern (eigen populatie)
- Extern (eigen populatie)
- Model performantie op getrainde, test, validatie set



Model validatie

¬ De **business relevantie** van het analytisch model moet worden gegarandeerd

¬ De **Statistische performantie en validatie** moet gebalanceerd zijn tegen statistische performantie: verantwoordbaar en interpreteerbaar

¬ **Operationele efficiëntie en economische kosten** moeten in aanmerking komen

¬ **Gereguleerde observatie** wordt toenemend belangrijk

https://medium.com/dataminingapps-articles/critical-success- factors-for-analytical-models-be35e2cbdef2#.lpxh6v89u

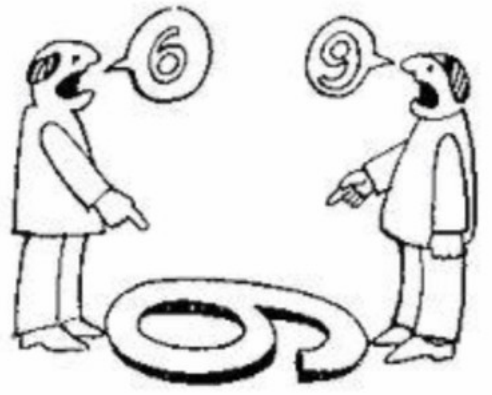
http://www.dataminingapps.com/dataminingapps-newsletter/

Business relevantie

- Het analytisch model zou het model moeten oplossen waarvoor het gemaakt is
- Hiervoor is grondige business kennis en verstand van het aan te nemen probleem nodig voor enige analyse kannen worden gestart
- Voorbeeld start vragen zijn: Wat is het, Hoe meet men het, hoe beheart men het, …

Vaak zijn er verschillende manieren om een probleem te definiëren:

- Bv. Voorspelling van welke klant volgend jaar ‘Gold status’ zal behalen
- Klassificatie van de huidige groeps tijd op tijdstip t
- Klassificatie op het verschil van t -> t+1
- Regressie op de hoeveelheid die de ‘Gold status’ bepaalt
- Tijdreeksen voorspellen op de bepaling van de ‘Gold status’ (hoelang duurt het)
- Hoeveel is genoeg? Hoeveel wil je erover weten?



Model performantie en juistheid

Juist label	Voorspelling
Nee	0.11
Nee	0.20
Ja	0.85
Ja	0.84
Ja	0.80
Nee	0.65
Ja	0.44
Nee	0.10
Ja	0.32
Ja	0.87
Ja	0.61
Ja	0.60
Ja	0.78
Nee	0.61

Voorspeld label	Correct?
nee	juist
nee	juist
ja	juist
ja	juist
ja	juist
ja	fout
nee	fout
nee	juist
nee	fout
ja	juist
ja	juist
ja	juist
ja	juist
	fout

Toename 0.50

¬ Verwarrings matrix

Referentie	Voorspelling: Nee	Voorspelling: Ja
Nee	3	2
Ja	2	7

Model performantie en juistheid

		Predicted condition		
	Total population	Predicted Condition positive	Predicted Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$
True condition	condition positive	**True positive**	**False Negative** (Type II error)	True positive rate (TPR), Sensitivity, Recall = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$
	condition negative	**False Positive** (Type I error)	**True negative**	False positive rate (FPR), Fall-out = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	True negative rate (TNR), Specificity (SPC) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$
	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Test outcome positive}}$	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Test outcome negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$
		False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Test outcome positive}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Test outcome negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$	

Model performantie en juistheid

Juist label	Voorspelling
Nee	0.11
Nee	0.20
Ja	0.85
Ja	0.84
Ja	0.80
Nee	0.65
Ja	0.44
Nee	0.10
Ja	0.32
Ja	0.87
Ja	0.61
Ja	0.60
Ja	0.78
Nee	0.61

Voorspeld label	Correct?
nee	juist
nee	juist
ja	juist
ja	juist
ja	juist
ja	fout
nee	fout
nee	juist
nee	fout
ja	juist
ja	juist
ja	juist
ja	juist
ja	fout

Toename 0.50

¬ Verwarrings matrix

Referentie	Voorspelling: Nee	Voorspelling: Ja
Nee	3	2
Ja	2	7

¬ **Accuracy** = (tp + tn) / total = (3 + 7) / 14 = 0.71
¬ **Balanced accuracy** = (recall + specificity) / 2 = (0.5 * tp) / (tp + fn) + (0.5 * tn) / (tn + fp) = 0.5 * 0.78 + 0.5 * 0.60 = 0.69
¬ **Recall (sensitivity)** = tp / (tp + fn) = 7 / 9 = 0.78 “Hoeveel juistheden hebben we gevonden?”
¬ **Precision** = tp / (tp + fp) = 7 / 9 = 0.78 “Hoeveel hebben we er fout?”

Model performantie en juistheid

Juist label	Voorspelling
Nee	0.11
Nee	0.20
Ja	0.85
Ja	0.84
Ja	0.80
Nee	0.65
Ja	0.44
Nee	0.10
Ja	0.32
Ja	0.87
Ja	0.61
Ja	0.60
Ja	0.78
Nee	0.61

Voorspeld label	Correct?
nee	juist
nee	juist
ja	juist
ja	juist
ja	juist
ja	fout
nee	fout
nee	juist
nee	fout
ja	juist
ja	juist
ja	juist
ja	juist
	fout

Toename 0.50

¬ Verwarrings matrix

Referentie	Voorspelling: Nee	Voorspelling: Ja
Nee	3	2
Ja	2	7

¬ **Recall en Precision zijn meestal aanvullend:** om meer goede resultaten uit te komen moet je bereid zijn om meer fouten te maken

Model performantie en juistheid

Juist label	Voorspelling
Nee	0.11
Nee	0.20
Ja	0.85
Ja	0.84
Ja	0.80
Nee	0.65
Ja	0.44
Nee	0.10
Ja	0.32
Ja	0.87
Ja	0.61
Ja	0.60
Ja	0.78
Nee	0.61

Toename 0.50

- Voor elke mogelijke threshold $t \in T$ met T de set van alle voorspelde mogelijkheden, kunnen we een verwarrings matrix vormen

- Dus welke threshold is de beste?

Threshold	Tp	Fp	tr	fn
0.10	9	5	0	0
0.11	9	4	1	0
0.20	9	3	2	0
0.32	9	2	3	0
0.44	8	2	3	1
0.60	7	2	3	2
0.61	6	2	3	3
0.65	5	1	4	4
0.78	5	0	5	4
0.80	4	0	5	5
0.84	3	0	5	6
0.85	2	0	5	7
0.87	1	0	5	8

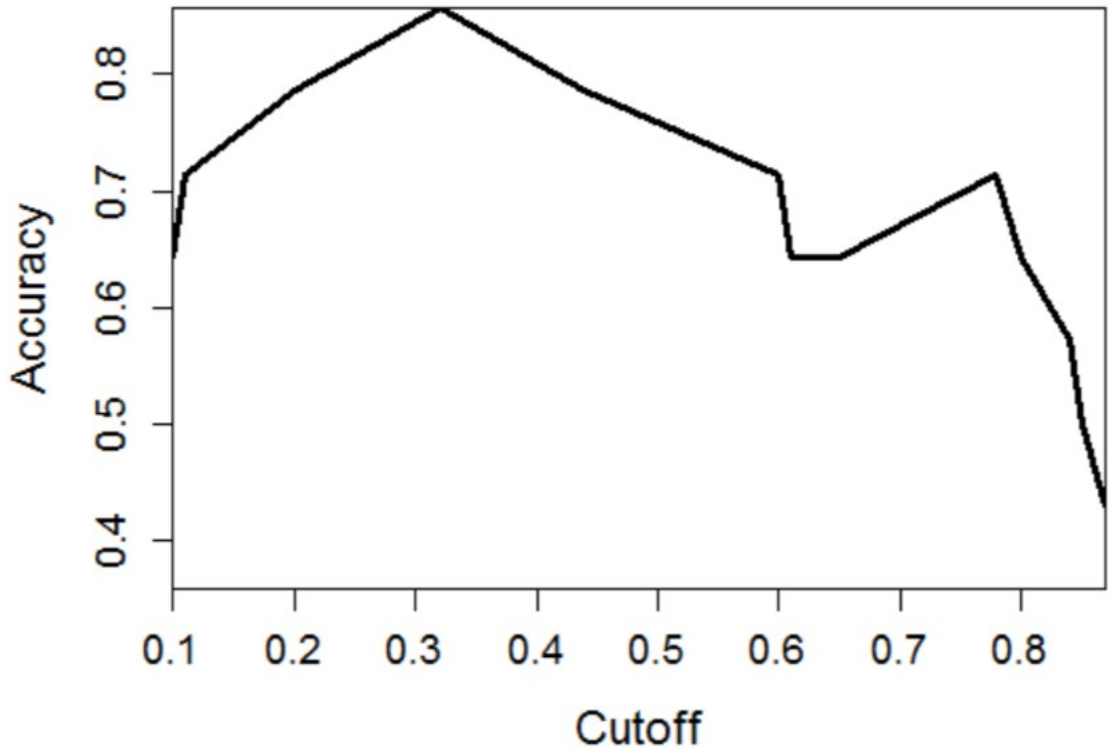
Model performantie en juistheid

Juist label	Voorspelling
Nee	0.11
Nee	0.20
Ja	0.85
Ja	0.84
Ja	0.80
Nee	0.65
Ja	0.44
Nee	0.10
Ja	0.32
Ja	0.87
Ja	0.61
Ja	0.60
Ja	0.78
Nee	0.61

Toename 0.50

¬ Voor elke mogelijke threshold $t \in T$ met T de set van alle voorspelde mogelijkheden, kunnen we een verwarrings matrix vormen

