

Konkursmodeller for de danske landbrugserhverv – version 1 (2016) + 2 (sept. 2017)

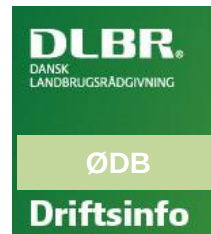
Klaus Kaiser, erhvervsøkonomisk chef
SEGES Erhvervsøkonomi

Philipp Trénel, ph.d., seniokonsulent
Søren Kjærgaard Boldsen, ph.d., specialist
Teknologisk Institut – division AgroTech

Formål

- At udvikle konkursmodeller til prædiktions af konkurssandsynlighed indenfor fire 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.

Økonomidata

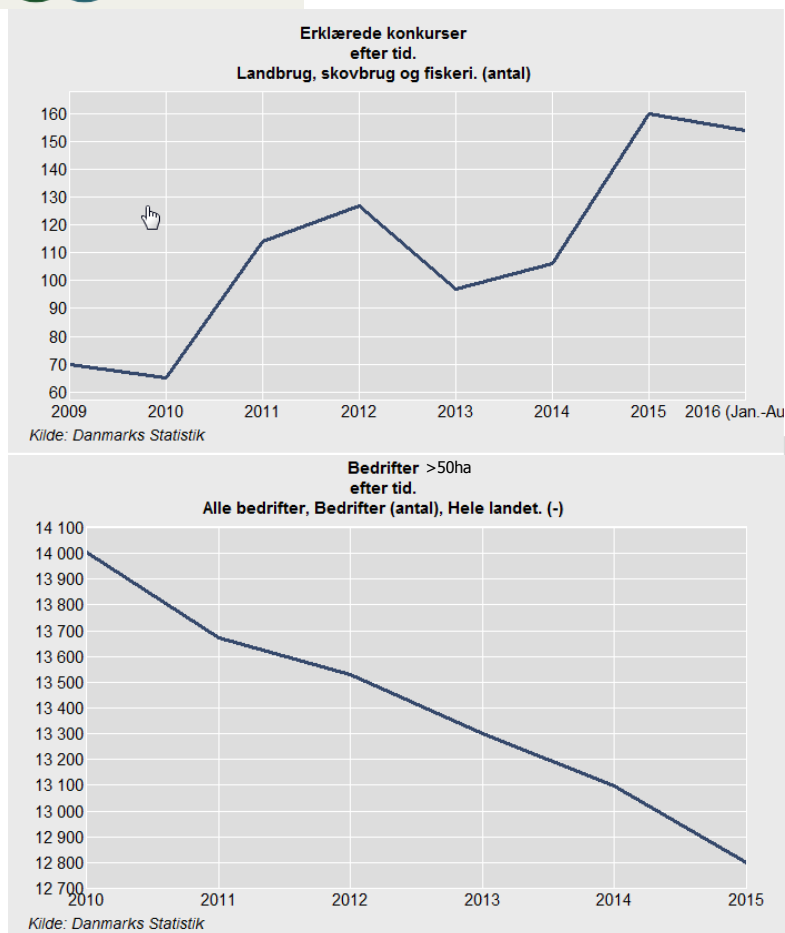


Konkursdata

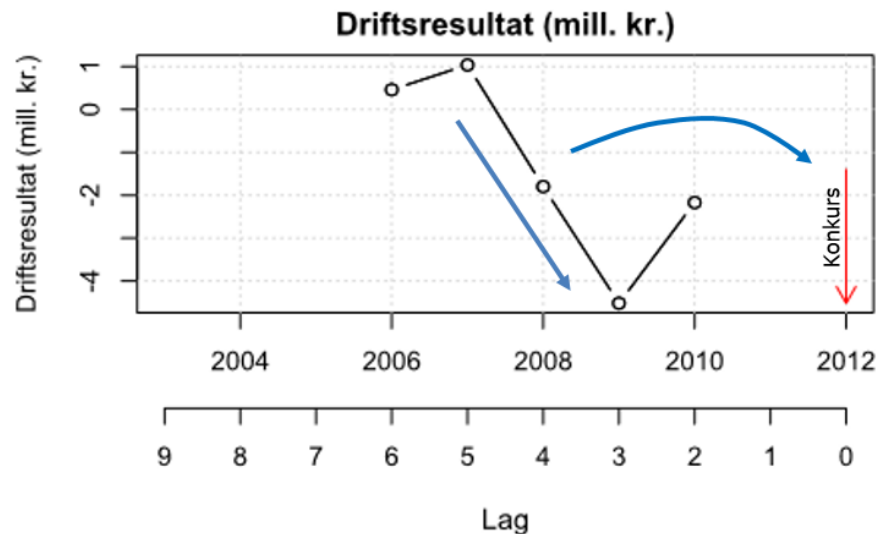
Konkurssandsynlighed
for bedrift i
 $\Pr(Y_i = 1)$

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 ($n = 420$ bedrifter)
- → mindsket statistisk power og reliability
(generaliserbarhed)







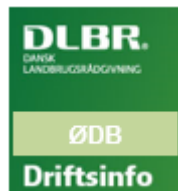
Udfordring 3

- Konkursandsynligheden 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)

