



Notat v. 2018: Konkursmodeller for de danske landbrugserhverv - modelperformance

24-01-2019

Philipp Trénel, ph.d., seniorkonsulent, specialist
phtr@teknologisk.dk, +45 72 20 33 92

Formål

- At udvikle version 2.1 af en model til prædiktionsaf **misligholdelsessandsynlighed (probability of default, PD)** på baggrund af data fra SEGES økonomi-database (ØDB) og konkursdata indhentet af SEGES fra statstiderne.dk. Modellen er videreudviklet ud fra modellen beskrevet i notat af samme navn fra 2016 (bilag 1) og 2017 (bilag 2), men har udvidet konkurskategorien til en misligholdelseskategori, der inkluderer bedrifter, der er gået konkurs eller misligholdt på anden vis, udtrykt ved at bedriften er pålagt morarenter eller er i gældssanering. Den udviklede model sigter på at indgå i SEGES early-warning system rettet mod rettidig rådgivning af landmænd med henblik på forebyggelse af konkurser.

Overordnede konklusioner

- Modellens prædiktionsperformance kunne betydeligt forbedres (Sensitivitet = 0.94, PPV = 0.99) ved at i) udvide targetkategorien fra konkurs til misligholdt, ii) foretage random forest-baseret variable-hunting og iii) udnytte en to-steps model.

Metoder

- ØDB data og konkurslabellen fra statstiderne.dk er sammenflettet og leveret af SEGES i form af et dataset (table 1 - data.csv).
- Inputvariablerne er klassificeret i overordnede kategorier jf. bilag 2.
- Der er foretaget følgende datahåndtering:
 - Banker med <10 bedrifter indgår i bankkategorien "Ukendt".
 - Banker med <50 bedrifter indgår i samlekategori "små banker".
 - Landsdelene "Østsjælland", "København by" og "Københavns omegn" slås sammen og betegnes "Østsjælland".
 - Driftsgren bestemmes for hver bedrift jf. proceduren beskrevet i bilag 2.

Dataafgrænsning:

- Ikke-heltidsbedrifter fjernes fra datasættet.
- Bedrifter med færre end 3 års ØDB-data eller med mere end halvdelen af årene mellem første og sidste registrering i ØDB med manglende data fjernes fra datasættet.

Kontekstuelle variabler:

- Variabler genereres jf. den kontekstuelle metode beskrevet i bilag 2, dvs. der beregnes for alle numeriske variabler et populationsgennemsnit for hvert år og for hver bedrift beregnes forskellen til populationsgennemsnittet.

Imputering vha. variabel-interpolering:

- For alle bedrifter og variabler foretages der en Stineman-interpolering for numeriske variabler og last-observation-carried-forward for alle klassevariabler for alle år med manglende data, der ligger mellem bedriftens første og sidste registrering i ØDB eller konkurshændelse.

Feature beregning:

- Der beregnes features for hver inputvariabel i datasættet som beskrevet i bilag 1, dvs. lag1, lag2, ..., SD1, SD2, ..., etc.
- Der beregnes features for forecast-årene mellem 2014 og 2017 for alle relevante bedrifter.
- Derudover beregnes der variabler, der angiver:
 - antal år med manglende registreringer i ØDB mellem aktuel år og sidste ØDB-registrering.
 - driftsgrenskift ja/nej
 - økologiskift ja/nej
 - bankskift ja/nej
 - kreditforeningsskift ja/nej

Det endelige datasæt:

- Det endelige datasæt består af 20.651 observationer (bedrift-år-kombinationer) fra i alt 6.074 bedrifter i de fem driftsgren-kategorier "Kvæg" (2.677 bedrifter), "Planteavl" (1.349), "Sobesætninger" (1.027), "Slagtesvinsbesætninger" (959) og "Øvrige" (510).
- Det endelige datasæt indeholder i alt 4.686 variabler og features.
- Der estimeres en samlet model på tværs af alle fem driftsgrene, med driftsgren som en potentiel forklarende variabel i modellen.

Variabel hunting:

- Der foretages en indledende variabelselektion vha. random forest metoden variabel-importance-based variabel hunting, se Ishwaran & Kogalur (2016), ved at køre 100 repetitioner af Random Forest (RF) modeller (hver med 1000 træer, 10-fold krydsvalidering og ellers defaultindstillinger). Variabler med en

selektionshyppighed på 0 blev ekskluderet fra datasættet, hvilket resulterede i 2.450 variabler og features.

