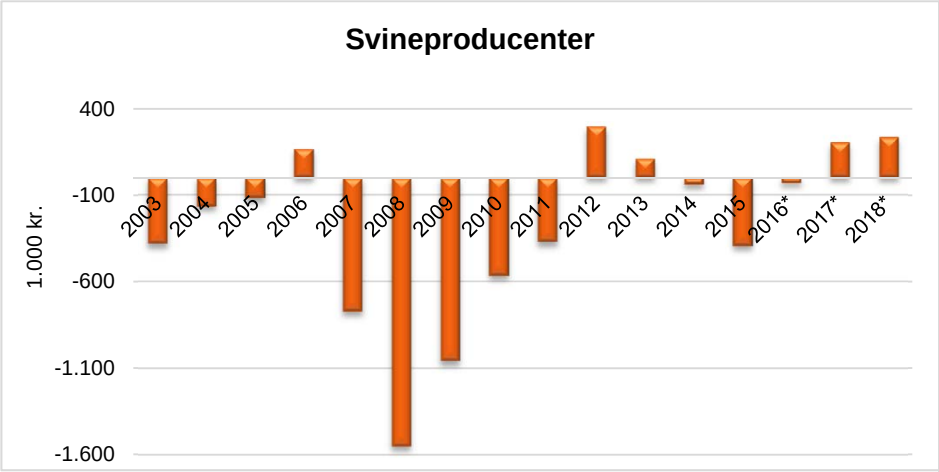
Økonomikonference

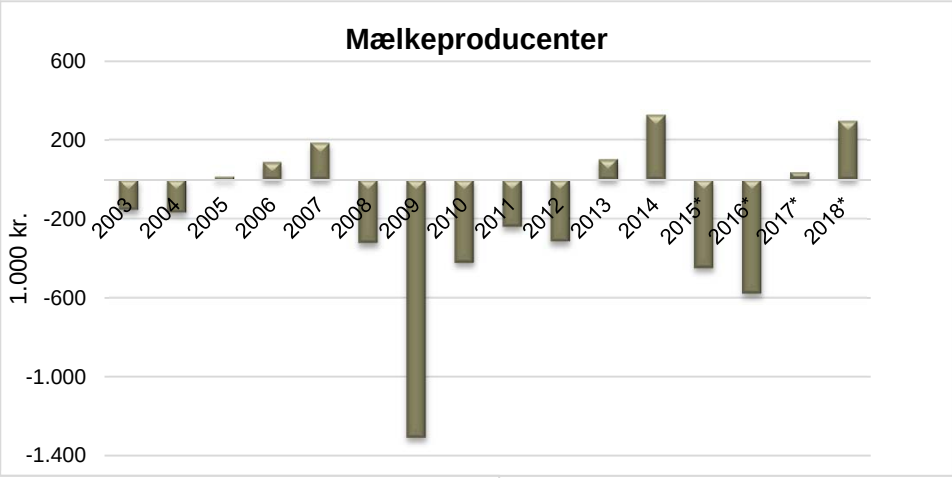
7. oktober 2016

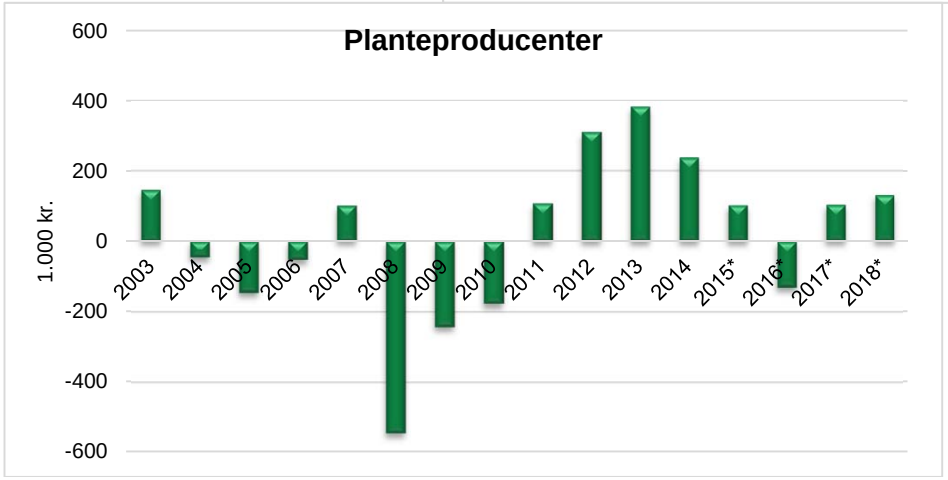
v/ Philipp Trénel, DTI
og
Klaus Kaiser, SEGES P/S