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

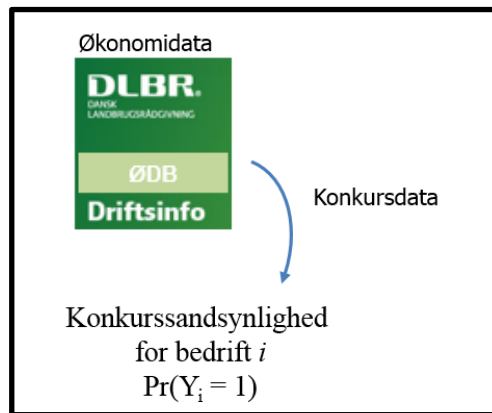
Udfordring 4

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

Black box



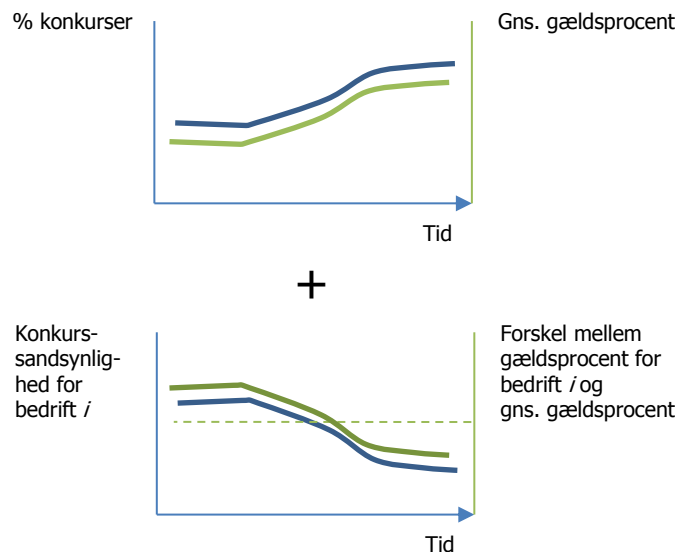
Identificering af **højrisiko** bedrifter
(early warning)

Individuel rådgivning på
baggrund af bedriftsspecifikke
økonomital

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)

Udfordring 6



- En enkelt bedrifts konkurssandsynlighed er en blandet funktion af:
 - Konjunkturen, som påvirker alle bedrifter og som ændrer sig igennem tiden
 - Bedriftsspecifikke forhold relativt til omverdens tilstand.
- Populationens konkurssandsynlighed er en dynamisk størrelse.
- En bedrifts konkurssandsynlighed er kontekstuel, dvs. relativt til populationens ændrede konkurssandsynlighed.



TEKNOLOGISK
INSTITUT

Version 1

Datamodel/strategi version 1 (2016)

Data



Økonomidata
for $m = 70$
variabler
($n \times m$)



flet

Konkursdata
(0/1)

($n \times 1$)

Datahåndtering

Feature ekstraktion for hver
af de m variabler

- lag 1, lag 2, ..., lag 6
- Ændring (1^o difference, d)
- Acceleration (dd)
- Løbende variance (SD), 3-års tidsvindue

Case-Control up-/down-
sampling indenfor år

- Generering af $N=10$ dataset for hver af de 4 definerede driftsgrene D
Søer, Slattesvin
Planter, Kvæg, Øvrig
- ved at:
- Bootstrap-sample case-observationer vægtet efter prop. af cases indenfor år

For hver af $D \times N \times 2 = D \times 20$

Missing Data Imputering

- Unsupervised random forest ($2 \times n_{tree} = 200$)

Variable pre-selection på
baggrund af eksterne kriterier

- $NA\% < \text{cutoff}$ (50%)
- Lineære kombinationer
- Gns. korrelation $< \text{cutoff}$ (75%)

For hver af $D = 4 \times N = 10$

- Tilfældigt sample control-observationer vægtet efter cases indenfor år og omvendt vægtet efter prop. af missing data (MAR)

Model

Split data i test-
og trainset

- testset: 5% af case-, og 5% af control-observationer

Random forest
klassifikations-
model

- Case-weights omvendt propor. med prævalensen
- Sampling stratificeret efter år

Prædiktio og modelperformance

For hver af $D \times N \times 2$

OOB – validering

Ekstern testset
validering

ROC-curve til
vurdering af
sensitivitet og
specificitet

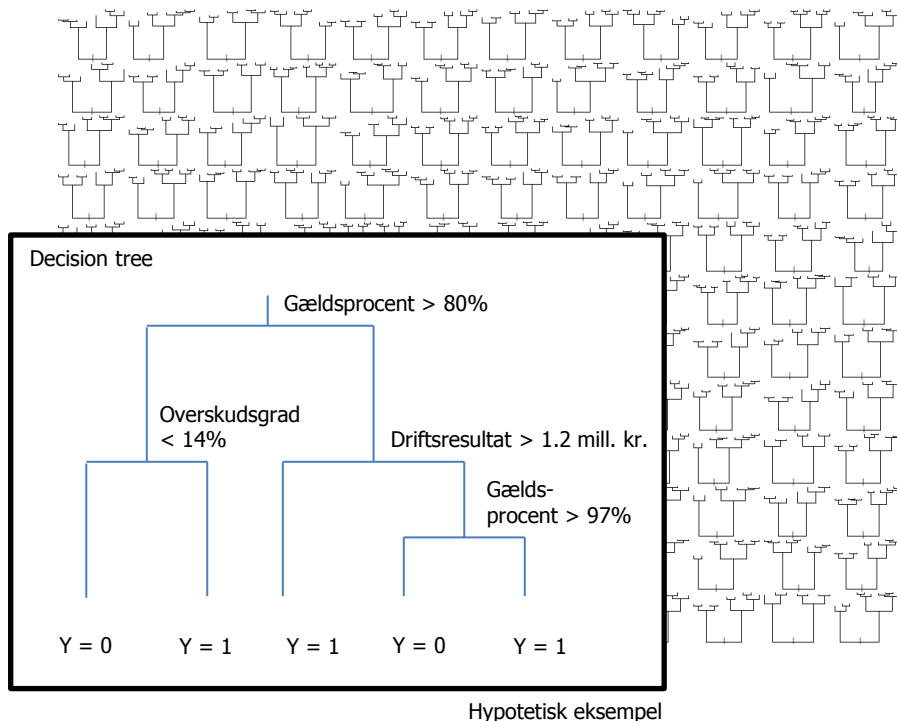
Per D over alle $N \times 2$

Samlet vurdering
for hver driftsgren

Konkurssandsynlighed
for bedrift i
 $\Pr(Y_i = 1)$

Hvad er en random forest?