To-steps model:

- Modellen består af en to-steps RF model, se bilag 1 og 2: Step 1 inkluderer alle observationer (N = 20.651) og sigter på at opnå en høj specificitet. Step 2 inkluderer kun de observationer, hvis prædikterede PD_{step1} opfylder $0.1 < PD_{step1} < 0.8$ (N = 10.769). Den endelige prædiktions over begge steps er:

$$PD_{step1+2} = \begin{cases} PD_{step1} & \text{hvis } 0.1 < PD_{step1} < 0.8 \\ PD_{step2} & \text{ellers} \end{cases}$$

Resultater

- Resultater fremgår af tabel 1 og figurerne 1 til 3.
- Der opnås en høj prædiktionsnøjagtighed, sensitivitet og predictive value positive (PPV) med to-steps-modellen.

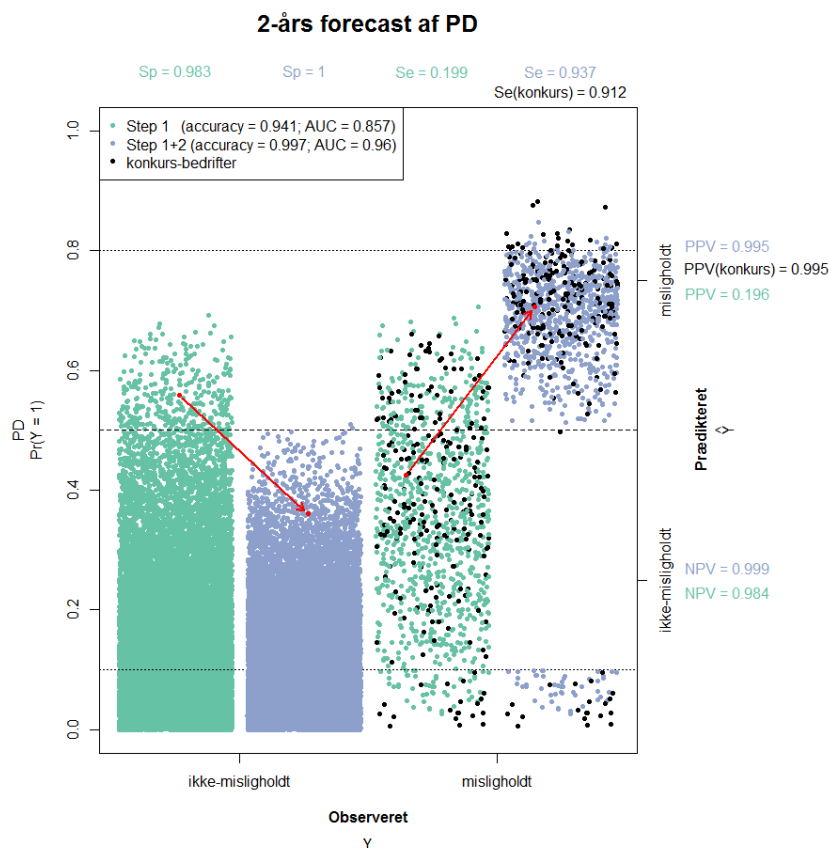


Fig. 1. Out-of-bag modelperformance af 2-års-forecast-modellen for prædiktions af Probability of Default (PD) eller misligholdelses-sandsynligheden trænet over forecast-årene 2014 til 2017. Modellen består af to steps: Step 1 (grøn) inkluderer alle observationer (N = 20.651) og sigter på at opnå en høj specificitet. Step 2 inkluderer kun de observationer, hvis prædikterede PD_{step1} opfylder $0.1 < PD_{step1} < 0.8$ (N = 10.769). Den endelige prædiktions over begge steps er $PD_{step1+2} = \begin{cases} PD_{step1} & \text{hvis } 0.1 < PD_{step1} < 0.8 \\ PD_{step2} & \text{ellers} \end{cases}$. De rødt markerede observationer

viser to eksempler, hvis prædikterede klasse skifter fra step 1 til step 2. Sort indikerer de misligholdte bedrifter, der er gået konkurs. I beregningen af PPV og NPV er der anvendt en forventet prævalens på 2%. AUC står for "area under the curve" og er et cutoff-fri mål for modelperformance.



Tabel 1: Out-of-bag modelperformance af 2-års-forecast-modellen (2-step-modellen) for prædiktionen af Probability of Default (PD) på tværs af driftsgrene (alle) og for hver af driftsgrenene.