KUNSTEN AT FORUDSIGE KONKURSER

LANDBRUGET ER MEGET KONJUNKTURFØLSOM

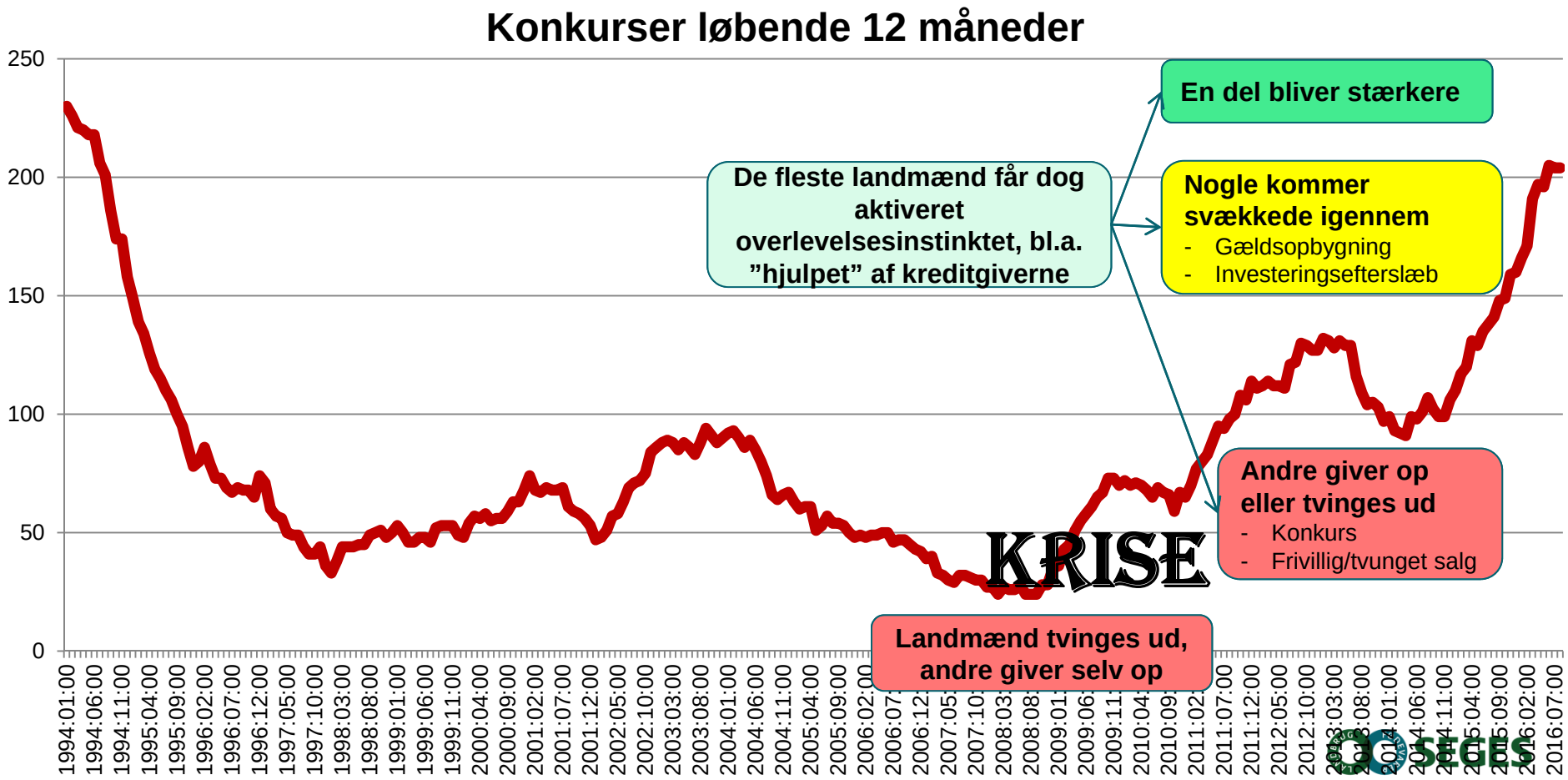






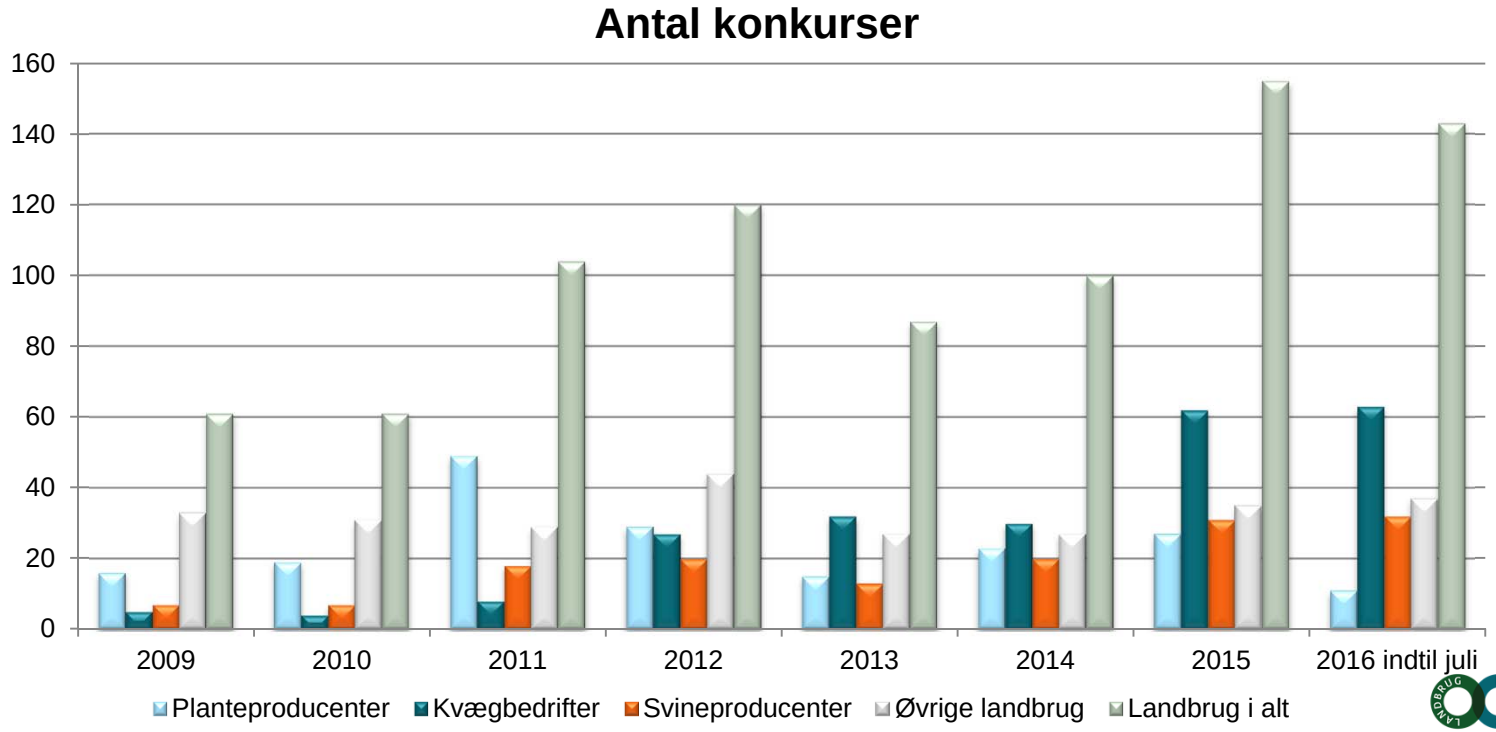
SEGES

928 KONKURSER I LANDBRUGET SIDEN FINANSKRISENS START I 2008



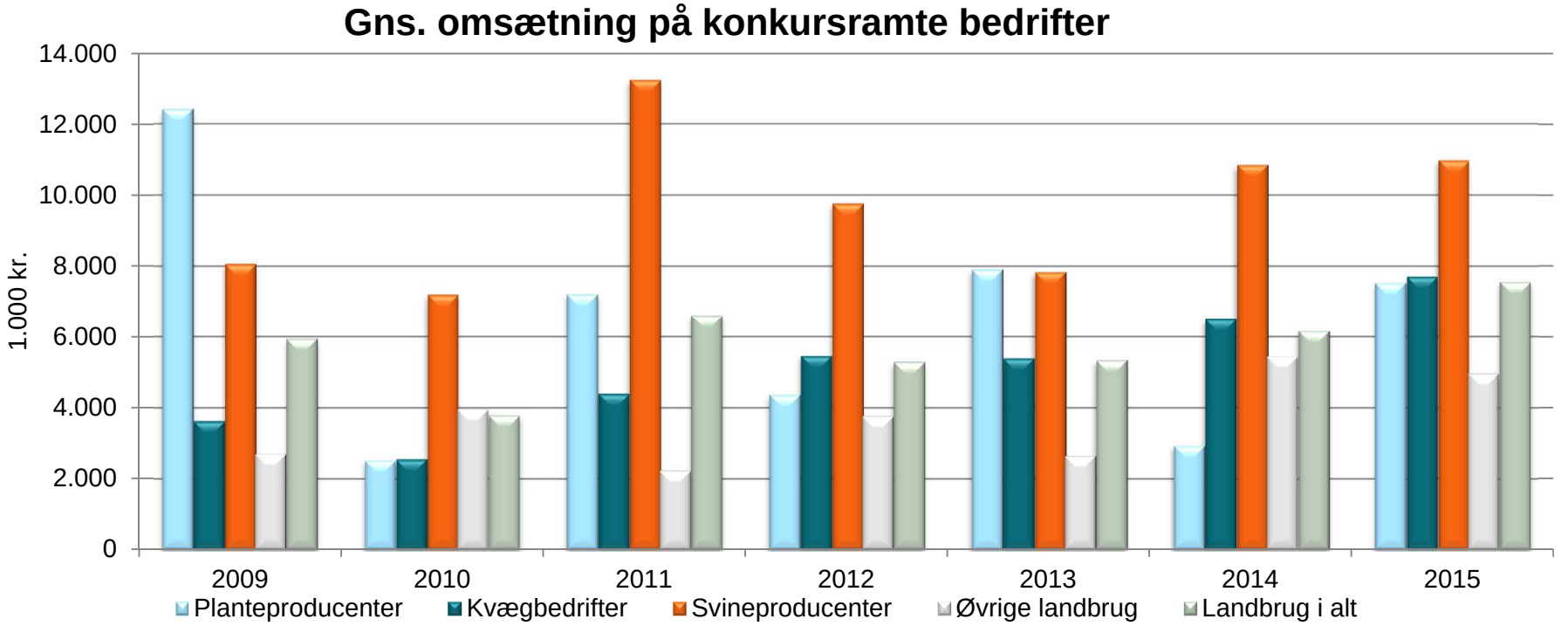
KONKURSER – OPDELT PÅ DRIFTSGRENE