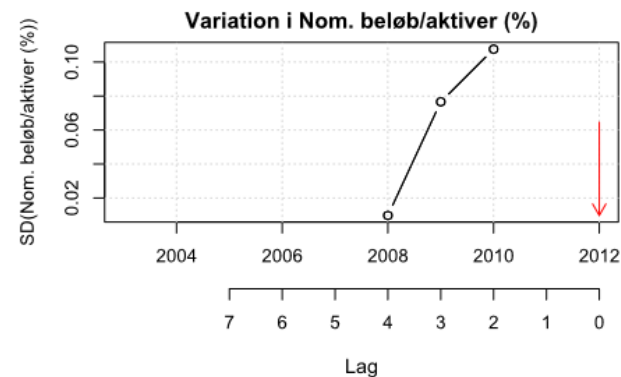
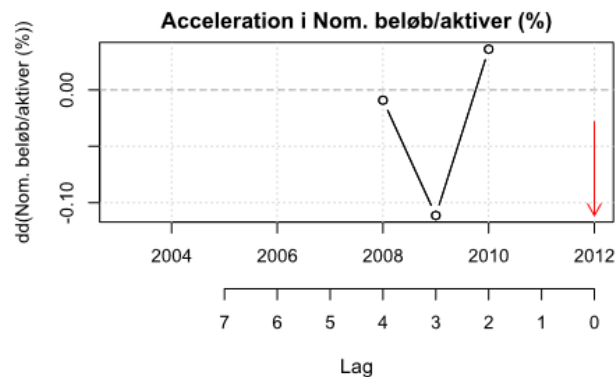
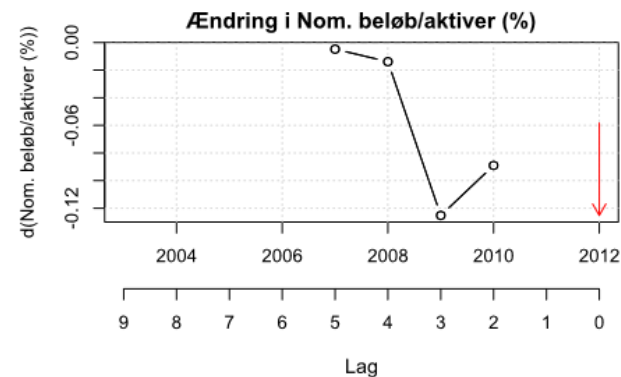
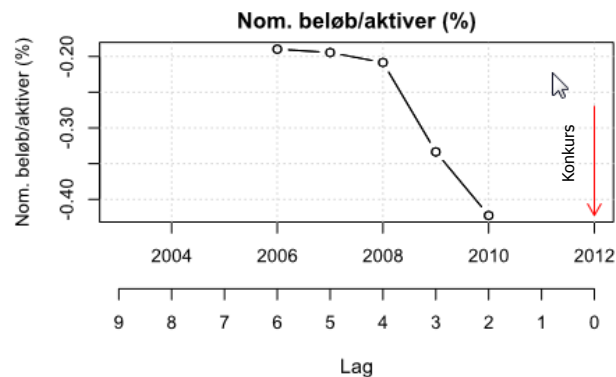
- Breiman 2001
- Et af de mest populære og succesfulde **machine learning** teknikker
- Bygger et stort antal (derfor 'forest') nær-uafhængige (derfor 'random') decision-tree modeller.
- Hver decision-tree prædikerer outputtet på baggrund af split-regler i inputparametrene.
- Den endelige prædiktion 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).



Rådgivningsintensitet

Eksempel på

Feature ekstraktion



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

 m inputvariable = 70

 k = features pr. m = 19

 $m \times k$ = 1330 variable 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/ netto rentebærende gæld - Div. indtjeningsmål/ passiver - Renteomk./omsætning - Egenkapitalforrentning - Rentedækningsgrad	Produktivitet - Udbytte pr. ha. - Mælkeydelse pr. ko - Grise pr. årsko	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		

Resultater version 1 - 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

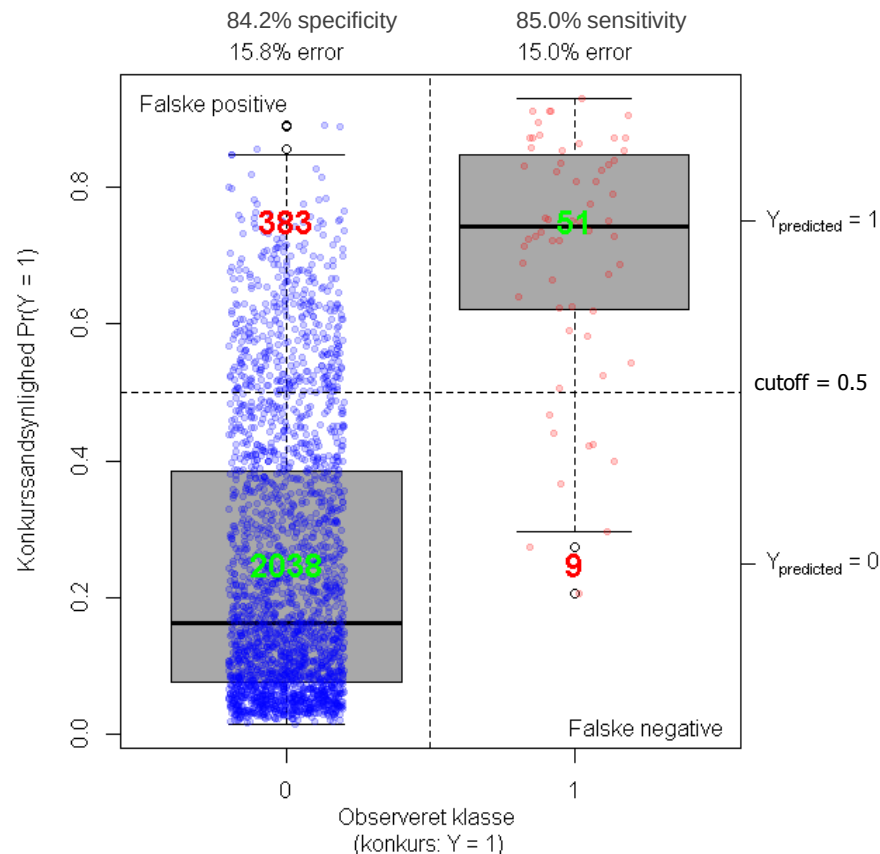
¹:cutoff-sandsynlighed= 0.5



Kvæg, model 1a

Sensitivitet og specificitet [version 1](#)

- 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).