¬ Dus welke threshold is de beste? **ACC-threshold curve**



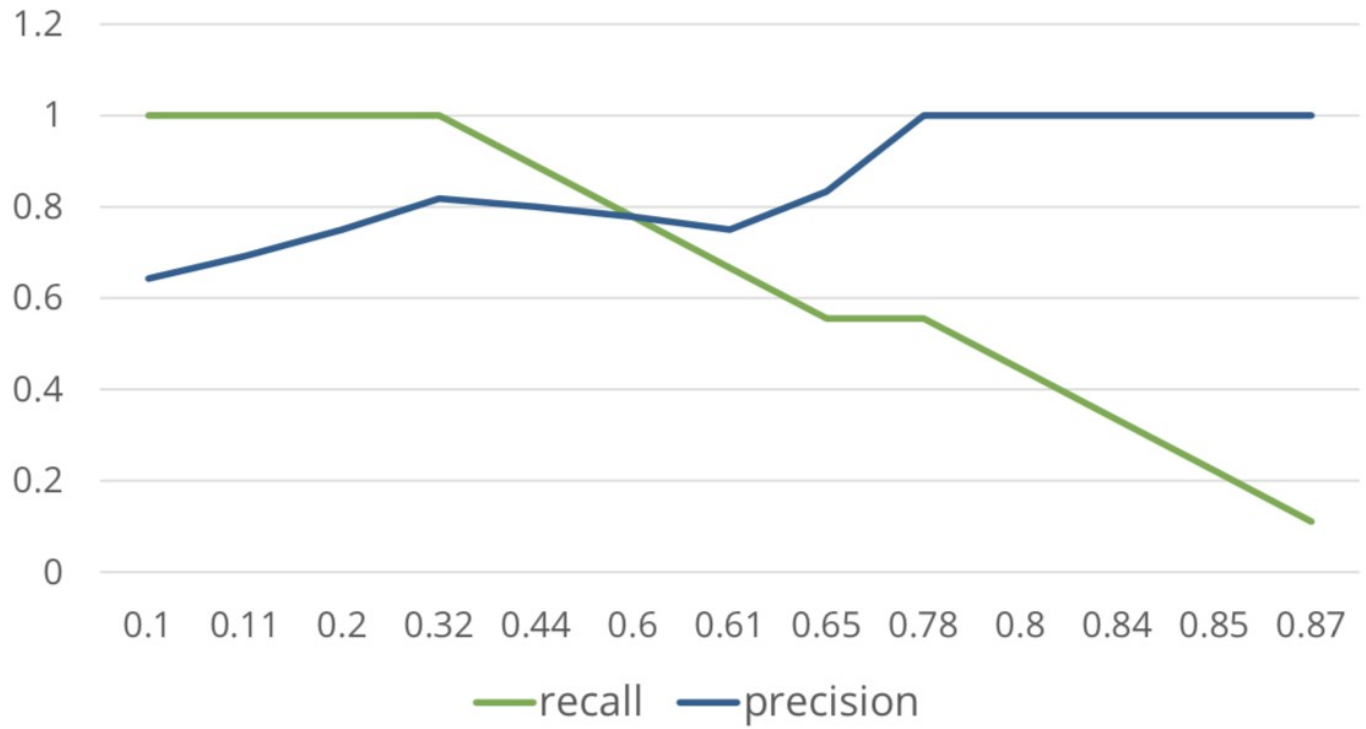
Model performantie en juistheid

Juist label	Voorspelling
Nee	0.11
Nee	0.20
Ja	0.85
Ja	0.84
Ja	0.80
Nee	0.65
Ja	0.44
Nee	0.10
Ja	0.32
Ja	0.87
Ja	0.61
Ja	0.60
Ja	0.78
Nee	0.61

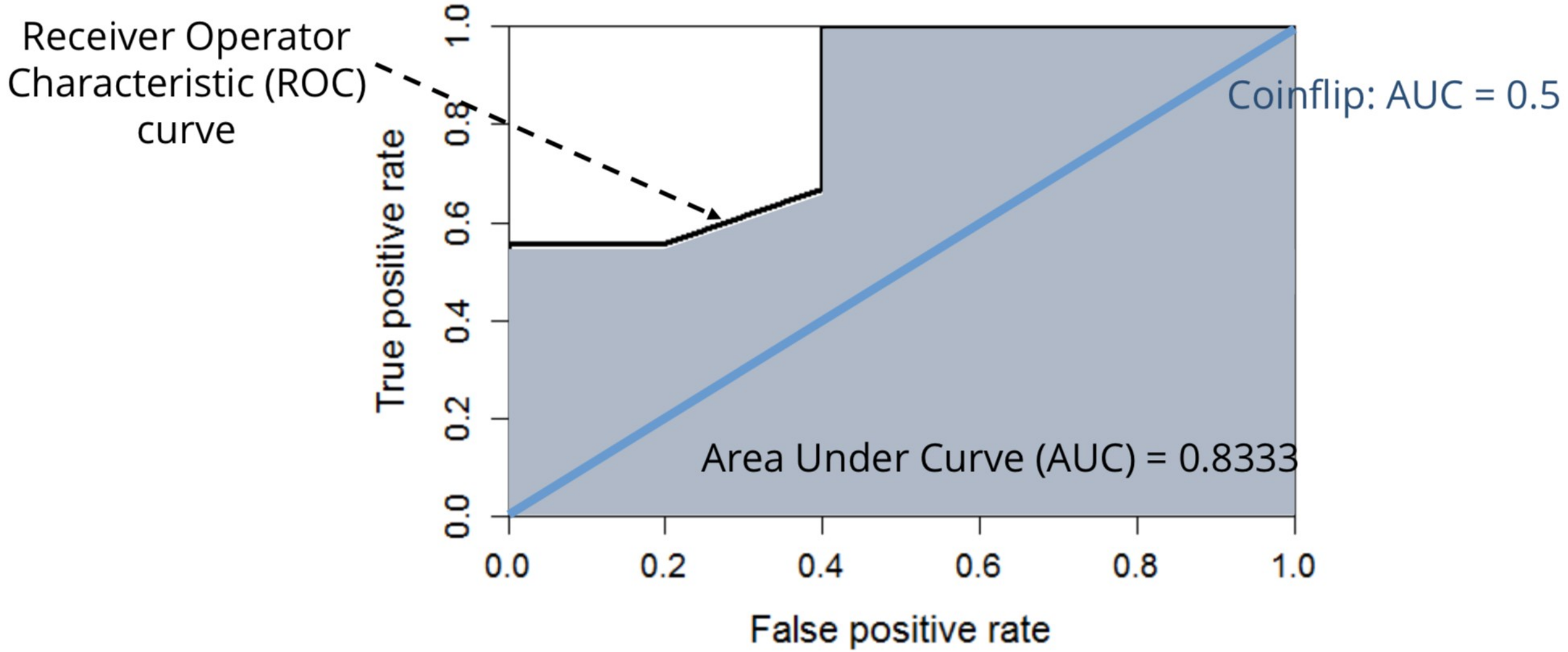
Toename 0.50

¬ Voor elke mogelijke threshold $t \in T$ met T de set van alle voorspelde mogelijkheden, kunnen we een verwarrings matrix vormen

¬ Dus welke threshold is de beste? **Recall-precision curve**



Model performantie en juistheid



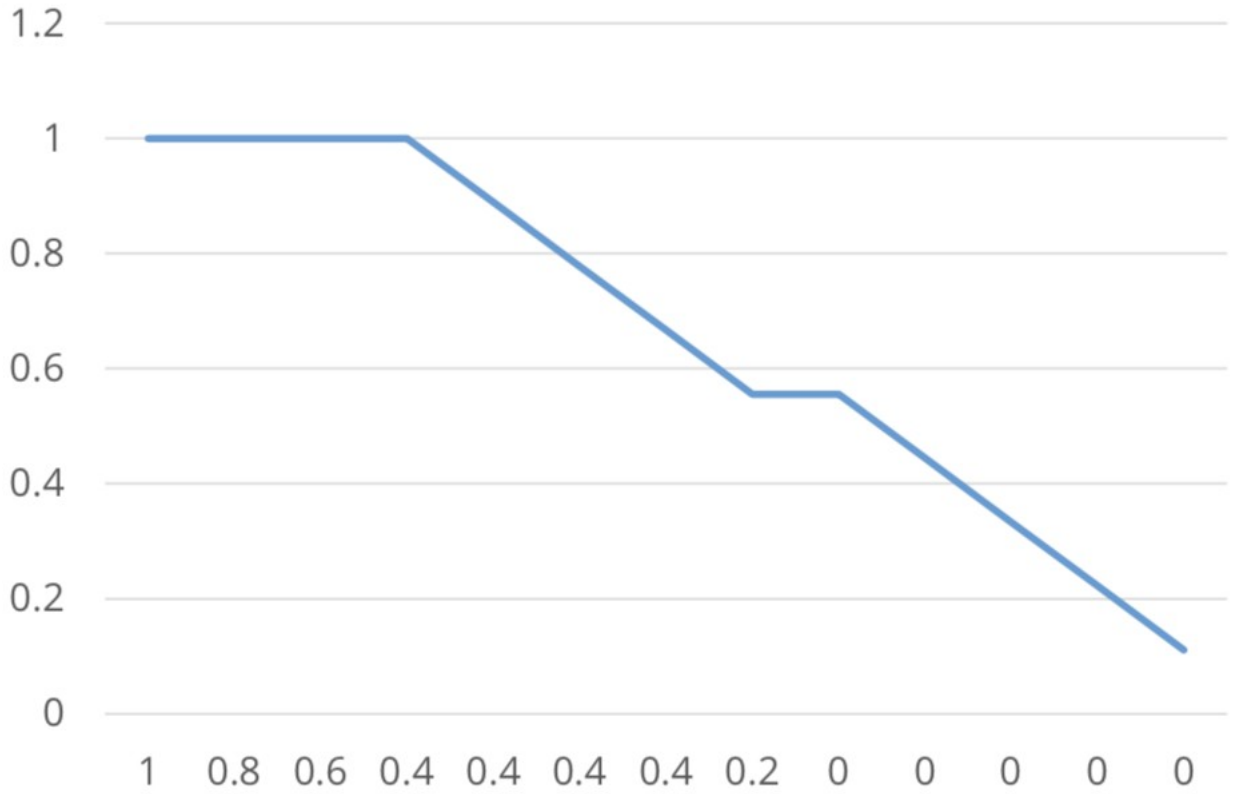
Model performantie en juistheid

Juist label	Voorspelling
Nee	0.11
Nee	0.20
Ja	0.85
Ja	0.84
Ja	0.80
Nee	0.65
Ja	0.44
Nee	0.10
Ja	0.32
Ja	0.87
Ja	0.61
Ja	0.60
Ja	0.78
Nee	0.61