- Kvægbedrifter er særligt hårdt ramt
- Men også høje konkurstal blandt svineproducenter og øvrige

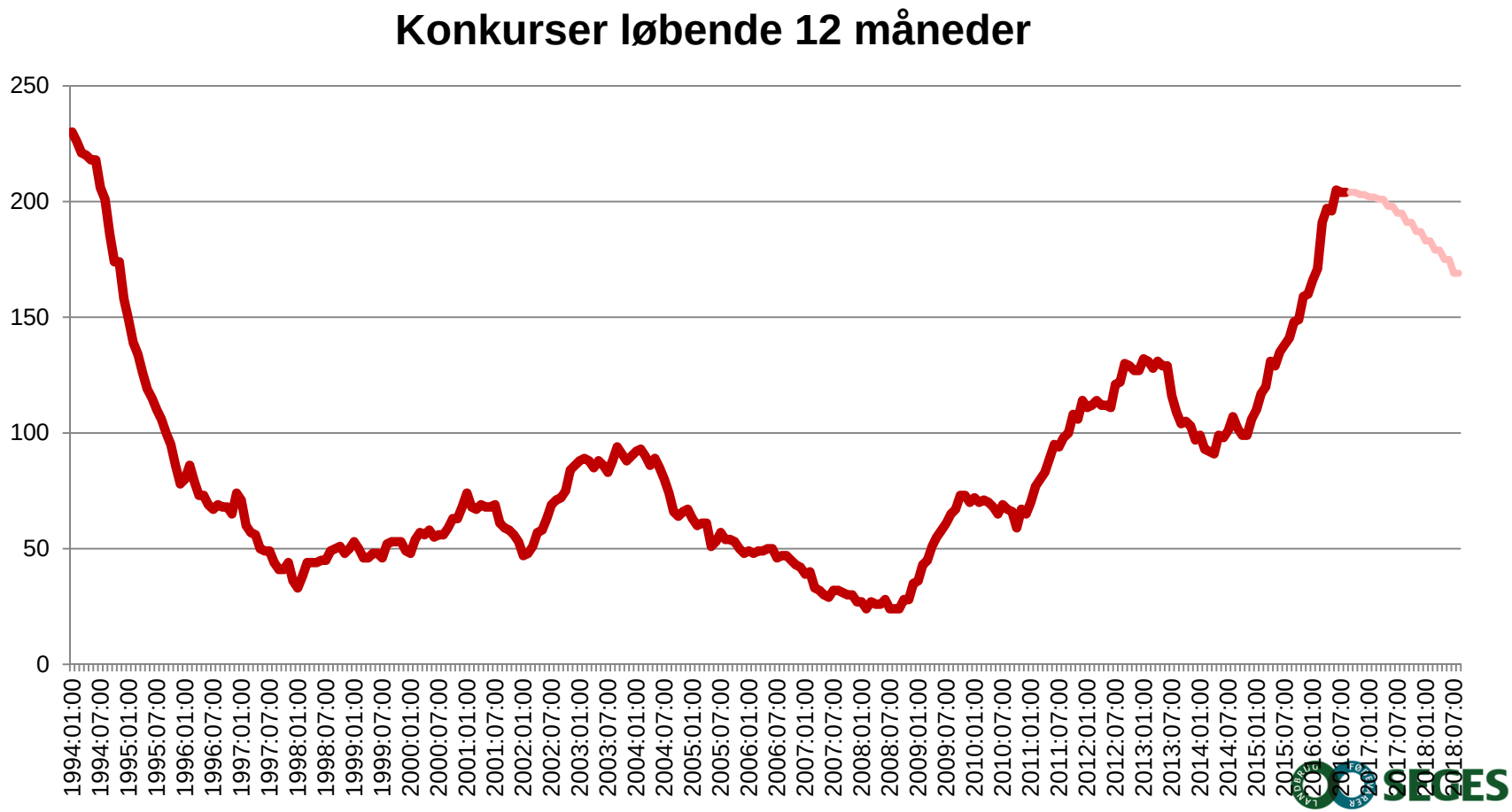


KONKURSRAMT OMSÆTNING

- Konkursramte svineproducenter er gns. større



FORTSAT HØJT KONKURSNIVEAU FORVENTES, MEN HVILKE BEDRIFTER?