Sensitivitet vs.

Positive Predicted Value (PPV) [version 1](#)

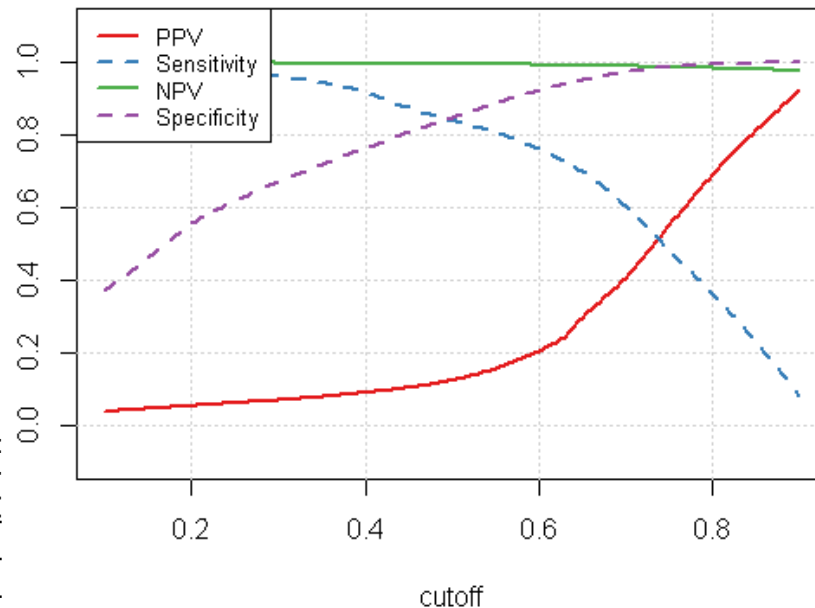
- "Hvad er sandsynligheden for konkurs for en bedrift, der af modellen forudsiges til at gå konkurs?"
- Dette afhænger af prævalensen!
- PPV er en prævalens-korrigeret analog til sensitivitet; NPV til Specificitet.

	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

Høj-risiko bedrifter
n = 21

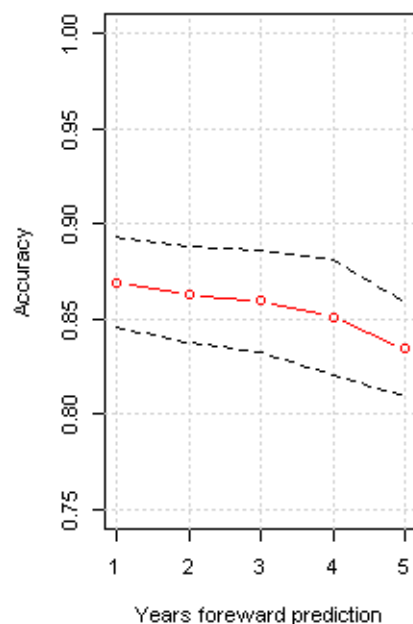
Kvæg, model 1a

Prævalens = 2.7%

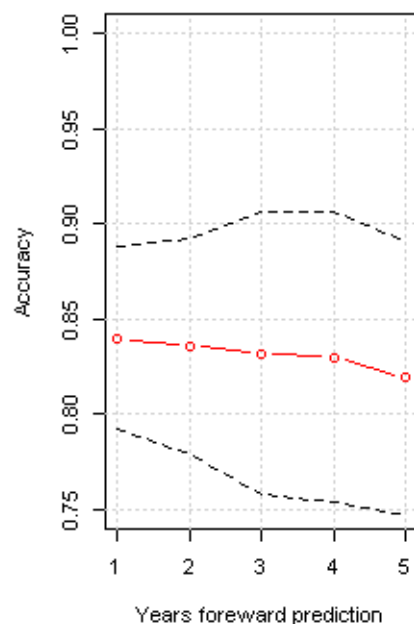


Resultater [version 1](#) - prædiktionsnøjagtighed som funktion af forecast-tidsvinduet