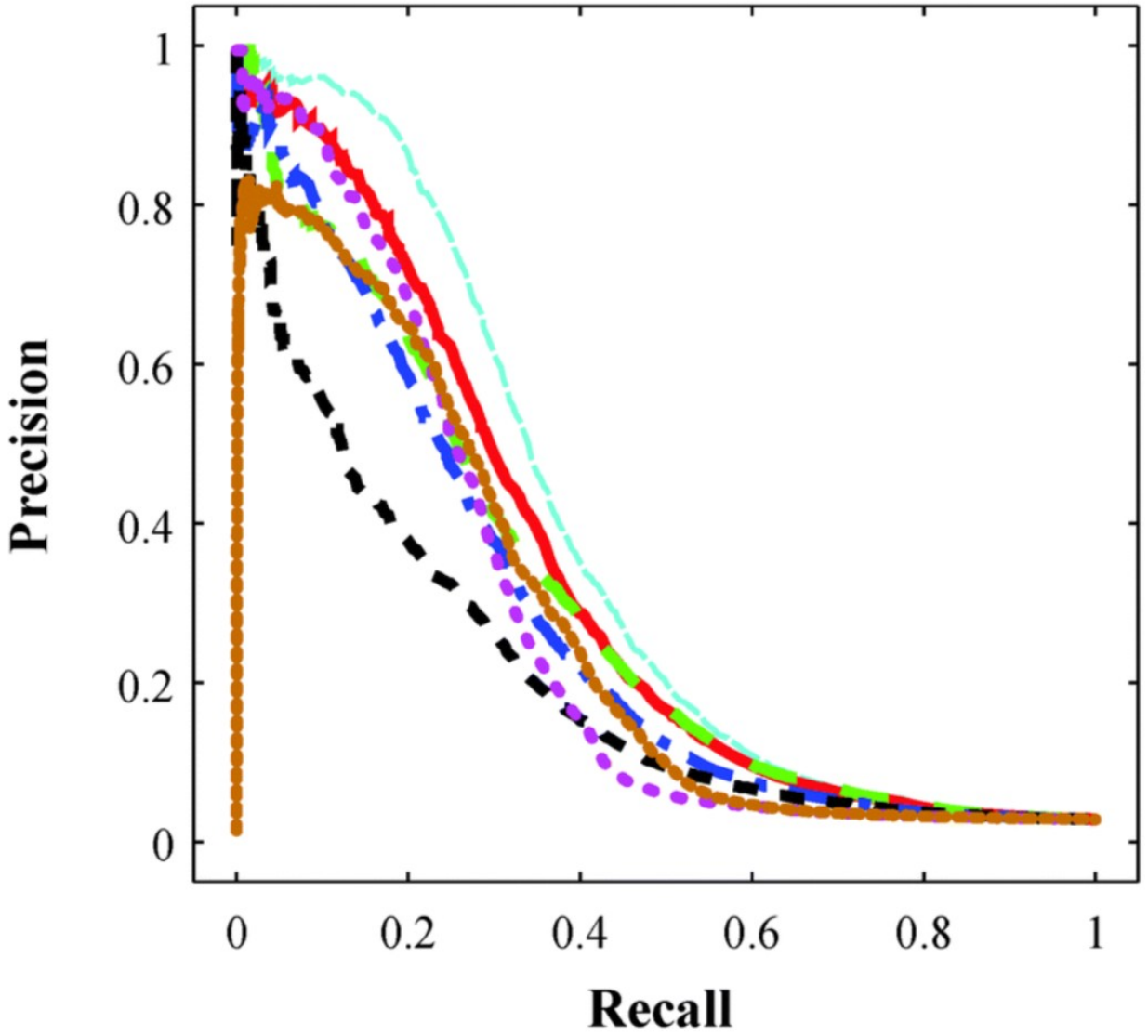
Toename 0.50

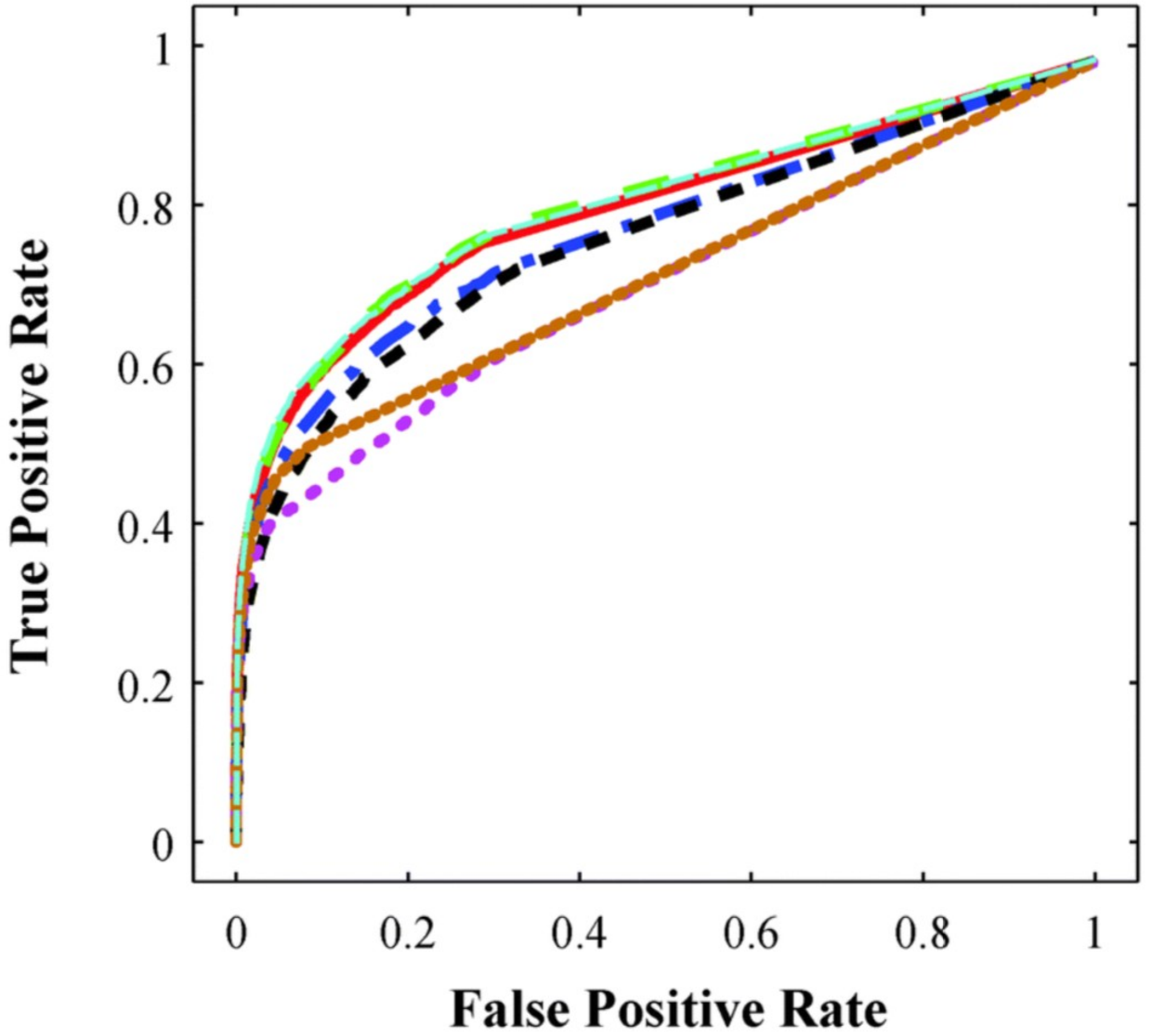
¬ Voor elke mogelijke threshold $t \in T$ met T de set van alle voorspelde mogelijkheden, kunnen we een verwarrings matrix vormen

¬ Dus welke threshold is de beste? **Roc curve en AUC**



Model performantie en juistheid





Model performantie en juistheid

Juist label	Voorspelling
Nee	0.11
Nee	0.20
Ja	0.85
Ja	0.84
Ja	0.80
Nee	0.65
Ja	0.44
Nee	0.10
Ja	0.32
Ja	0.87
Ja	0.61
Ja	0.60
Ja	0.78
Nee	0.61

Toename 0.50

- Voor elke mogelijke threshold $t \in T$ met T de set van alle voorspelde mogelijkheden, kunnen we een verwarrings matrix vormen

- Dus welke threshold is de beste? **Hangt af van het doel**

Threshold	Tp	Fp	tn	fn
0.10	9	5	0	0
0.11	9	4	1	0
0.20	9	3	2	0
0.32	9	2	3	0
0.44	8	2	3	1
0.60	7	2	3	2
0.61	6	2	3	3
0.65	5	1	4	4
0.78	5	0	5	4
0.80	4	0	5	5
0.84	3	0	5	6
0.85	2	0	5	7
0.87	1	0	5	8

Recall	precision	Tp-rate	Fp-rate
1	0.624857	1	1
1	0.692308	1	0.8
1	0.75	1	0.6
1	0.818182	1	0.4
0.888889	0.8	0.888889	0.4
0.777778	0.777778	0.777778	0.4
0.666667	0.75	0.666667	0.4
0.555556	0.833333	0.555556	0.2
0.555556	1	0.555556	0
0.444	1	0.44444	0