TEKNOLOGISK INSTITUT

Konkursmodeller for de danske landbrugserhverv

Philipp Trénel, ph.d., seniokonsulent
Teknologisk Institut – division AgroTech

TEKNOLOGISK INSTITUT

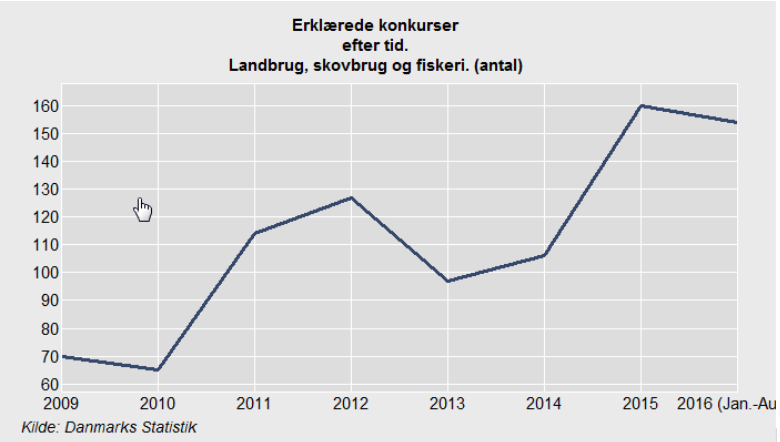
Formål

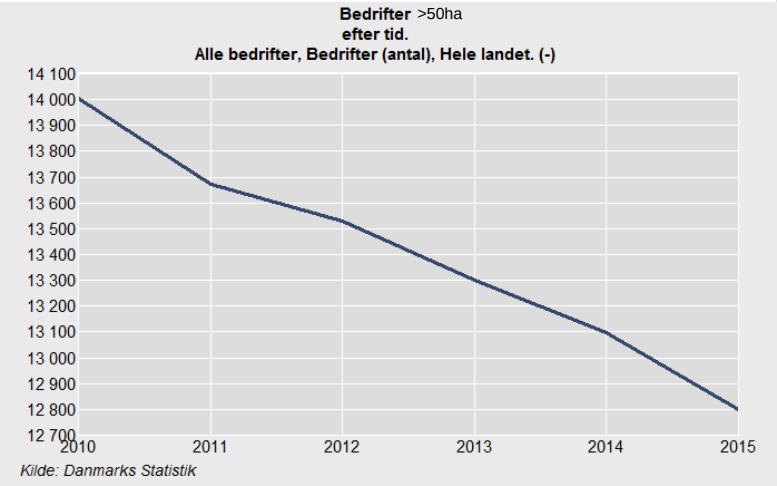
- At udvikle konkursmodeller til prædiktion af konkurssandsynlighed indenfor fem af det danske landbrugserhvervs sektorer (søer, slagtesvin, planteavl, kvægdrift) på baggrund af data fra DLBRs økonomi-database (ØDB) og konkursdata indhentet af SEGES.

- Konkursmodellerne skal levere både en høj prædiktionsnøjagtighed (accuracy) og en nem faglig tolkning mhp. en senere anvendelse i rådgivningen.

- Konkursmodellerne udvikles med henblik på opbygning af et early-warning system.

TEKNOLOGISK INSTITUT





Udfordring 1

- Det er svært at spå – især om fremtiden
- ... og i særdeleshed om sjældne hændelser i fremtiden.

- Konkurser i landbruget i DK:
  → lav prævalens (mellem 0.5 og 2.5%)
  (i 2015: 1.2 %, Danmarks Statistik)
- Dette afspejler sig i ØDB data brugt her
  (prævalens ~ 0.8%)

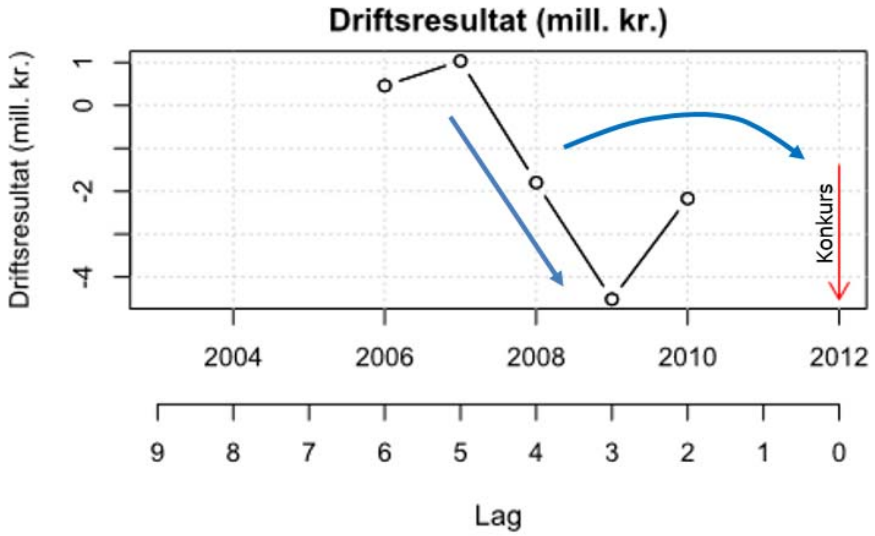
- Mange control-observationer (n = 6034 bedrifter)
- Få case-observationer (n = 110 bedrifter)
- → mindsket statistisk power og reliability (generaliserbarhed)

TEKNOLOGISK INSTITUT

Udfordring 2

- Der er mange mulige forklarende variabler i ØDB, + deres features, til at bygge modellen på. (70 variabler, i alt 1.330 features)
- Få af dem forventes at have en høj forklaringsværdi
- → Behov for variable selection/model reduktion i et høj-dimensionelt parameterrum
- ”The curse of dimensionality”

TEKNOLOGISK INSTITUT



Feature	Acronym
Niveau	lag
Hældning, 1º difference	d
Acceleration, 2º difference	dd
Løbende variation i 3 års tidsvindue	SD

Udfordring 3

- Konkurssandsynligheden er en funktion af hændelser igennem tid
- Input-variabler til modellen skal derfor indeholde de features/egenskaber i økonomi-nøgletallenes historik, som har en sammenhæng til en fremtidig konkurshændelse.
- Dette gælder både korttids- og langtids- memory-processer

- → ”omitted variable bias” (OVB)

TEKNOLOGISK INSTITUT

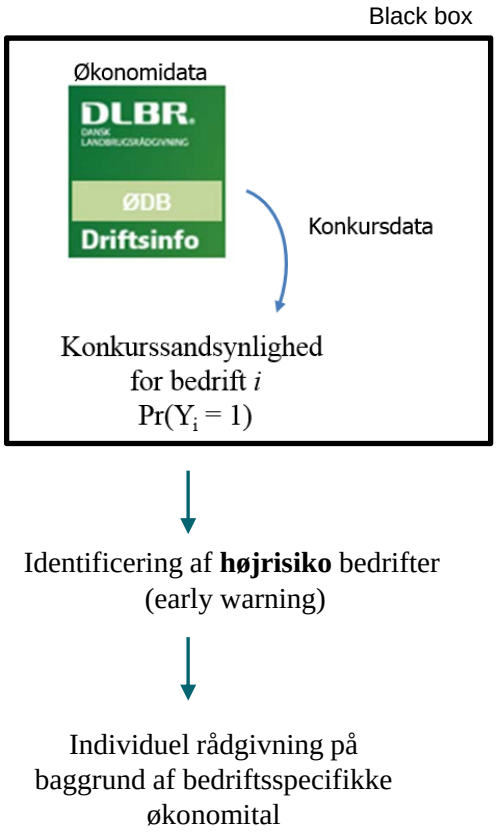
Udfordring 4

Datakvalitet i ØDB

- Støj i data
  - manuelle indtastninger
  - personafhængige vurderinger
  - manglende oplysninger
  - varierende præcision af oplysninger, m.m.
- → ikke-reducerbar støj → mindsket power

- Missing data
  - En stor andel af bedrifter viser store andele af manglende oplysninger
    (36.6% missing data i datasættet)
  - Kan missing data antages at være **missing at random** (MAR)?
  - Missing not at random (MNAR)
    → sampling bias