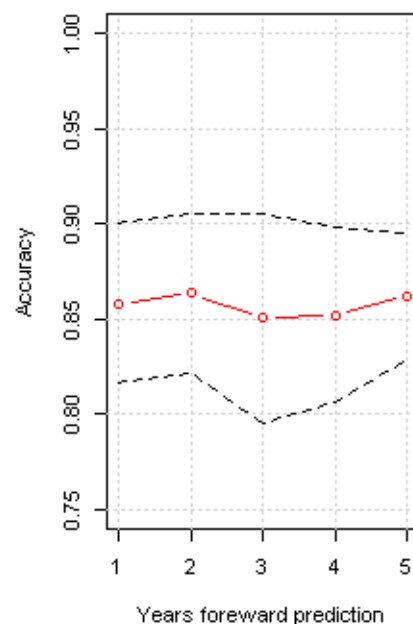
Kvægbedrifter



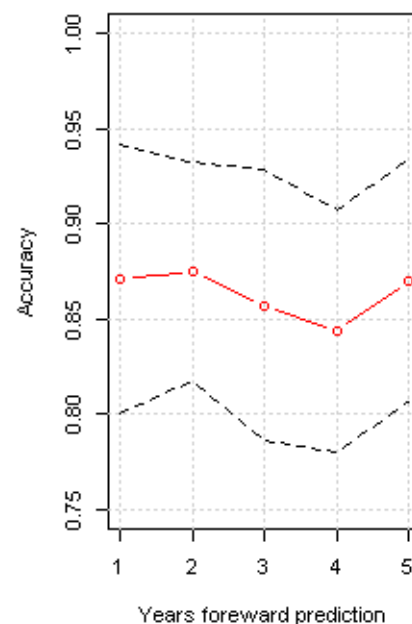
So-besætninger



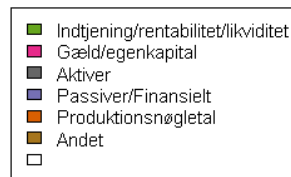
Slagtesvinsproduktion



Planteavl

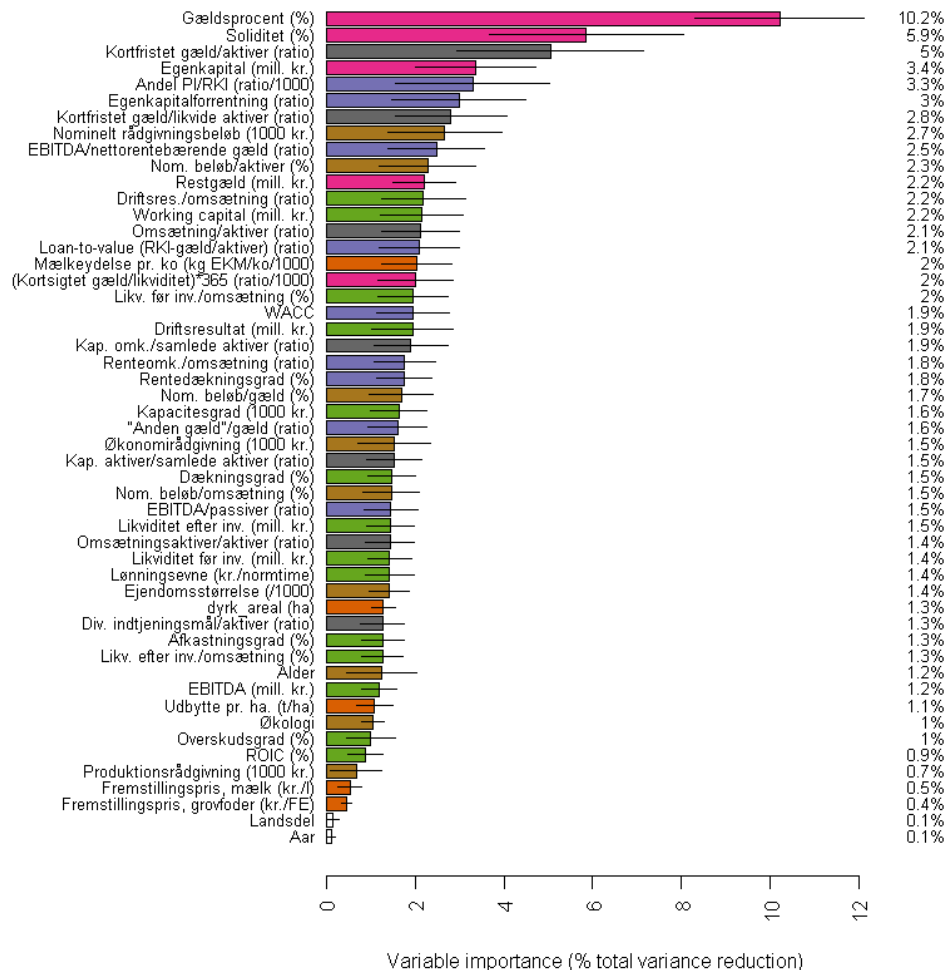


Grouped variable importance



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

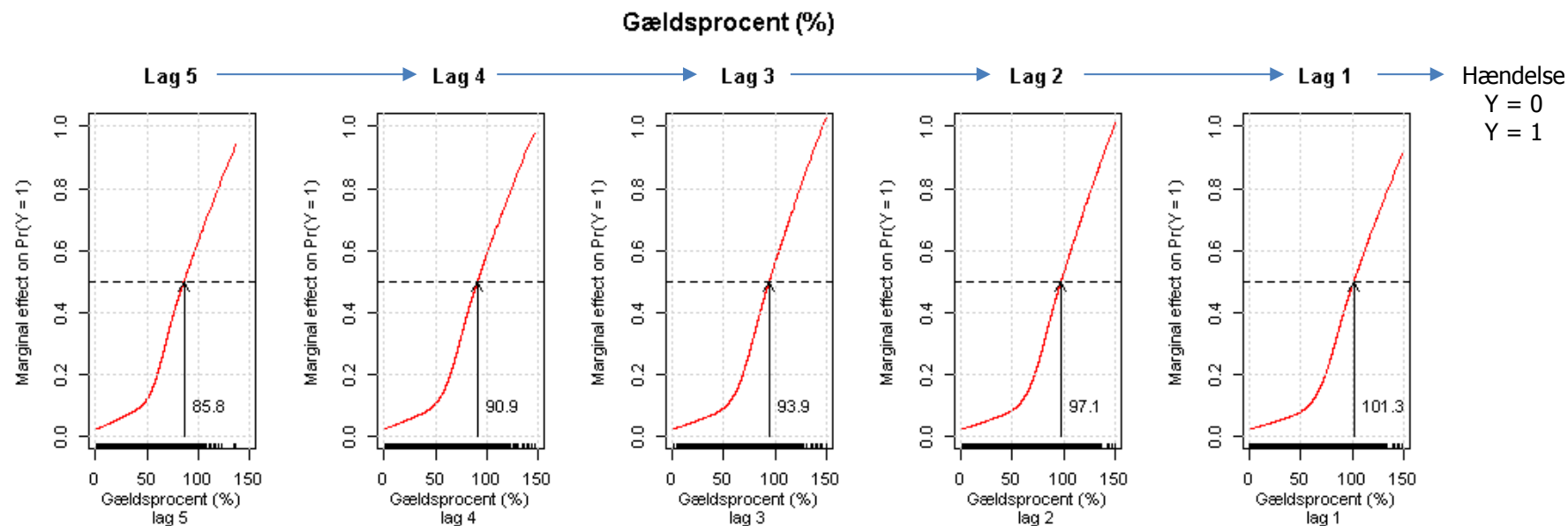


TEKNOLOGISK
INSTITUT

Resultater version 1

- variable importance

Resultater version 1 - marginale effekter af de mest betydende variabler



Konklusioner fra analyse version 1

- 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 højrisko-bedrifter (f.eks. PPV $\sim$ 0.75 og sensitivitet $\sim$ 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.