Driftsgren	Statistic	Estimate	s.e.	Driftsgren	Statistic	Estimate	s.e.	Driftsgren	Statistic	Estimate	s.e.
Alle	N	20.651		Slagtesvins- produktion	N	2.768		So-besæt- ninger	N	3.232	
	heraf misligholdte (prævalens)	1.103 (5,3%)			heraf misligholdte (prævalens)	147 (5,3%)			heraf misligholdte (prævalens)	208 (6,4%)	
	heraf konkurser (prævalens)	238 (1,2%)			heraf konkurser (prævalens)	146 (1,6%)			heraf konkurser (prævalens)	49 (1,5%)	
	Accuracy	0,997	<0,001		Accuracy	0,996	<0,001		Accuracy	0,997	<0,001
	AUC	0,961	0,001		AUC	0,937	0,004		AUC	0,957	0,002
	Se	0,939	0,002		Se	0,921	0,007		Se	0,949	0,002
	Sp	0,999	<0,001		Sp	>0,999	<0,001		Sp	>0,999	<0,001
	2%-prævalens PPV	0,992	0,004		2%-prævalens PPV	>0,999	<0,001		2%-prævalens PPV	>0,999	<0,001
	2%-prævalens NPV	0,999	<0,001		2%-prævalens NPV	0,998	<0,001		2%-prævalens NPV	0,999	<0,001
	Sensitivitet _{konkurser}	0,913	0,004		Sensitivitet _{konkurser}	0,913	0,013		Sensitivitet _{konkurser}	0,918	<0,001
Kvæg- bedrifter	PPV _{konkurser}	0,991	0,004	Planteavl	PPV _{konkurser}	>0,999	<0,001	Øvrige	PPV _{konkurser}	>0,999	<0,001
	N	8.996			N	4.043			N	1.612	
	heraf misligholdte (prævalens)	618 (6,9%)			heraf misligholdte (prævalens)	109 (2,7%)			heraf misligholdte (prævalens)	21 (1,3%)	
	heraf konkurser (prævalens)	146 (1,6%)			heraf konkurser (prævalens)	15 (0,4%)			heraf konkurser (prævalens)	4 (0,2%)	
	Accuracy	0,997	<0,001		Accuracy	0,997	<0,001		Accuracy	0,995	<0,001
	AUC	0,974	0,002		AUC	0,953	0,008		AUC	0,819	0,014
	Se	0,960	0,003		Se	0,894	0,023		Se	0,581	0,020
	Sp	0,999	<0,001		Sp	>0,999	<0,001		Sp	>0,999	<0,001
	2%-prævalens PPV	0,981	0,009		2%-prævalens PPV	>0,999	<0,001		2%-prævalens PPV	>0,999	<0,001
	2%-prævalens NPV	0,999	<0,001		2%-prævalens NPV	0,998	<0,001		2%-prævalens NPV	0,992	<0,001
	Sensitivitet _{konkurser}	0,908	0,006		Sensitivitet _{konkurser}	0,987	0,038		Sensitivitet _{konkurser}	0,750	<0,001
	PPV _{konkurser}	0,980	0,010		PPV _{konkurser}	>0,999	<0,001		PPV _{konkurser}	>0,999	<0,001

AUC: area under the curve; Se: sensitivitet; Sp: specificitet; PPV: positive predictive value; NPV: negative predictive value; cutoff PD = 0,50; s.e. er beregnet ud fra 10 genberegninger af random forest modellen.