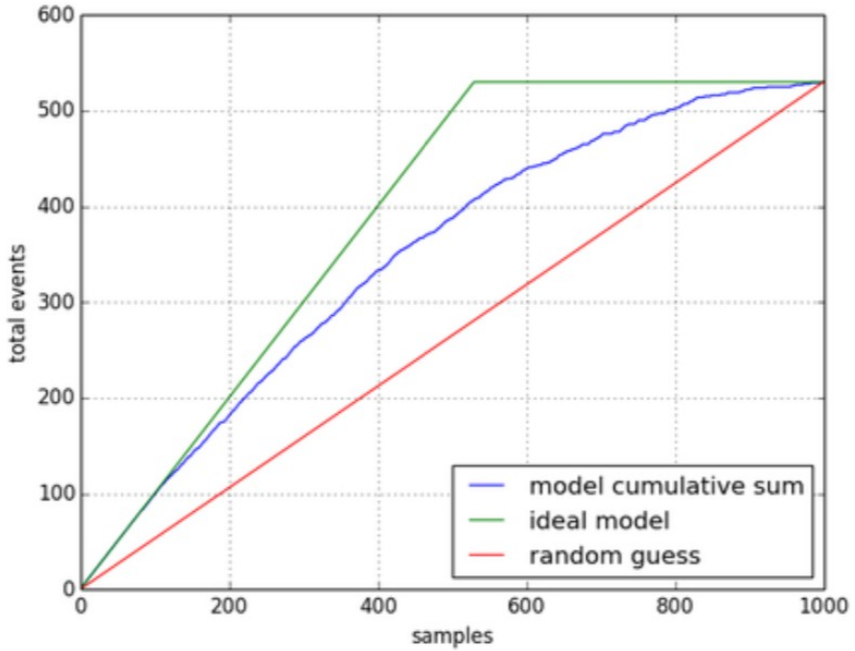
Model performantie en juistheid

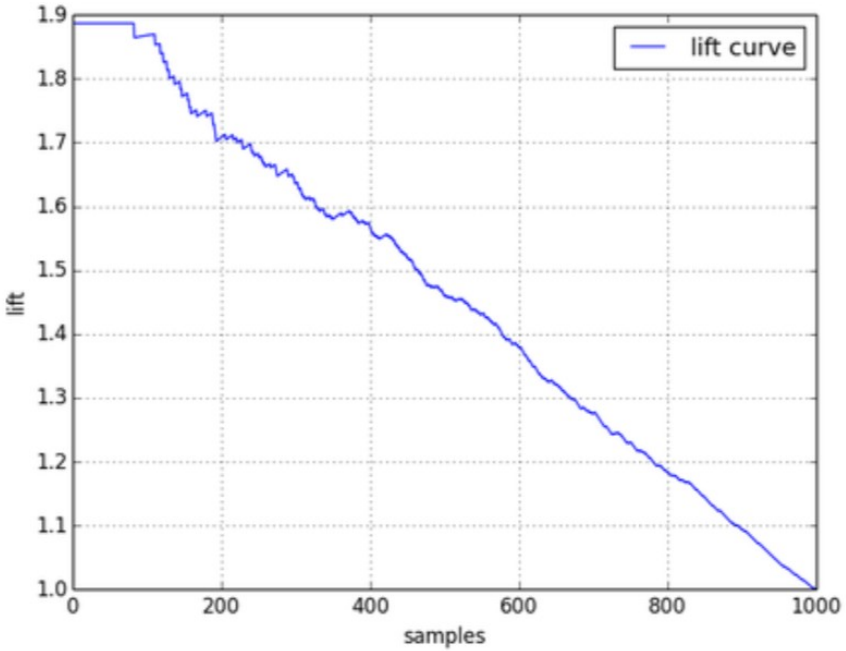
Juist label	Voorspelling
Nee	0.11
Nee	0.20
Ja	0.85
Ja	0.84
Ja	0.80
Nee	0.65
Ja	0.44
Nee	0.10
Ja	0.32
Ja	0.87
Ja	0.61
Ja	0.60
Ja	0.78
Nee	0.61

Toename 0.50

¬ Voor elke mogelijke threshold $t \in T$ met T de set van alle voorspelde mogelijkheden, kunnen we een verwarrings matrix vormen

¬ Dus welke threshold is de beste? **Lift (ratio van model tot een willekeurige gok)**





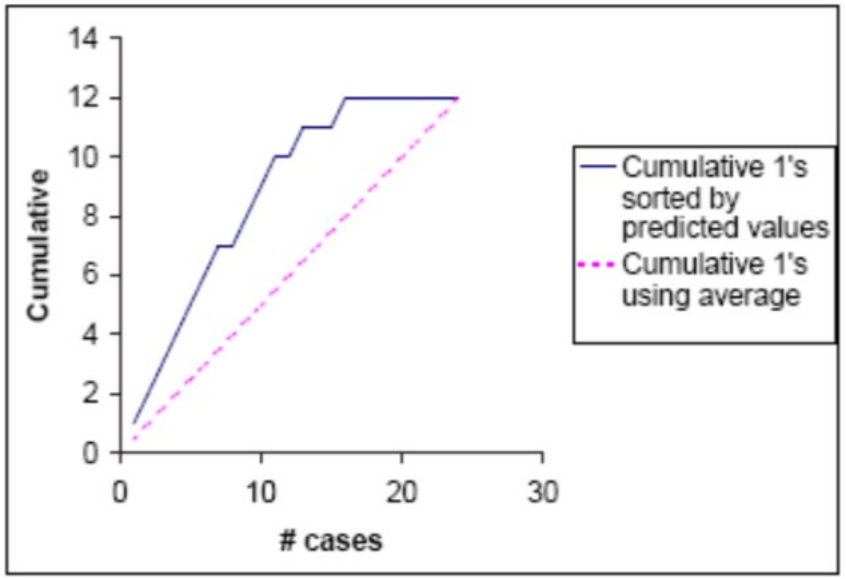
Model performantie en juistheid

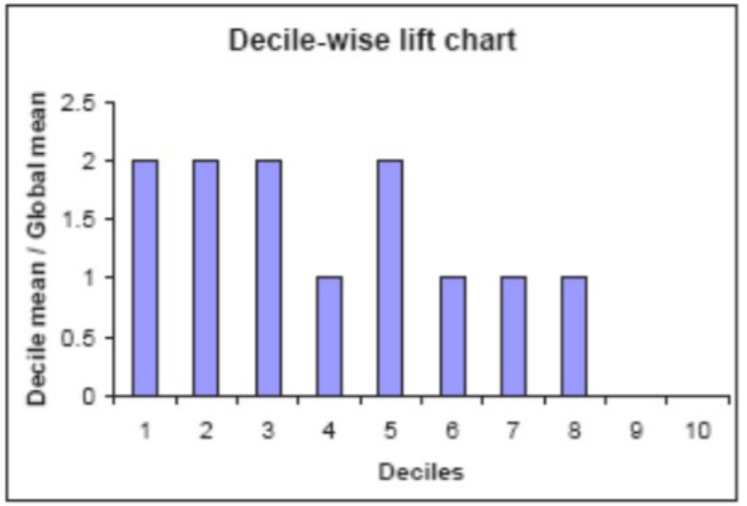
Juist label	Voorspelling
Nee	0.11
Nee	0.20
Ja	0.85
Ja	0.84
Ja	0.80
Nee	0.65
Ja	0.44
Nee	0.10
Ja	0.32
Ja	0.87
Ja	0.61
Ja	0.60
Ja	0.78
Nee	0.61

Toename 0.50

¬ Voor elke mogelijke threshold $t \in T$ met T de set van alle voorspelde mogelijkheden, kunnen we een verwarrings matrix vormen

¬ Dus welke threshold is de beste? **Decile lift (ratio van een model tot een willekeurige gok)**

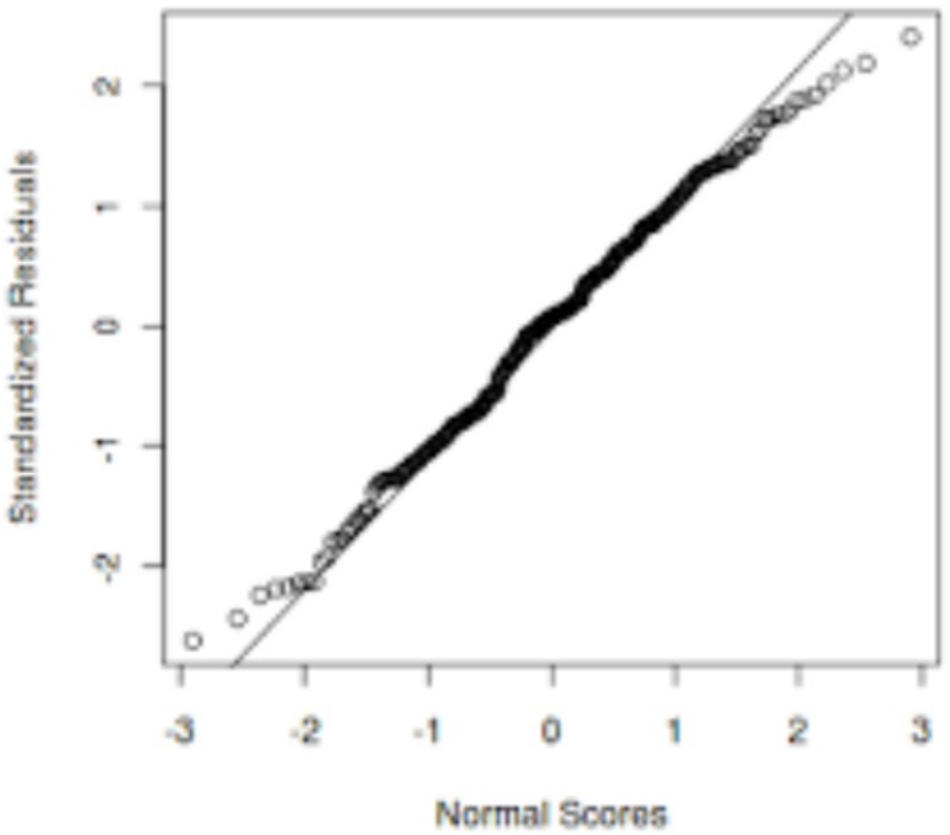




Model performantie en juistheid

En meer

- Residuals plot
- Correlatie tussen voorspelling en juist label
- Etc.