TEKNOLOGISK
INSTITUT

Version 2

Nye tiltag i version 2

- Flere data (2016 og 2017)
- Inkludering af nye variabler, f.eks. morarente, kreditinstitut, m.m.
- Random Forest model-parameter tuning
- Kontekstuel analyse, ved at inkludere års-gennemsnit og for hver bedrift den standardiserede afvigelse fra års-gennemsnittet for hvert feature og år.
- Undersøge om prædiktionen kan forbedres i en **samlet model på tværs af driftsgrene** med driftsgren som forklarende variabel.
- **2-steps model** for at undersøge om antal falske positive kan nedbringes og PPV forøges
 - Step 1: den hidtidige model
 - Step 2: ny model til prædiktion af konkurssandsynlighed for kun de bedrifter der er prædikeret til at gå konkurs under step 1.

$$\Pr(Y_i = 1 \mid \Pr(\hat{Y}_i = 1)_{step\ 1})$$

for at forbedre **prædiktions-nøjagtigheden**

- Gruppering af variabler
- Ekstrahere for hver konkurs-prædikeret bedrift de mest betydende konkurs-drivers.
- Grafisk og tabularisk visning af betydende variabler (konkurs-drivers, variable importance)
 - på populationsniveau
 - for hver konkurs-prædikeret bedrift

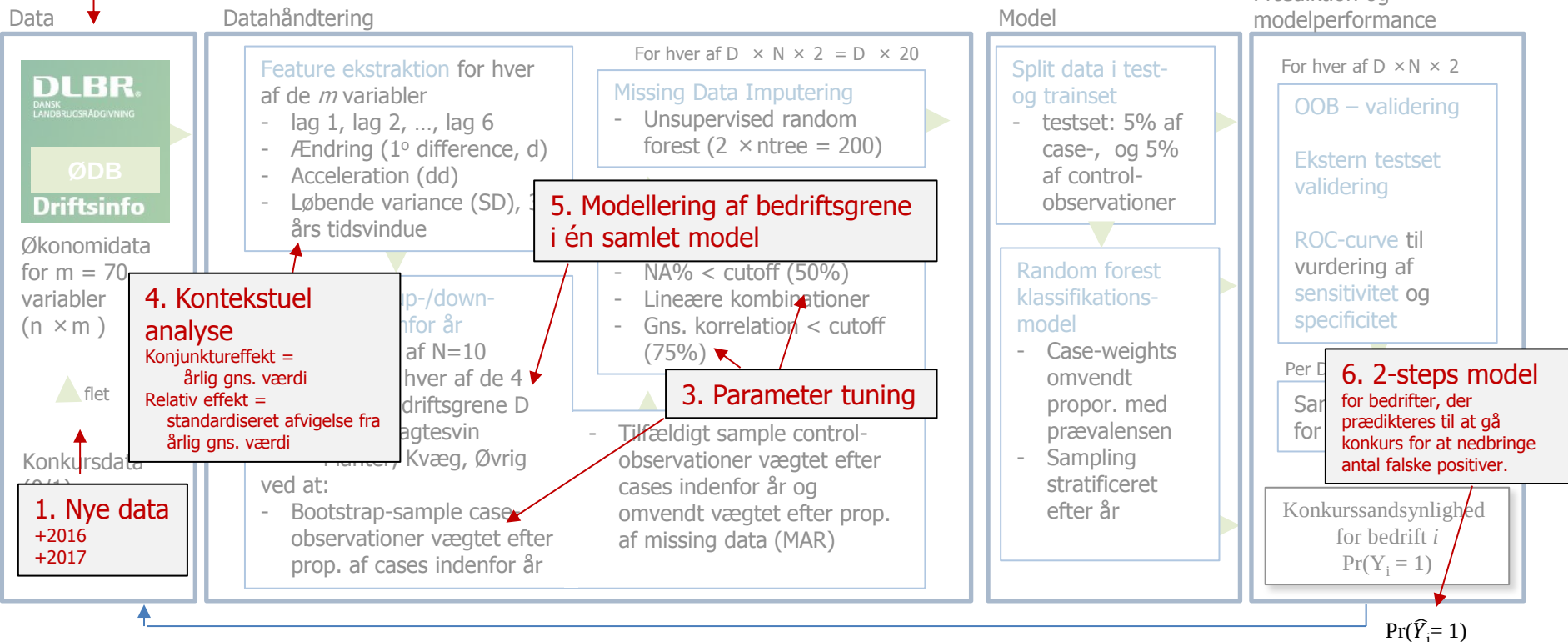
for at lette **tolkningen** af konkurs-drivers