TEKNOLOGISK INSTITUT

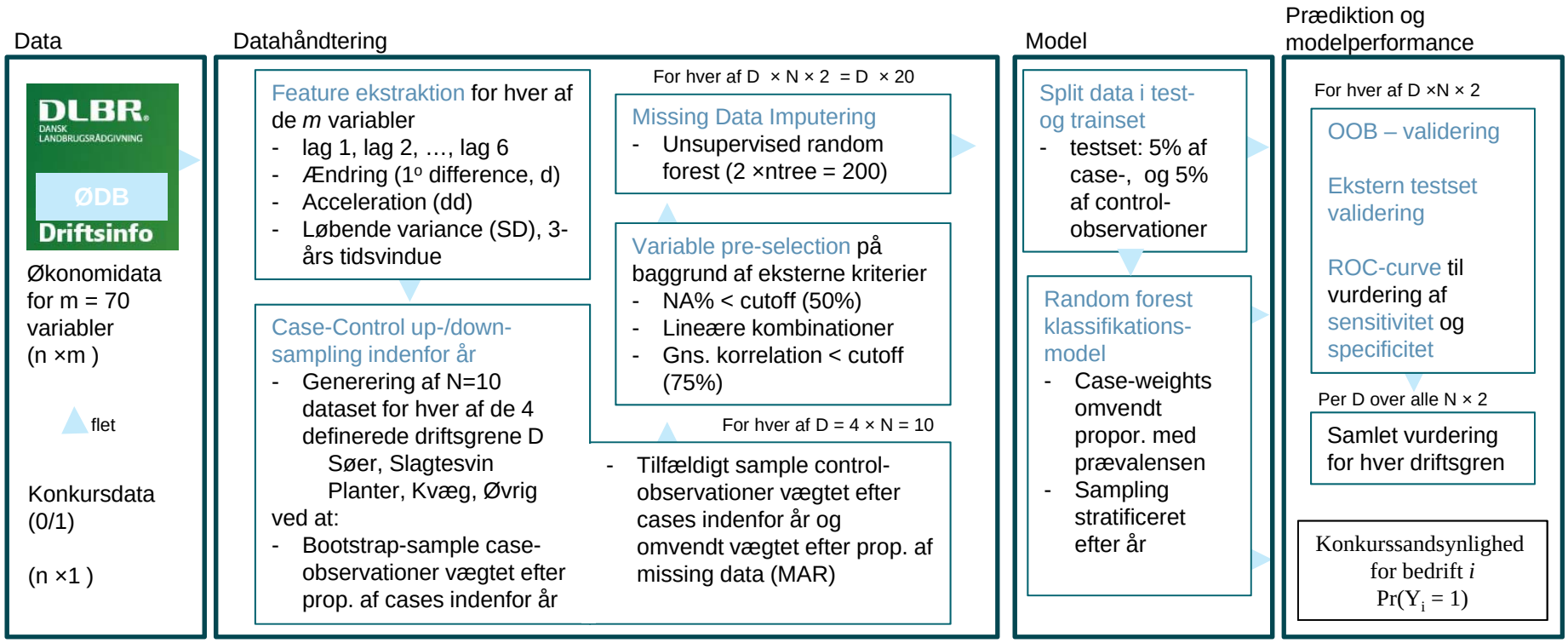


Udfordring 5

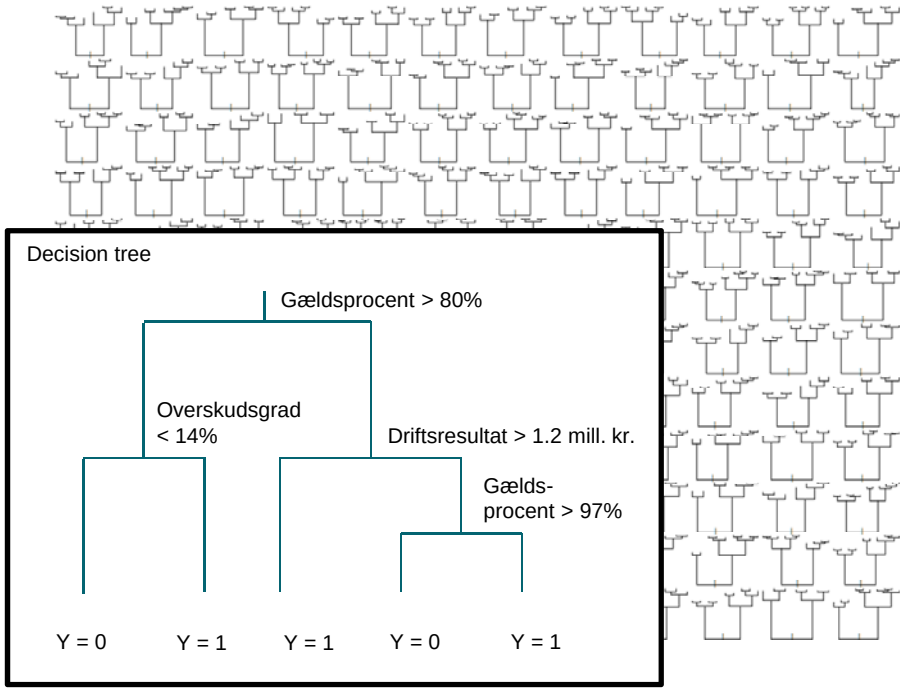
- Høj prædiktionsnøjagtighed opnås typisk med **black-box** modeller
  (machine learning, e.g. random forest, deep learning neurale netværk, boosting, m.m.)
- En høj grad af menneskelig fortolkbarhed opnås typisk ved (ofte urealistisk) simple modeller (e.g., lineær regression, logistisk regression, m.m.)
- ”The complexity monster” (Pedro Domingos)

TEKNOLOGISK INSTITUT

Datamodel/strategi



TEKNOLOGISK INSTITUT



Hypotetisk eksempel

Hvad er en random forest?

- Breiman 2001
- Et af de mest populære og succesfulde machine learning teknikker
- Bygger et stort antal (derfor 'forest') nær-uafhængige (derfor 'random') decision-tree modeller.
- Hver decision-tree prædikterer outputtet på baggrund af split-regler i inputparametrene.
- Den endelige prædiktion er majority-voten over alle træer i skoven.
- Nær-uafhængighed opnås vha. resampling-teknikker vedr. både observationer og inputvariabler.
- Metoden inkorporerer automatisk variable-selection/model reduktion og validering (out-of-bag, OOB).