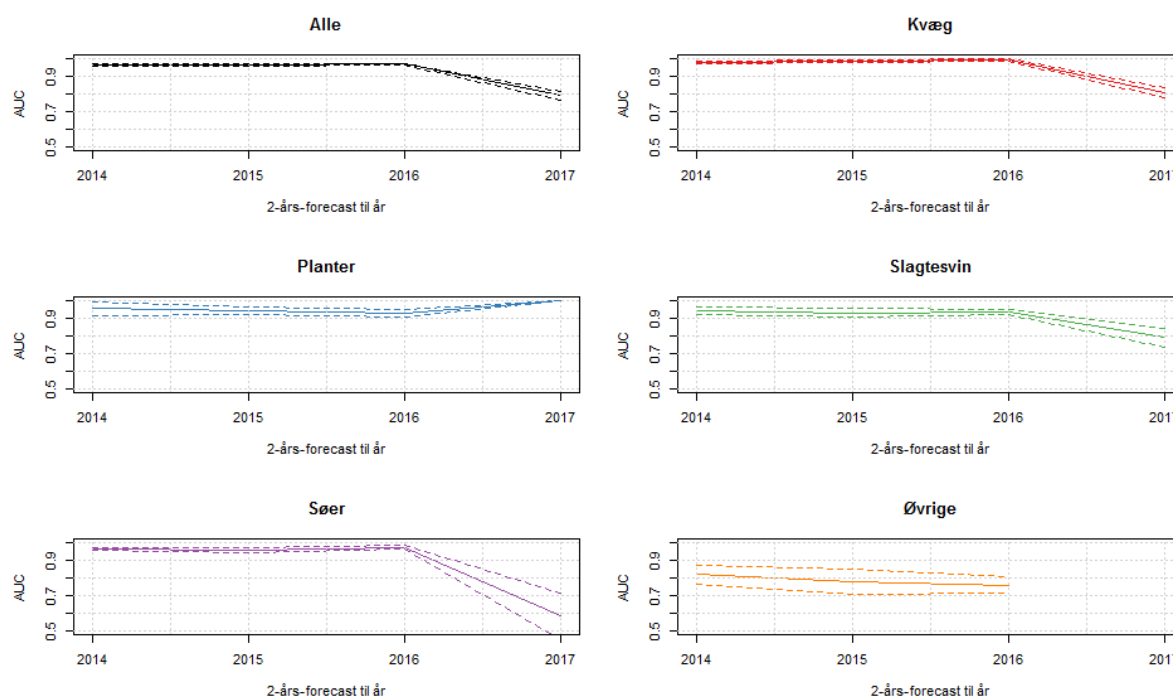


Fig. 2: To-års forecast model-performance udtrykt som AUC over årene 2014 til 2017 på tværs af driftsgrenene og for hver af driftsgrenene. 95% konfidensintervallerne er beregnet ud fra 10 genberegninger af random forest modellen.

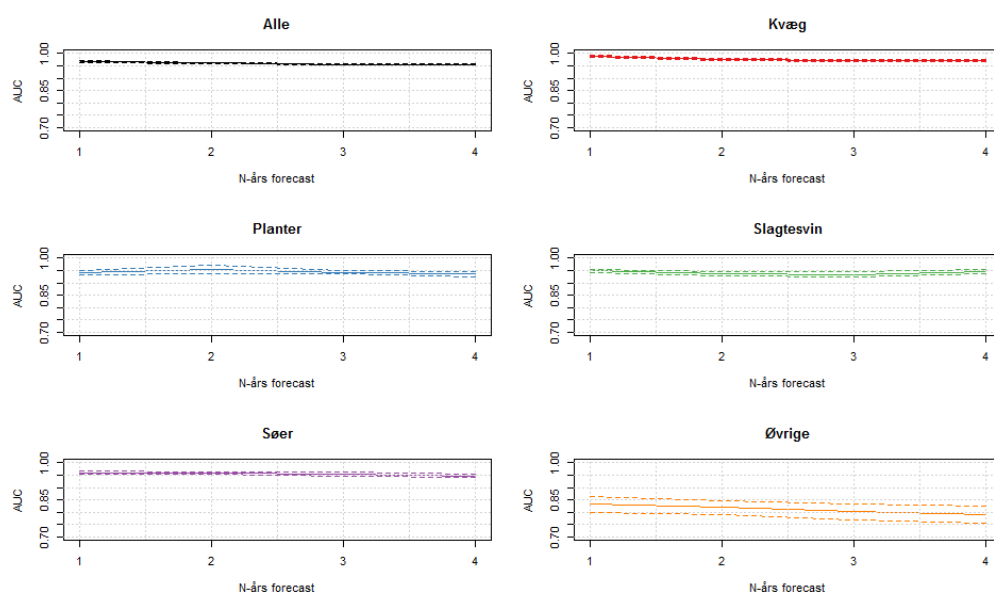


Fig. 2: Model-performance udtrykt som AUC for prædiktionen N år frem på tværs af driftsgrenene og for hver af driftsgrenene. 95% konfidensintervallerne er beregnet ud fra 10 genberegninger af random forest modellen.

Referencer

Ishwaran, H. & Kogalur, U.B. 2016 randomForestSRC: Random Forests for Survival, Regression and Classification (RF-SRC). <https://cran.r-project.org/web/packages/randomForestSRC/index.html>