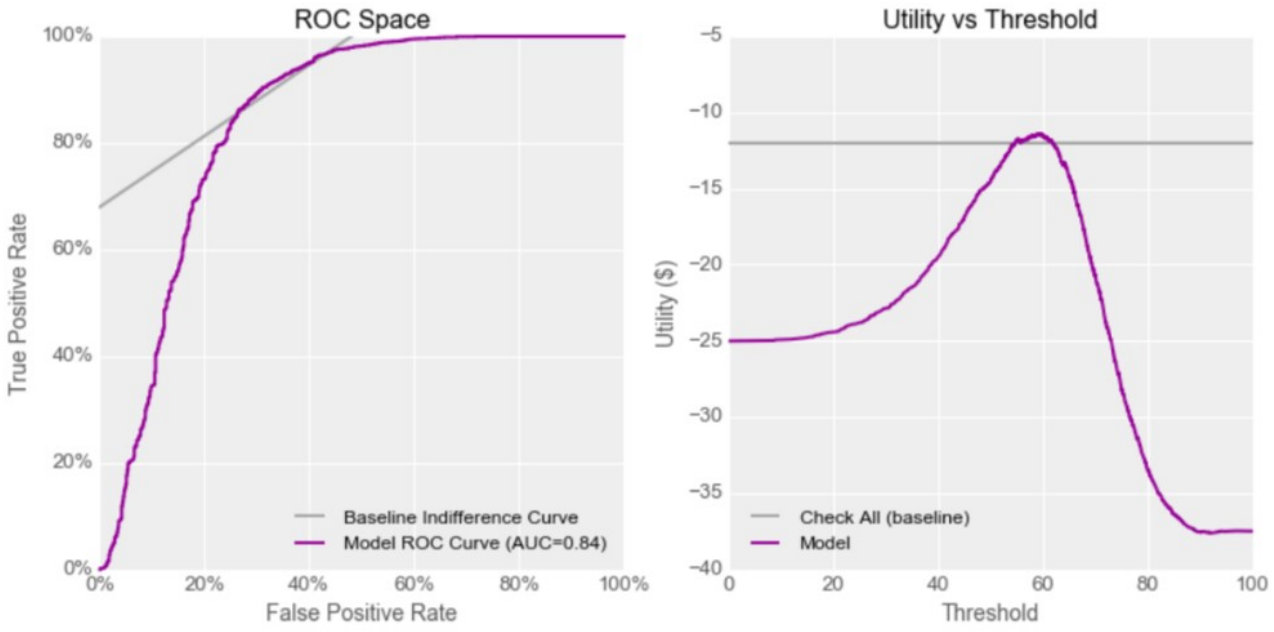


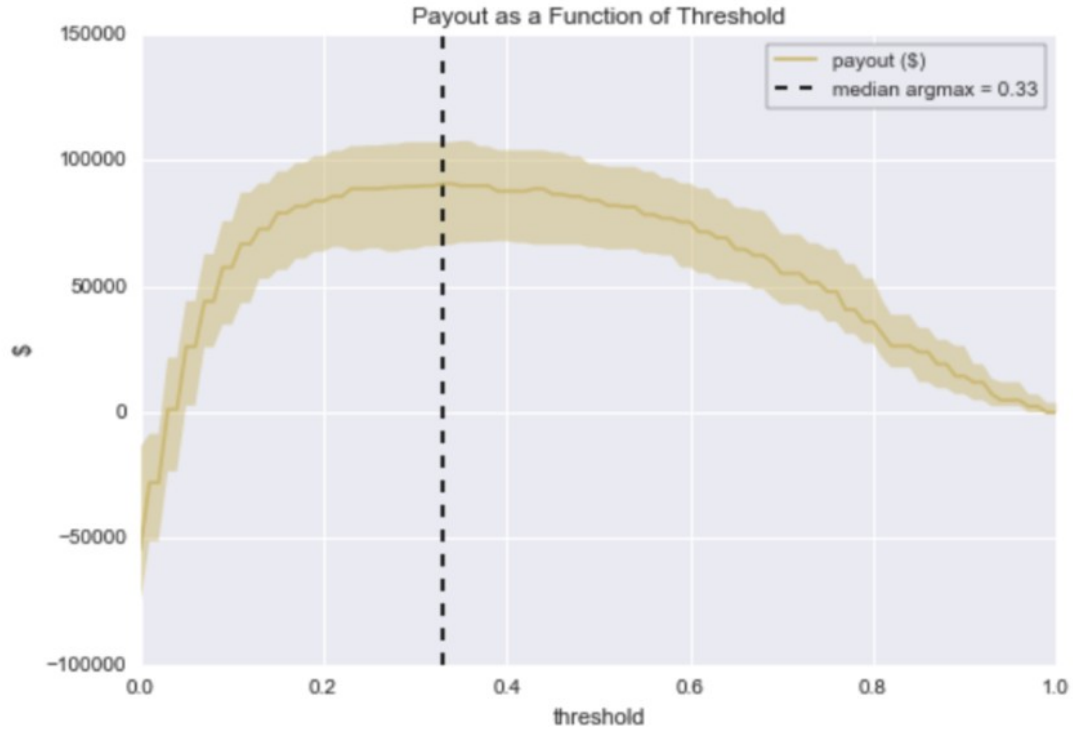
Model performantie en juistheid

En meer

- Zetten van misclassificatie kosten
- Berekening gebruikskost / voordeel

	Voorspelling		
Referentie		Nee	Ja
	Nee	3	2
	Ja	2	7

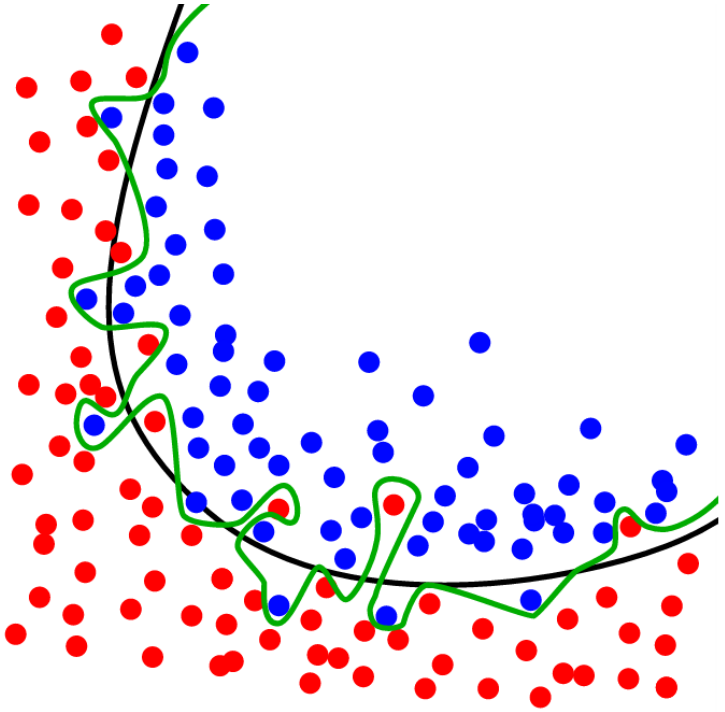




Model performantie en juistheid

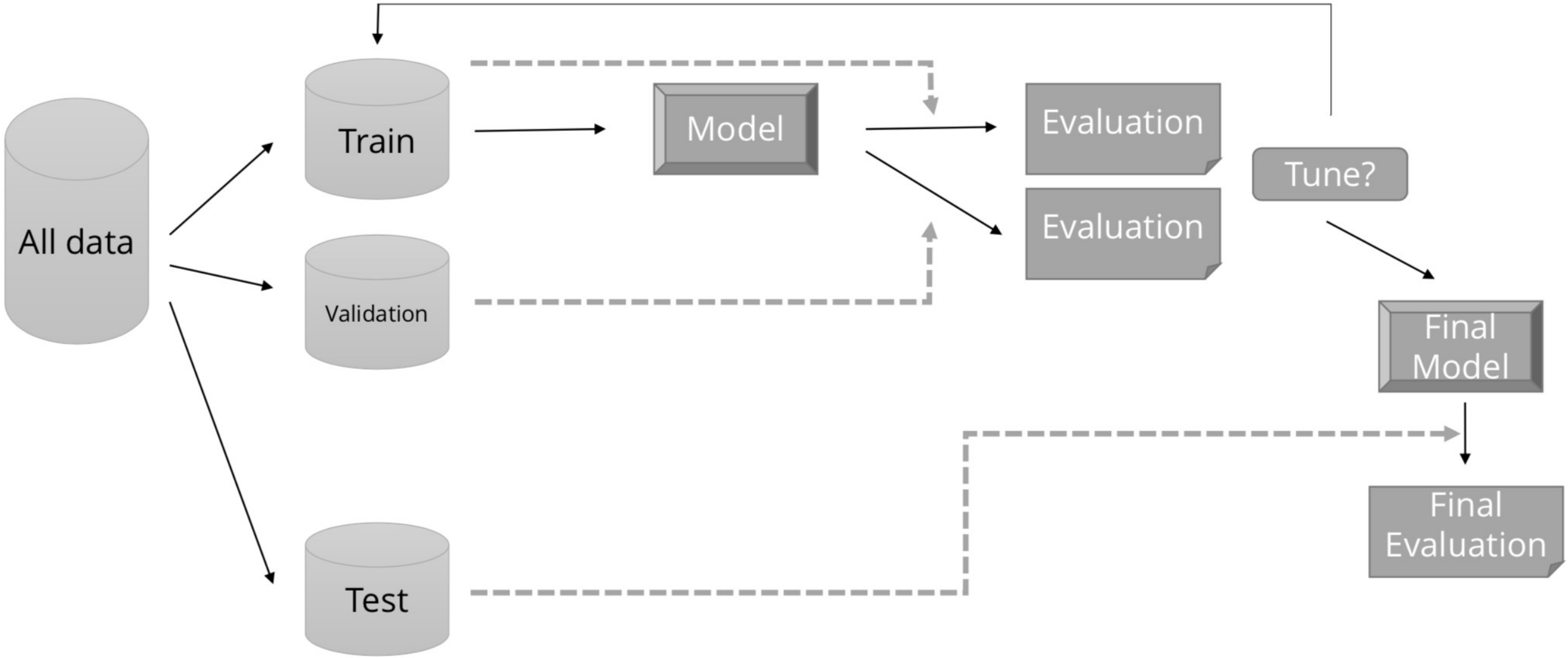
Preventie van te gepast te zijn, mogelijkheid tot generalisatie