TEKNOLOGISK INSTITUT

Input-variabler: ØDB historik 2006 - 2015, Konkurs-historik 2012 - 2015

m inputvariabler = 70 *k* = features pr. m = 19 *m* × *k* = 1330 variabler i alt

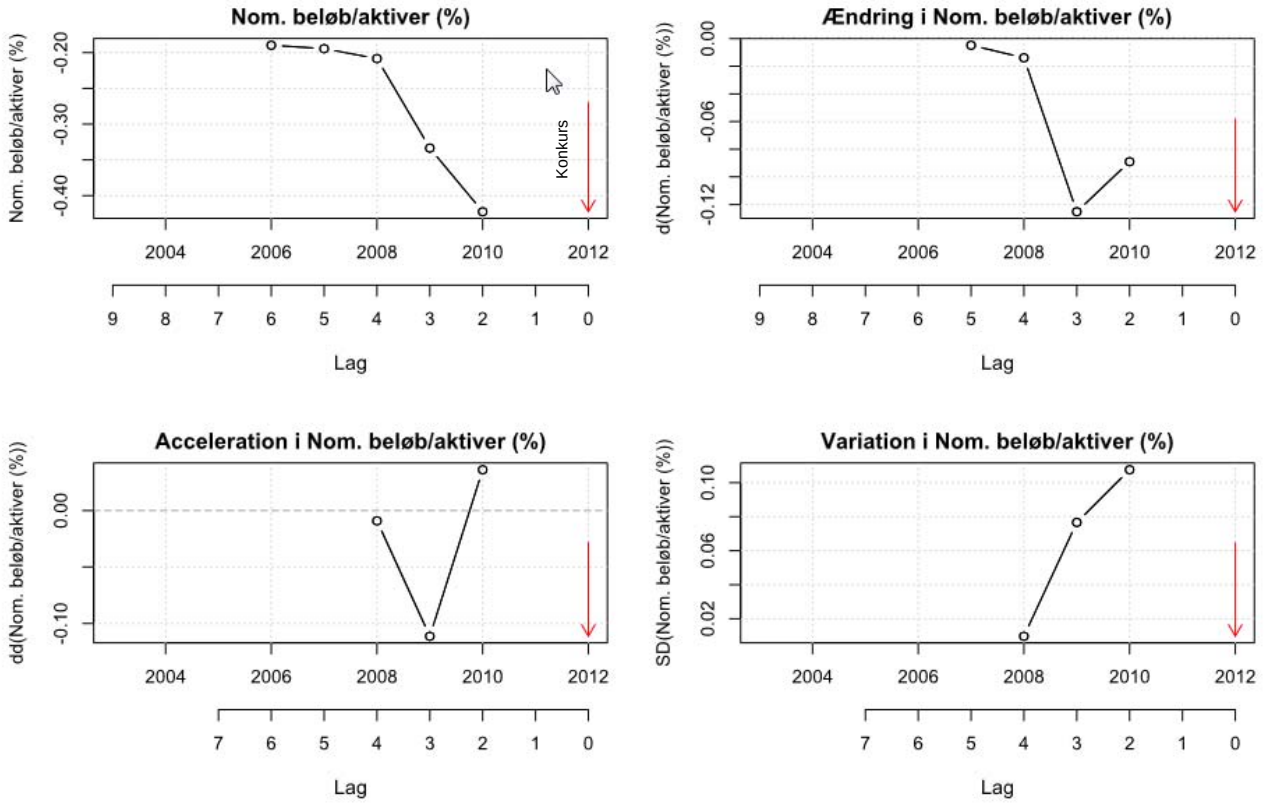
Indtjening/rentabilitet/ likviditet	**Gæld/egenkapital**	**Aktiver**	**Passiver/Finansielt**	**Produktionsnøgletal**	**Andet**
Nominelle beløb: - EBIT(DA) - Driftsresultat - Likviditet - Working capital	Nominelle beløb: - Restgæld - Egenkapital	Aktivudnyttelse: - Div. Indtjeningsmål/aktiver - Omsætning/aktiver - Kortfristet gæld/aktiver - Kortfristet gæld/likvide aktiver - Omsætningsaktiver/aktiver	Passivudnyttelse: - Div. indtjeningsmål/ nettorentebærende gæld - Div. indtjeningsmål/ passiver - Renteomk./omsætning - Egenkapitalforrentning - Rentedækningsgrad	Produktivitet - Udbytte pr. ha. - Mælkeydelse pr. ko - Grise pr. årsso	Demografi mv.: - Etableringsår - Alder - Ejendomsstørrelse - Landsdel/kommune - Driftsgren - Økologi - Hvilken DLBR-virks.
	Soliditet		WACC		Nominelt rådgivningsbeløb
Rentabilitetsmål: - Kapacitetsgrad - Dækningsgrad - Overskudsgrad - Afkastningsgrad - ROIC - Likviditet/omsætning - Driftsres./omsætning - Lønningsevne	Gældsprocent	Kapacitet: - Kap.omk./samlede aktiver - Kap.aktiver/samlede aktiver	Gældssammensætning: - Andel afdragsfrihed - Andel rentetilpasning - Andel PI/RKI - "Anden gæld"/gæld - Loan-to-value (RKI-gæld/aktiver)	Effektivitet - DB pr. enhed - Fremstillingspris pr. enhed	Rådgivningsintensitet: - Nom. beløb/omsætning - Nom. beløb/aktiver - Nom. beløb/gæld
	(Kortsigtet gæld/likviditet) * 365		Kreditinstitut		Rådgivningsværktøjer: - Produktionsrådgivning - Økonomirådgivning - Ingen rådgivning
			Morarenter		

TEKNOLOGISK INSTITUT

Rådgivningsintensitet

Eksempel på

Feature ekstraktion



TEKNOLOGISK INSTITUT

Resultater - model performance og prædiktionsnøjagtighed (1 år frem)