2. Nye variabler

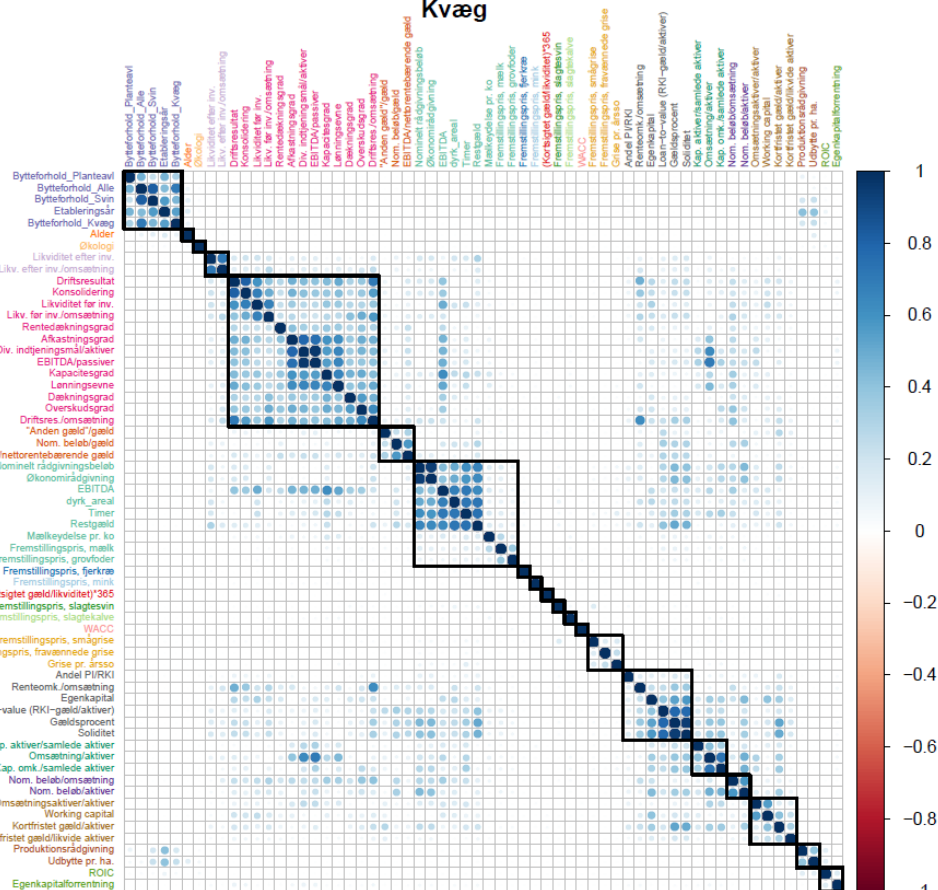
Kreditinstitut
Morarente
m.m.

Datamodel/strategi version 2 (2017)

Formål: - at forbedre prædiktionen



Kvæg



Correlogram af |input-variable|

Gruppering af input-variable

Gruppering af mulige forklarende variable i ranket rækkefølge – Mælkeproducenter (75 pct.)

Rentabilitet, likviditet, indtjening mv.	Finansielt:	Soliditet mv.	Produktivitet:	Rådgivning og ejendomsstørrelse:
Driftsresultat/omsætning	Egenkapitalforrentning	Gældsprocent	Udbytte pr. ha	Rådgivning (nominelt beløb)
Omsætning/aktiver	ROIC	Soliditet	Produktionsrådgivning	Økonomirådgivning
Likv. før inv./omsætning	Finansielt beredskab:	Kortfristet gæld/aktiver	Egenkapital	Dyrket areal
Driftsresultat (nominelt beløb)	(Kortsigtet gæld/likviditet)*365	Kortfristet gæld/likvide aktiver	Andel PIRKI	Timer (ny)
Renteomk./omsætning		EBITDA/nettorentebærende gæld	Kortfristet gæld/likvide aktiver	
Rentedækningsgrad		Aktiver (nominelt beløb)	EBITDA/nettorentebærende gæld	
Kapacitetsgrad	Finansieringsomkostninger:	Restgæld (nominelt beløb)	Working capital (kortf. Aktiver/kortf. gæld)	Effektivitet:
Dækningsgrad	WACC	Loan-to-value (RKI-gæld/aktiver)	Kap. omk./samlende aktiver	Mælkeydelse pr. ko
Omsætning (nominelt beløb)		Gæld (nominelt beløb)	Gæld (nominelt beløb)	Fremstillingspris, grovfoder
EBITDA/passiver		*Anden gæld/gæld	Kap. aktiver/samlende aktiver	Fremstillingspris, slagtekalve (ny)
Likviditet efter inv.		Omsætningsaktiver/aktiver		
Lønningsevne				
Div. Indtjeningsmål/aktiver				
Afkastningsgrad				
Likv. efter inv./omsætning				
EBITDA (nominelt beløb)				
Overskudsgrad				
Konsolidering (ny)				

cluster analyse + faglig vurdering

Grupperede input-variable

Gruppering af mulige forklarende variable i ranket rækkefølge – Mælkeproducenter (75 pct.)