- Gedurende het optimaliseren van het model
- Training, validatie en test set



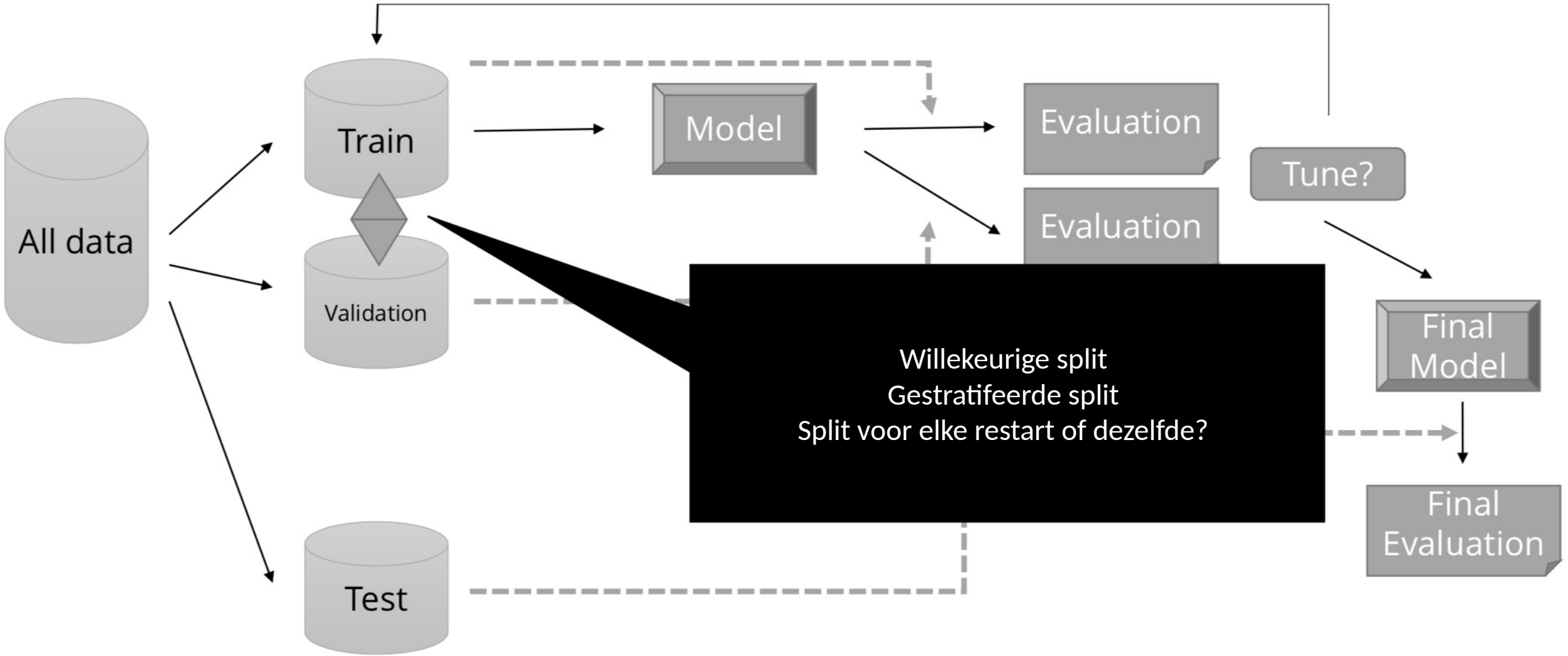
Model performantie en juistheid

Training, validatie en test set



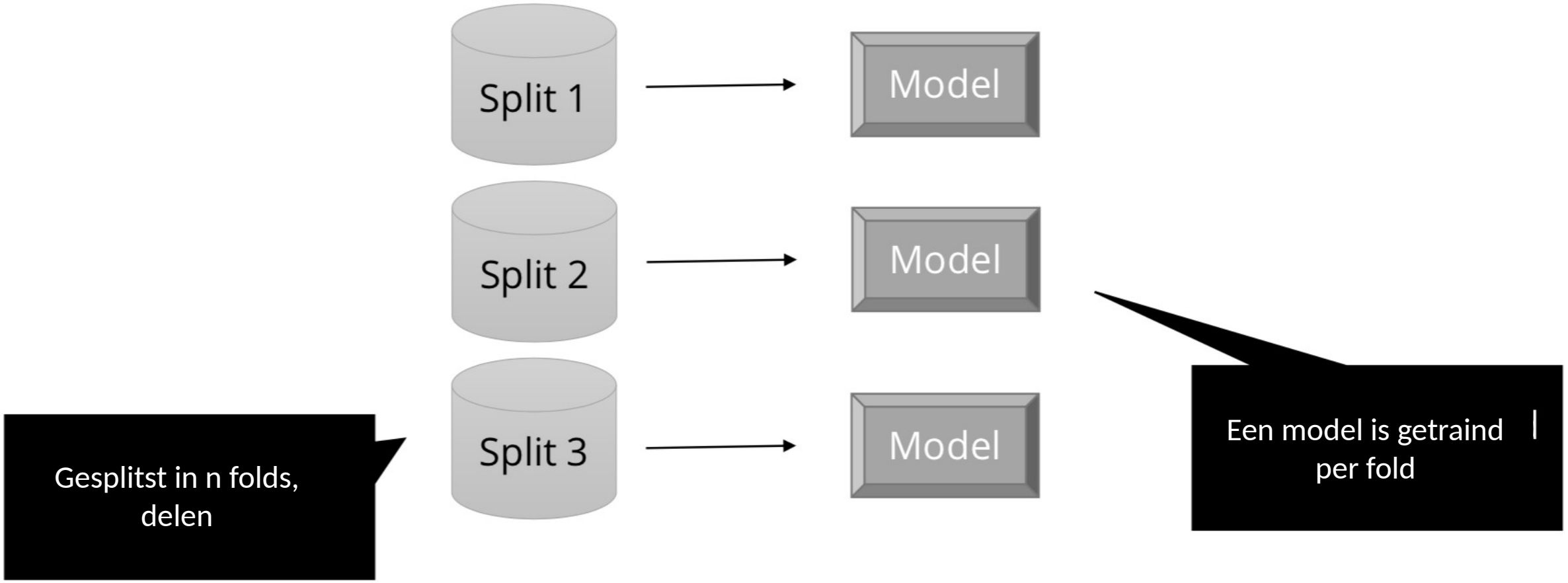
Model performantie en juistheid

Training, validatie en test set



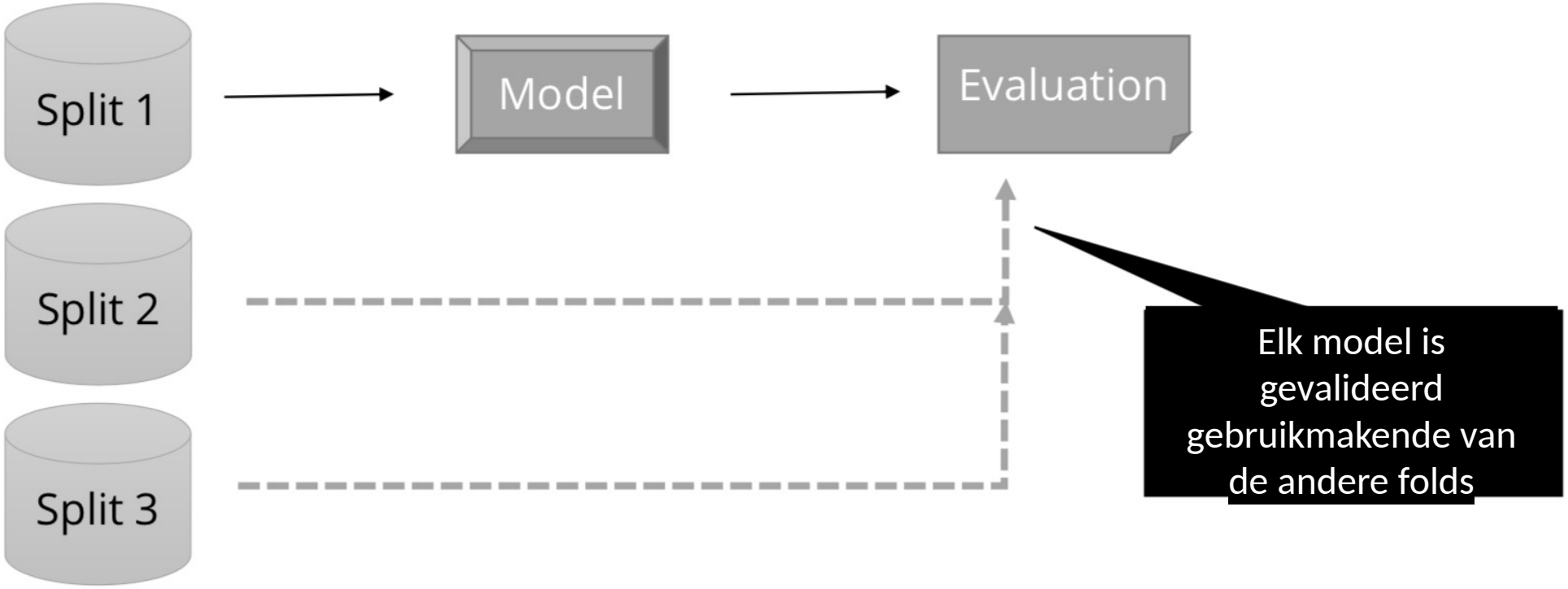
Model performantie en juistheid

Kruis-validatie



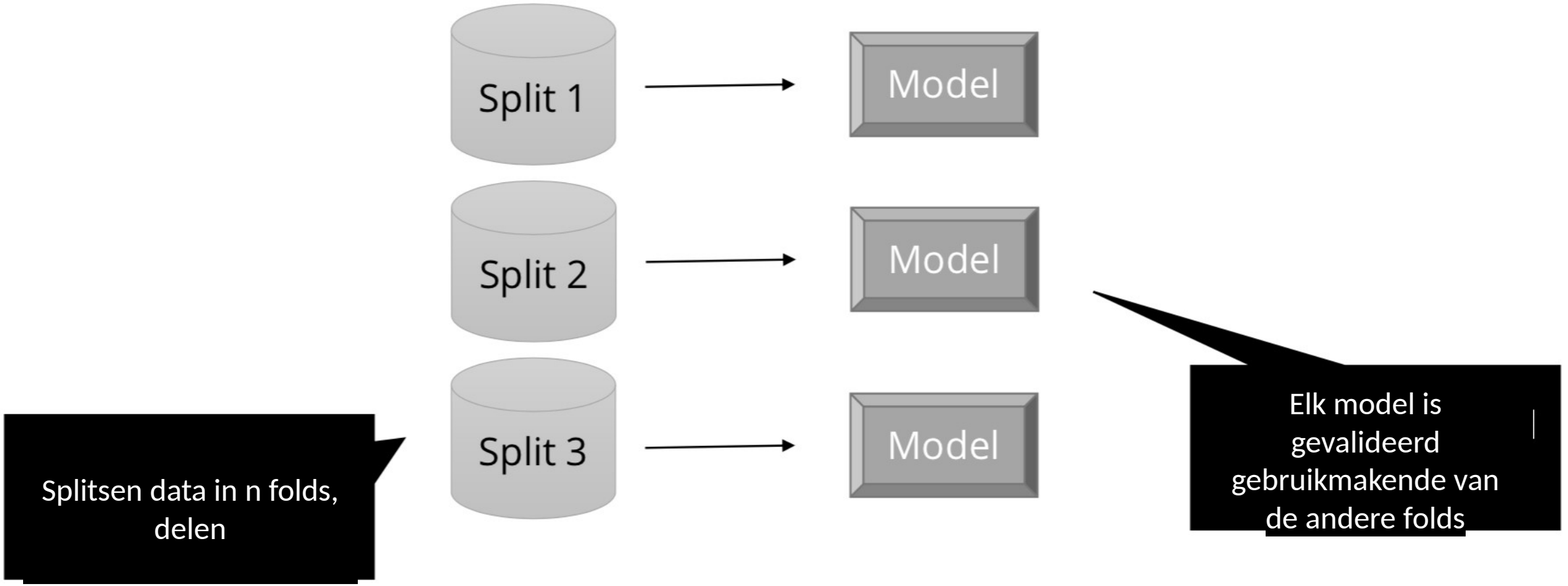
Model performantie en juistheid

Kruis-validatie



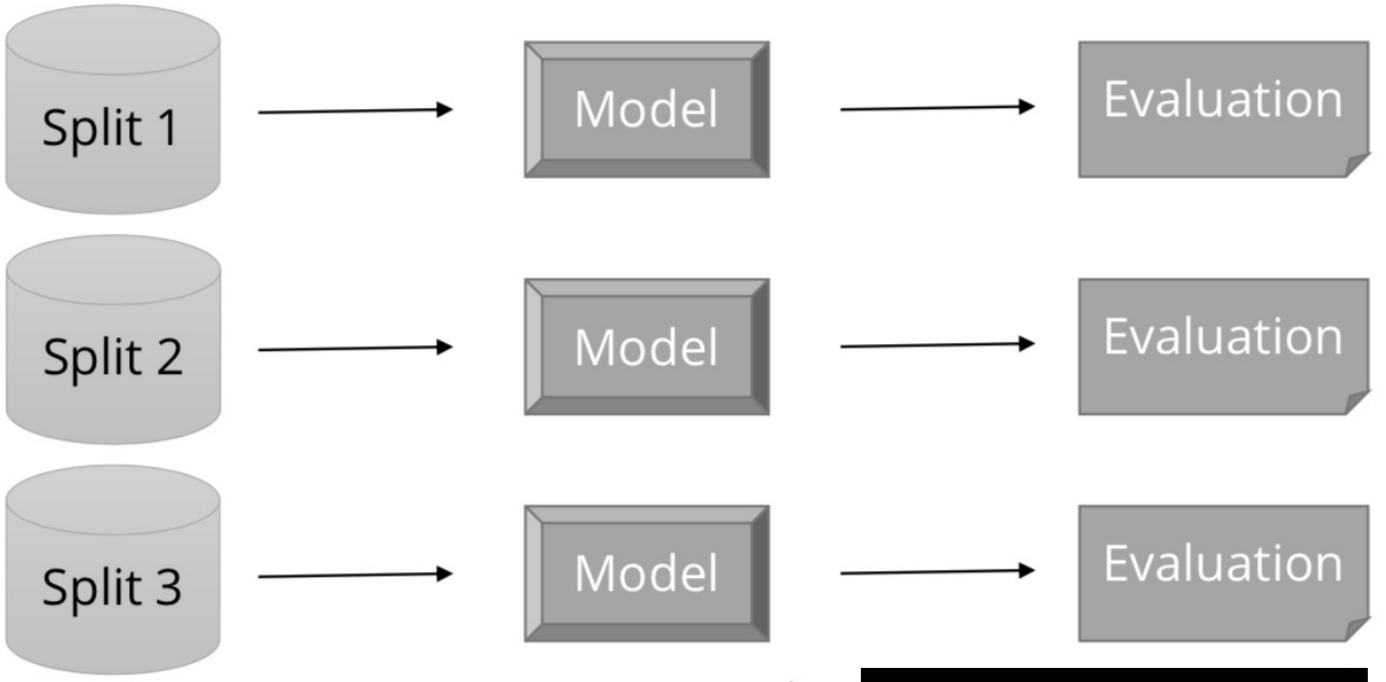
Model performantie en juistheid

Kruis-validatie



Model performantie en juistheid

Kruis-validatie



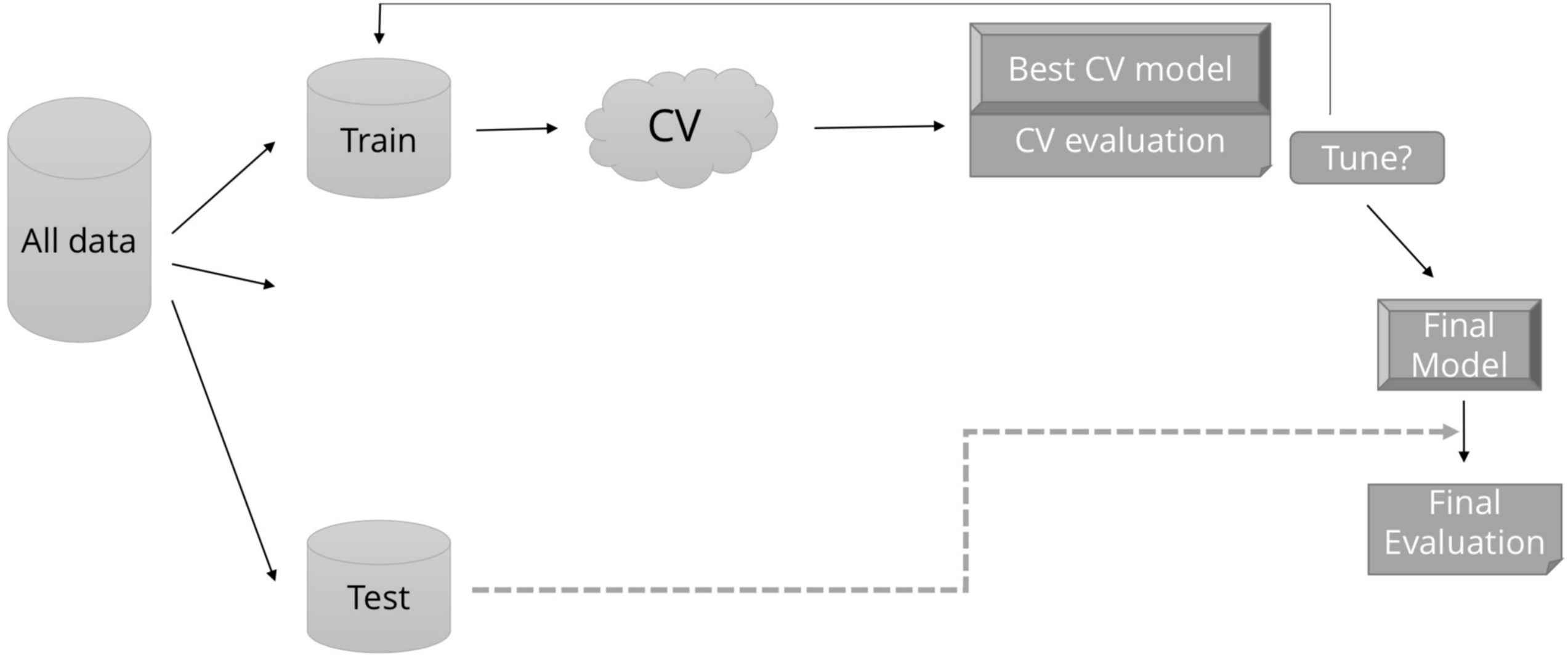
Splitsen data in n folds, delen