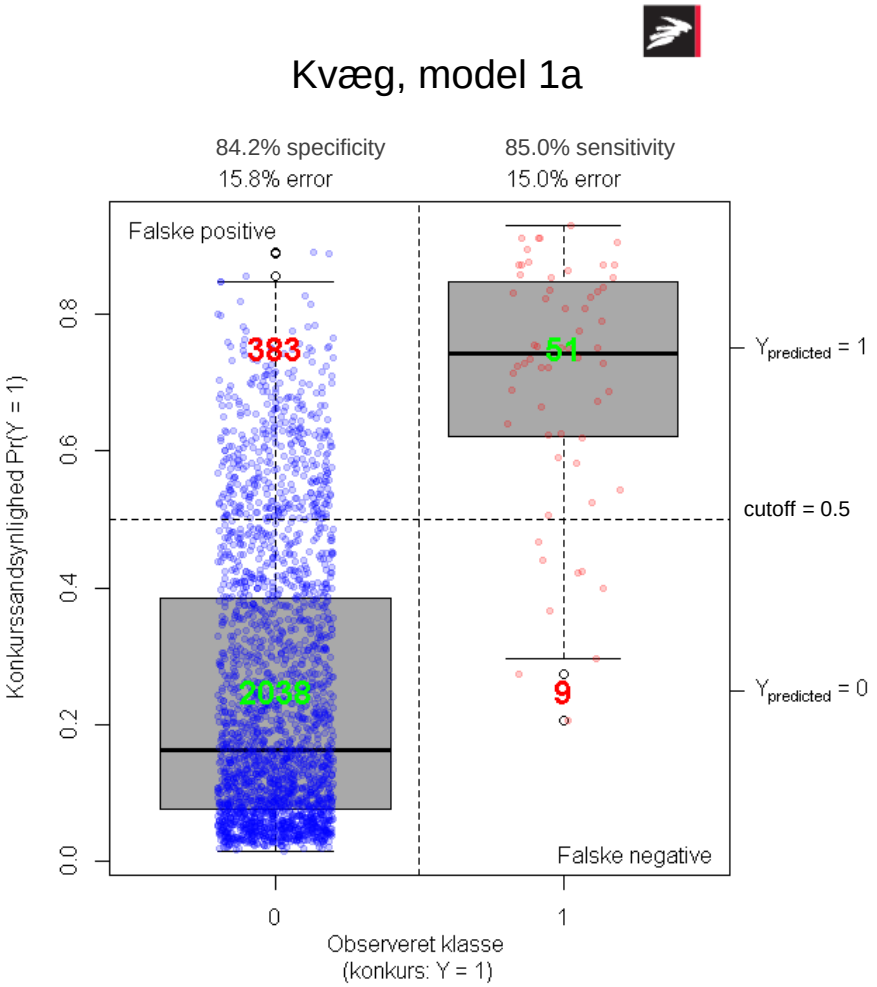
Driftsgren	Statistic	estimate	s.e.	N
Kvægbedrifter	Y=0	2687,6	1,6	20
	Y=1	73,0	2,2	20
	Prevalence (%)	2,64	0,08	20
	Accuracy (OOB)	**0,869**	0,012	20
	Sensitivity[1] (OOB)	0,879	0,031	20
	Specificity[1] (OOB)	0,869	0,012	20
	Accuracy (test set)	0,874	0,023	20
	Sensitivity[1] (test set)	0,882	0,098	20
	Specificity[1] (test set)	0,873	0,025	20
So-besætninger	Y=0	1015,0	3,9	20
	Y=1	19,0	0,0	20
	Prevalence (%)	1,84	0,01	20
	Accuracy (OOB)	**0,84**	0,024	20
	Sensitivity[1] (OOB)	0,865	0,081	20
	Specificity[1] (OOB)	0,839	0,025	20
	Accuracy (test set)	0,852	0,045	20
	Sensitivity[1] (test set)	0,948	0,112	20
	Specificity[1] (test set)	0,848	0,047	20

Driftsgren	Statistic	estimate	s.e.	N
Slagtesvinsproduktion	Y=0	994,7	8,0	20
	Y=1	14,0	0,0	20
	Prevalence (%)	1,39	0,01	20
	Accuracy (OOB)	**0,858**	0,021	20
	Sensitivity[1] (OOB)	0,819	0,098	20
	Specificity[1] (OOB)	0,859	0,021	20
	Accuracy (test set)	0,853	0,045	20
	Sensitivity[1] (test set)	0,95	0,224	20
	Specificity[1] (test set)	0,85	0,045	20
Planteavl	Y=0	1207,7	1,9	20
	Y=1	8,0	0,0	20
	Prevalence (%)	0,66	0,00	20
	Accuracy (OOB)	**0,871**	0,035	20
	Sensitivity[1] (OOB)	0,853	0,101	20
	Specificity[1] (OOB)	0,871	0,036	20
	Accuracy (test set)	0,9	0,031	4
	Sensitivity[1] (test set)	1	0	4
	Specificity[1] (test set)	0,899	0,032	4

[1]:cutoff-sandsynlighed= 0.5

Sensitivitet og specificitet

- En sensitiv model har få falske negative.
- En specifik model har få falske positive.
- Den ønskede balance mellem sensitivitet og specificitet er bruger/case-afhængig og opnås vha. den cutoff-prædiktions-sandsynlighed, som klassificerer observationerne i kategorierne
  Y = 1 (konkurs) eller
  Y = 0 (ikke-konkurs).

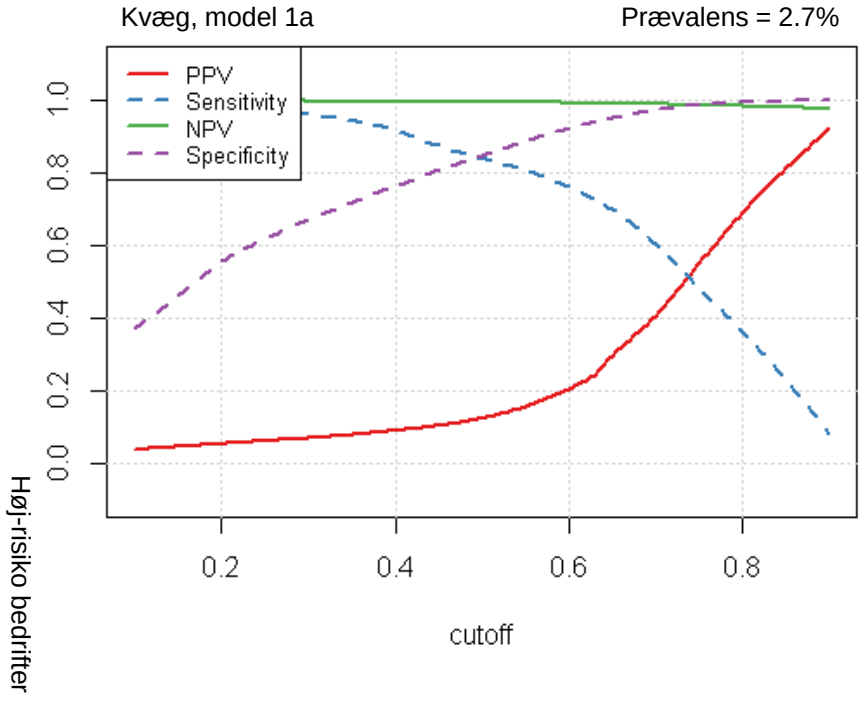


TEKNOLOGISK INSTITUT

Sensitivitet vs. Positive Predicted Value (PPV)

- ”Hvad er sandsynligheden for konkurs for en bedrift, der af modellen forudsiges til at gå konkurs?”
- Dette afhænger af prævalensen!
- PPV er en prævalens-korrigeret analog til sensitivitet; NPV til Specificitet.