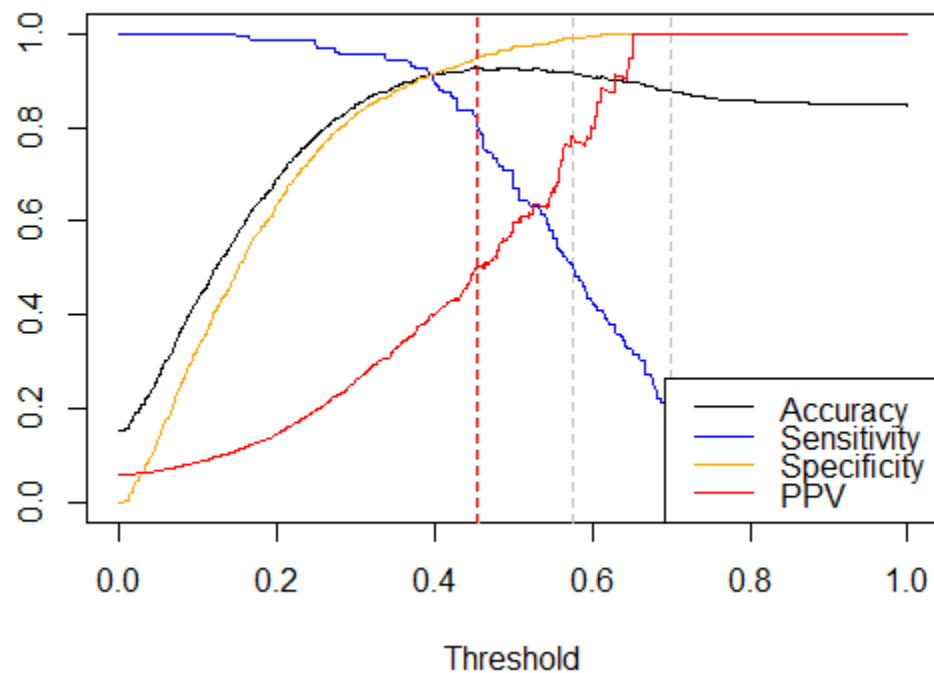
Rentabilitet, likviditet, indtjening mv. Driftsresultat/omsætning Omsætning/aktiver Likv. før inv./omsætning Driftsresultat (nominelt beløb) Renteomk./omsætning Rentedækningsgrad Kapacitetsgrad Dækningsgrad Omsætning (nominelt beløb) EBITDA/passiver Likviditet efter inv. Likviditet før inv. Lønningsevne Div. Indtjeningsmål/aktiver Afkastningsgrad Likv. efter inv./omsætning EBITDA (nominelt beløb) Overskudsgrad Konsoidering (ny)	Finansielt: Egenkapitalforrentning ROIC	Soliditet mv. Gældsprocent Soliditet Kortfristet gæld/aktiver Egenkapital Andel PI/RKI Kortfristet gæld/likvide aktiver EBITDA/nettorentebærende gæld Aktiver (nominelt beløb) Restgæld (nominelt beløb) Working capital (kortf. Aktiver/kortf. gæld) Loan-to-value (RKI-gæld/aktiver) Kap. omk./samlede aktiver Gæld (nominelt beløb) *Anden gæld*/gæld Kap. aktiver/samlede aktiver Omsætningsaktiver/aktiver	Produktivitet: Udbytte pr. ha. Produktionsrådgivning	Rådgivning og ejendomsstørrelse: Rådgivning (nominelt beløb) Økonomirådgivning Dyrket areal Timer (ny)
	Finansielt beredskab: (Kortsigtet gæld/likviditet)*365		Effektivitet: Mælkeydelse pr. ko Fremstillingspris pr. kg. mælk Fremstillingspris, grovfoder Fremstillingspris, Slagtekalve (ny)	Demografi Alder Etableringsår
	Finansieringsomkostninger: WACC			

Markedsforhold:
Bytteforhold, alle (Ny)
Bytteforhold, mælk (Ny)

Prædiktionsperformance (prædiktion 1 år frem)

	inkluderede trin		
		Se. ¹	Sp.
0. Version 1 (2015 og før)	0	0.93	0.92
1. Nye data (2016 + 2017)	0,1	0.83	0.83
2. Nye variabler	0,1,2	0.86	0.86
3. Kontekstuel analyse	0,1,2,3	0.87	0.87
4. Trans-driftsgren model	0,1,2,3,4	0.90	0.90
5. 2-steps model	0,1,2,3,4,5	0.90	0.90
6. Rullende model	0,1,2,3,4,6	0.90	0.90

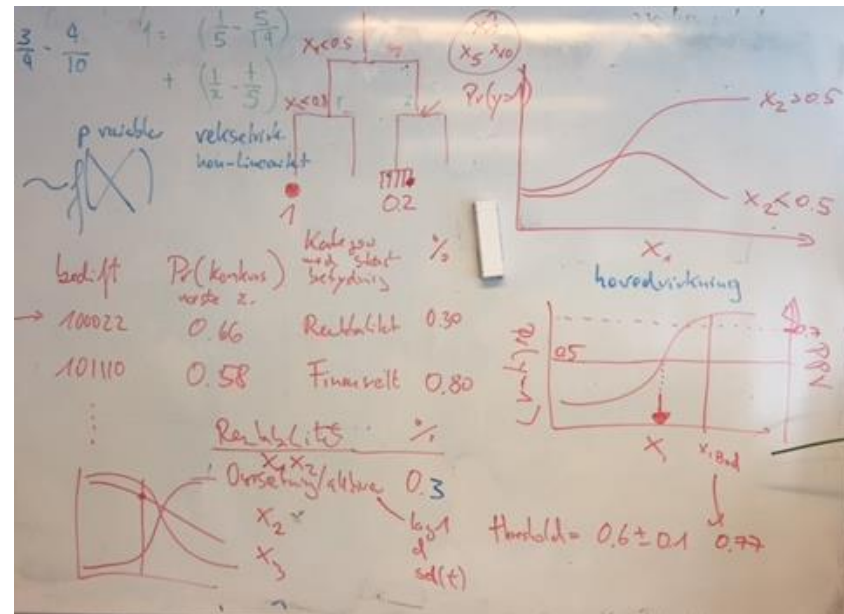
¹: OOB



Konkurs-drivers

De af SEGES udarbejdede, bedriftsgren-specifikke **variabelkategorier** vil blive **anvendt i modellens fortolkning** for den enkelte bedrift.

- Variabelkategorier f.eks. for mælkeproducenter er: Rentabilitet, likviditet, indtjening mv., Finansielt, Finansielt beredskab, Finansieringsomkostninger, Soliditet mv., Produktivitet, Effektivitet, Rådgivning og ejendomsstørrelse, Demografi, Markedsforhold, Økologi, Andet
- Ud fra modellens skov af klassifikationstræer bliver der for hver bedrift beregnet:
 - 1) en prædikeret konkurssandsynlighed, $\Pr(Y=1)$
 - 2) en PPV
 - 3) kategoriernes betydningsgrad (KBG i %), dvs. bidrag til $\Pr(Y=1)$, i rangeret orden
 - 4) kategoriernes gns. forskel til $\Pr(Y=1) = 0.5$ tærsklen (vi skal have fundet ud af hvordan)
 - 5) variabelernes betydningsgrad (i %) indenfor hver kategori, dvs. bidrag til KBG, i rangeret orden
 - 6) variabelernes gns. forskel til $\Pr(Y=1) = 0.5$ tærsklen (vi skal have fundet ud af hvordan)
 - 7) grafer over variabelernes effekt på $\Pr(Y=1)$ (kun hovedvirkning)





TEKNOLOGISK
INSTITUT