Beste model wordt bijgehouden

Finale evaluatie is verdeeld over alle deel modellen

Model performantie en juistheid

Training, validatie en test set



Model performantie en juistheid

Kruis-validatie

¬ Twee-voud, holdout: dis is de normale train/validatie deling

¬ K-voud

¬ Achterhouden-van-1, Jackknifing: verwijder telkens één instantie en train model op de overblijvende instanties

Model performantie en juistheid

Finale bedenkingen:

¬ Wat als de finale test evaluatie een slecht resultaat geeft? (*Weggooien heel het project?*)

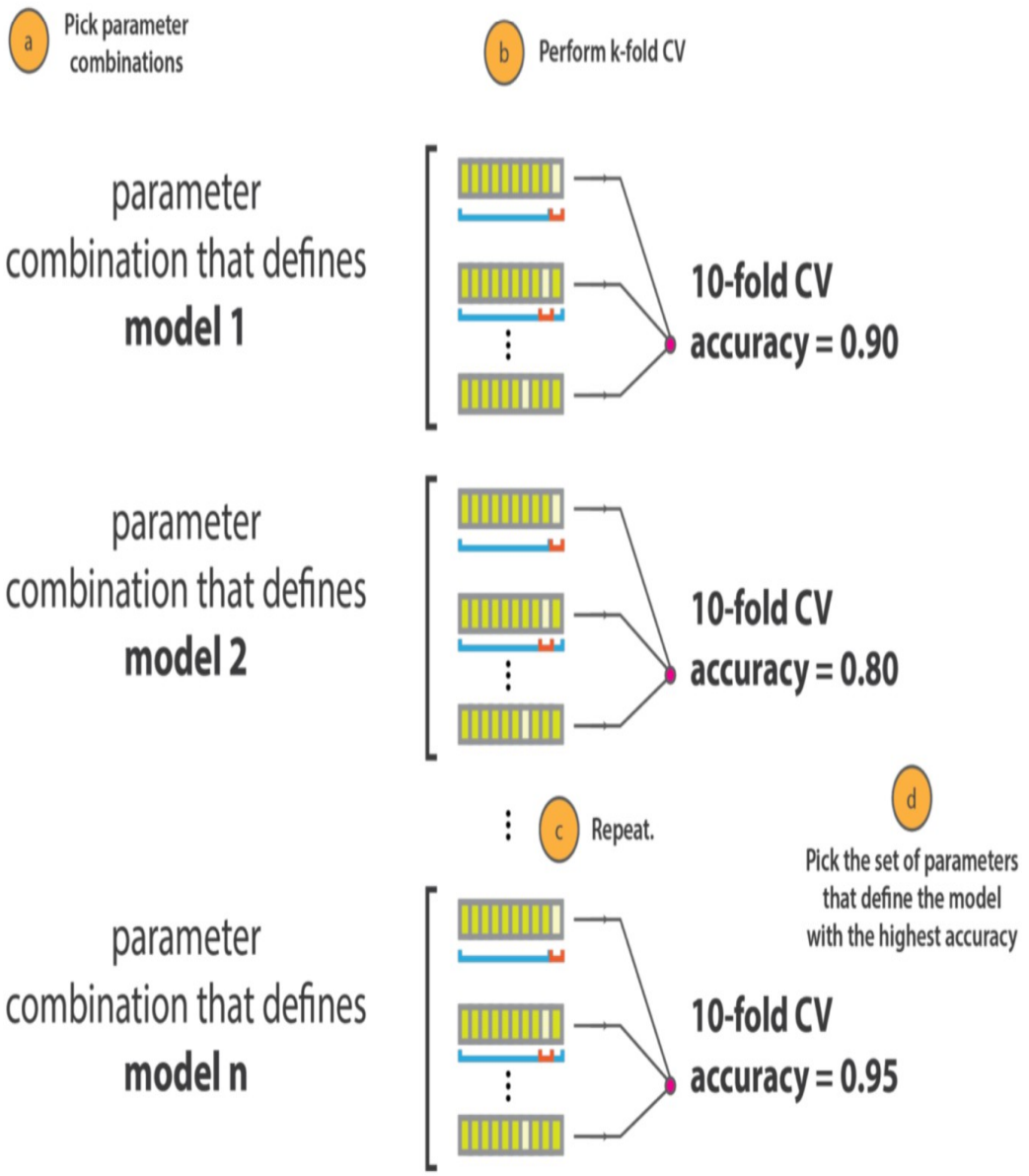
¬ Zou onderdeel engineering en transformative op de gehele data set moeten worden uitgevoerd? (*het is eenvoudig van het niet te doen*)