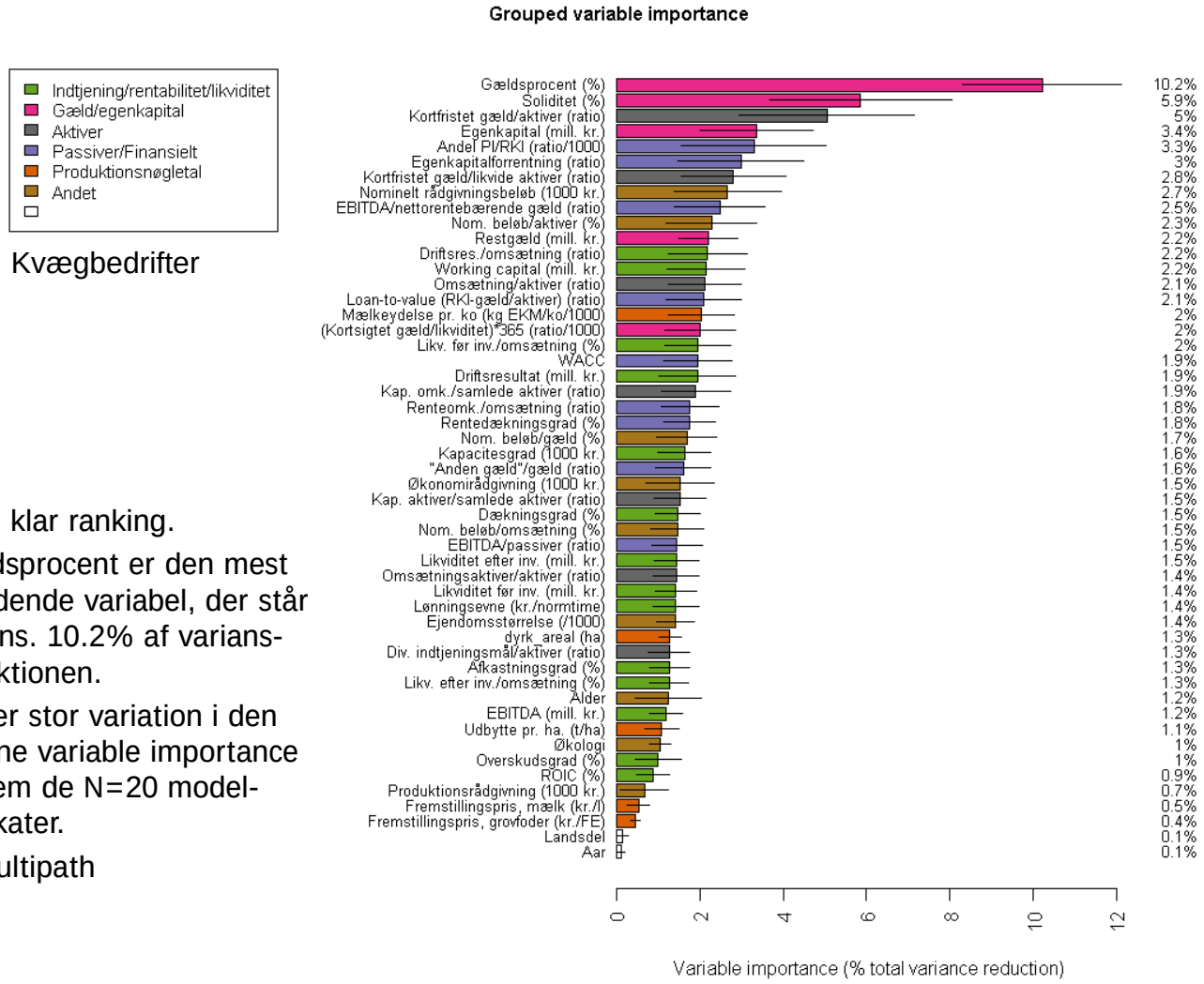
	Ex1	Ex2	Ex3
cutoff	**0.5**	0.74	0.82
Sensitivity	0.840	**0.509**	0.310
Specificity	0.858	0.986	0.997
PPV	**0.122**	**0.520**	**0.750**
NPV	0.995	0.988	0.983



TEKNOLOGISK INSTITUT

Resultater - prædiktionsnøjagtighed som funktion af forecast-tidsvinduet

Resultater - variable importance

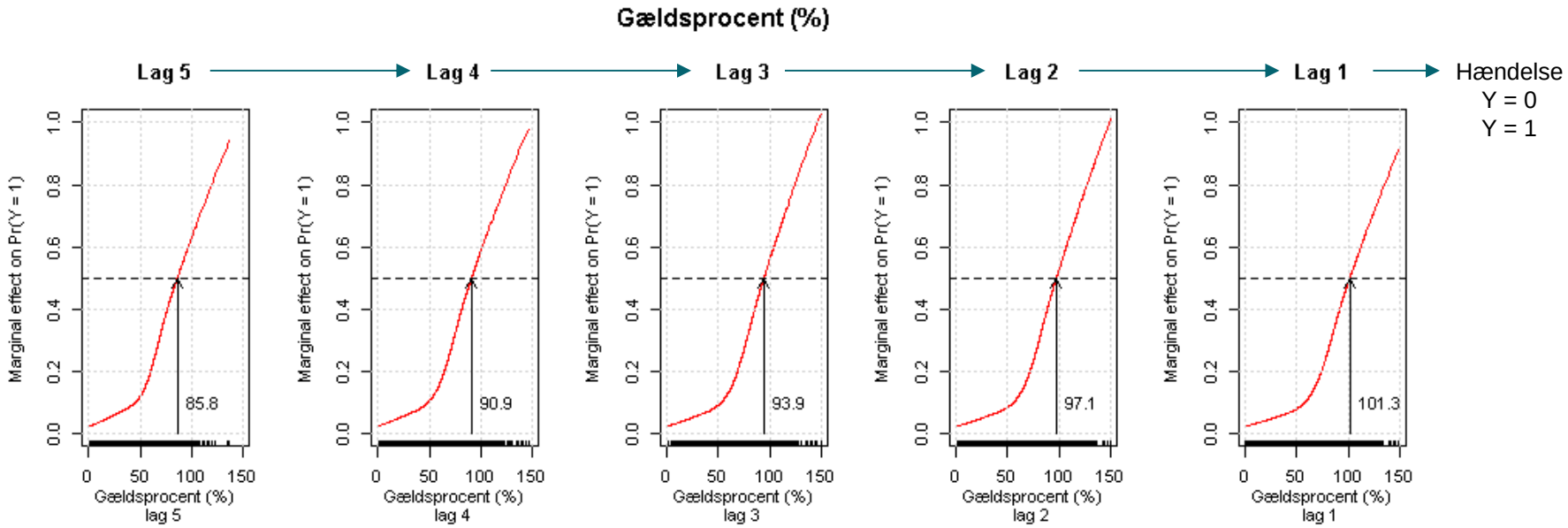


Kvægbedrifter

- Viser klar ranking.
- Gældsprocent er den mest betydende variabel, der står for gns. 10.2% af varians-reduktionen.
- Der er stor variation i den fundne variable importance mellem de N=20 model-replikater.
- → multipath

TEKNOLOGISK INSTITUT

Resultater - marginale effekter af de mest betydende variabler (1/2)



TEKNOLOGISK INSTITUT

Konklusioner

- Random Forest modeller kan med rimelig høj nøjagtighed (accuracy = [0.84, 0.87]) prædiktere en bedrifts konkurssandsynlighed på baggrund af DLBRs ØDB data.

- Den lave prævalens af konkurser er en udfordring.
  PPV'en kan øges på bekostning af sensitiviteten til en identifikation af højrisiko-bedrifter (f.eks. PPV ~ 0.75 og sensitivitet ~ 0.3 for kvægbedrifter).

- Der er mange udfordringer forbundet med en Big Data analyse.
  En succesfuld/brugbar model skal imødekomme udfordringer:
  - Prævalens
  - Data kvalitet og missing data
  - Features extraction
  - Variable selection
  - Honest test set based validation
  - Tæmning af "the complexity monster" – modellen skal være let at fortolke

- Den faglige tolkning er først begyndt.

KUNSTEN AT FORUDSIGE KONKURSER