¬ Té veel overgebruik of training/validatie verdeelde hinten op verborgen overtraiging (*Ik zal maar een kleinere parameter weerkeren*)

¬ Dus te veel parameter combinatie runs (over-gebruik van dezelfde data) (*Het is in orde, ik kruis-valideer na elk gebruik)*

¬ Sommige modellen proberen over gepastheid van zichzelf te vermeiden (denk aan: bootstrapping)

¬ Ook, als scores te goed om waar te zijn lijken, dan zijn ze dat meestal (doel variabele “lekkage”)



Operationele efficiëntie en economische kost

¬ Operationele efficiëntie is gerelateerd tot de hoeveelheid moeite nodig is om alles te evalueren, monitoren hertesten of herbouwen van een model

¬ Vanuit dit perspectief is het redelijk duidelijk dat een neural network of een random bos minder efficient is dan een leeg vanilla regressive model of keuze boom

¬ In bepaalde instellingen zoals kredietkaart fraude detectie, zijn operationele efficiëntie zeer belangrijk omdat een keuze binnen enkele luttele seconden moet worden gemaakt na een transactie is ingezet

¬ Economische kost wijst op de kost die nodig is om de model inputs-, het model te runnen en process uitkomsten te verzamelen

¬ Alsook de kosten van externe data en/of modellen horen hierbij

¬ De economische opbrengst op het analytisch model berekenen is geen makkelijke oefening

¬ Technische schulden en onderhoud… Zullen modellen over 1 jaar nog werken? Over 10?

Andere zorgen

¬ Reguliere opmerkingen: bv. Geen black-boxes, geen gebruik van bepaalde onderdelen

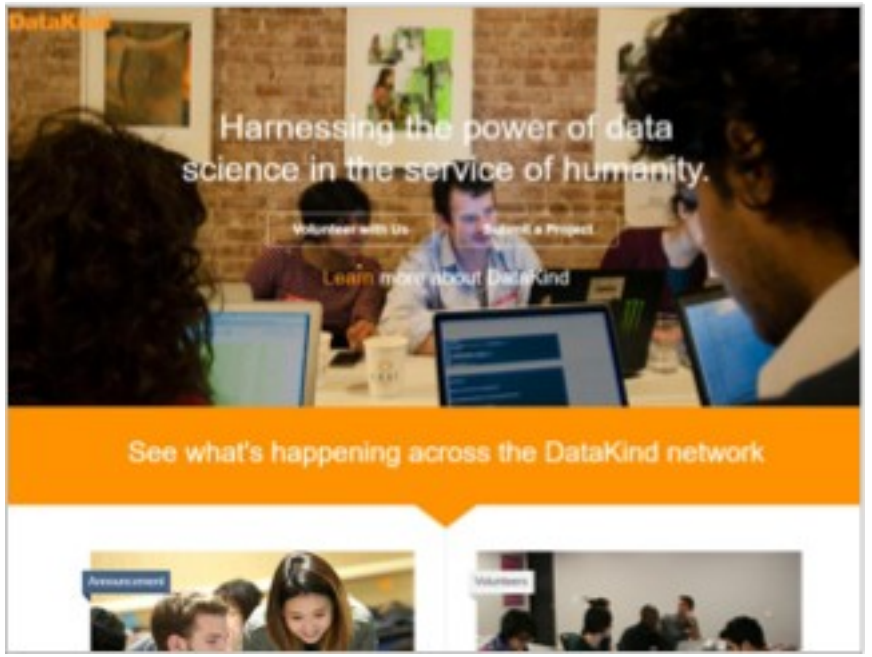
¬ Vraag over ethische warden… Questions on ethics...

¬ Kan een algorithme racistisch zijn? Seksistisch?

¬ “Zullen voorspellende modellen niet aansluiten bij de nieuwe gemeenschappen?”

¬ Bedrijven zoals DataKind en Bayes Impact

¬ Concept van “open modellen”



Data Mining: Where Legality and Ethics Rarely Meet

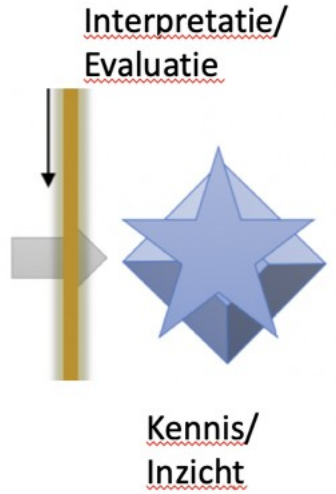
By Kelly Shermach
Aug 25, 2006 4:00 AM PT

Print
Email

More than ever, knowingly or unknowingly, consumers disseminate personal data in daily activities. Credit and debit card transactions, ATM visits, Web site browsing and purchases -- even mobile phone use -- all generate data downloaded for analysis and customer profiling. Collectors may use this data to enhance customers' experience, but may also share information with marketers more focused on customer acquisition.

Samenvatting

- Business relevantie
- Statistische performantie en juistheid
- Operationele efficiëntie en economische kost
- Reguliere observaties



Samenvatting

¬ During the first KDD Cup 1997, the goal was to select a subset of lapsed donors to contact. An entry from one well-known company selected the worst possible candidates for mailing - their results were significantly worse than random!
Apparently their data miners switched the sign somewhere. Fortunately for them, the names of the contestants were kept anonymous. (Gregory Piatetsky-Shapiro)

¬ In the 1980s, the Pentagon wanted to harness computer technology to make their tanks harder to attack.
The preliminary plan was to fit each tank with a digital camera hooked up to a computer. The computer would continually scan the environment outside for possible threats (such as an enemy tank hiding behind a tree), and alert the tank crew to anything suspicious. The only possible way to solve the problem was to employ a neural network. The research team went out and took 100 photographs of tanks hiding behind trees, and then took 100 photographs of trees - with no tanks. They took half the photos from each group and put them in a vault for safe-keeping, then scanned the other half into their mainframe computer. The huge neural network was fed each photo one at a time and asked if there was a tank hiding behind the trees. Over time it got better and better until eventually it was getting each photo correct. It could correctly determine if there was a tank hiding behind the trees in any one of the photos.
But the scientists were worried: had it actually found a way to recognize if there was a tank in the photo, or had it merely memorized which photos had tanks and which did not? So the scientists took out the photos they had been keeping in the vault and fed them through the computer. To their immense relief the neural net correctly identified each photo as either having a tank or not having one.
The Pentagon was very pleased with this, but a little bit suspicious. They commissioned another set of photos (half with tanks and half without) and scanned them into the computer and through the neural network. The results were completely random. For a long time nobody could figure out why. After all nobody understood how the neural had trained itself.
Eventually someone noticed that in the original set of 200 photos, all the images with tanks had been taken on a cloudy day while all the images without tanks had been taken on a sunny day. The neural network had been asked to separate the two groups of photos and it had chosen the most obvious way to do it - not by looking for a camouflaged tank hiding behind a tree, but merely by looking at the colour of the sky. The military was now the proud owner of a multi-million dollar mainframe computer that could tell you if it was sunny or not.

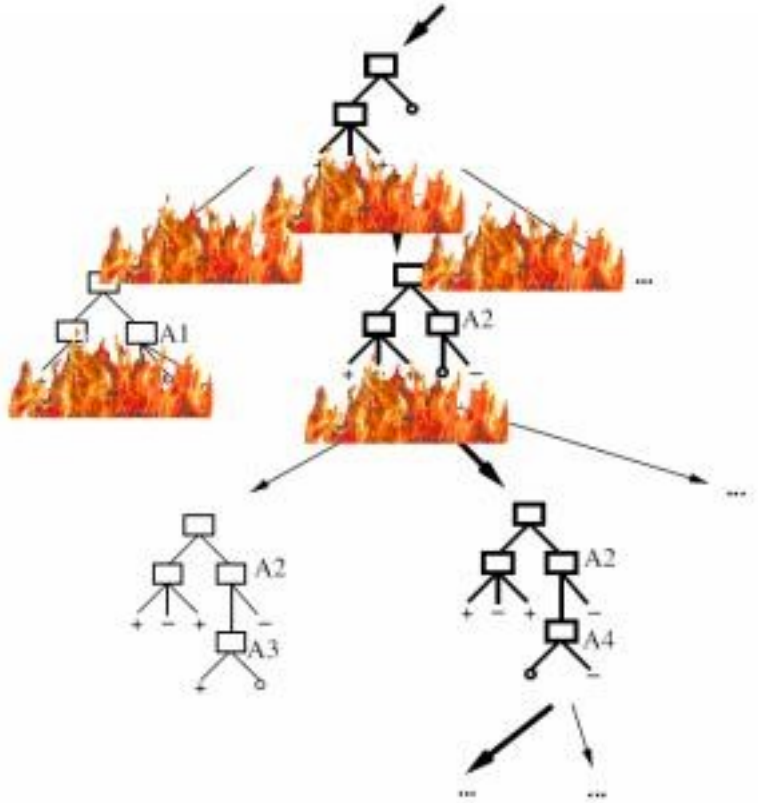
Samenvatting

¬ Occasionally researchers using pruning algorithms on their decision trees get carried away. Instead of pruning unnecessary branches in the interests of reducing overfitting. The experimenter just burns down the tree until it is a decision stump (Data Mining Disasters: a report)

¬ In pre-Big Data days, for example, a hotel chain used some pretty sophisticated mathematics, data mining, and time series analysis to coordinate its yield management pricing and promotion efforts. The forecasting models—which were marvels—mapped out revenues and margins by property and room type. The projections worked fine for about a third of the hotels but were wildly, destructively off for another third.

The forensics took weeks; the data were fine. Were competing hotels running unusual promotions that screwed up the model? Nope. For the most part, local managers followed the yield management rules.

Almost five months later, after the year's financials were totally blown and HQ's credibility shot, the most likely explanation materialized: The modeling group—the data scientists of the day—had priced against the hotel group's peer competitors. They hadn't weighted discount hotels into either pricing or room availability. For roughly a quarter of the properties, the result was both lower average occupancy and lower prices per room (Learn from Your Analytics Failures, HBR)



!Let op met overmatig gebruik van algorithmen want dan kan het snel mislopen!