

NHH



Norges Handelshøyskole

Bergen, Høst 2017

Konkursprediksjon med termindata

En empirisk studie av prediksjonsevnen til termindata fra Skatteetaten

Mats Sævig og Guri Husom Vonen

Veiledere: Jarle Møen og Jonas Andersson

Masterutredning i økonomi og administrasjon

Hovedprofil: Økonomisk styring

NORGES HANDELSHØYSKOLE

Dette selvstendige arbeidet er gjennomført som ledd i masterstudiet i økonomi- og administrasjon ved Norges Handelshøyskole og godkjent som sådan. Godkjenningen innebærer ikke at Høyskolen eller sensorer inntår for de metoder som er anvendt, resultater som er fremkommet eller konklusjoner som er trukket i arbeidet.

Forord

Denne masterutredningen er gjennomført som en del av masterstudiet i økonomi og administrasjon ved Norges Handelshøyskole høsten 2017.

Det var viktig for oss å finne et tema hvor kvantitative analyser stod sentralt. Konkursprediksjon ble ansett som et svært interessant tema hvor vi fikk anvendt vår kunnskap om blant annet regnskapsanalyse, skatt og statistiske modeller. Problemstillingen er utviklet i samarbeid med Norwegian Center for Taxation (NoCeT).

Vi ønsker å rette en stor takk til våre veiledere Jarle Møen og Jonas Andersson for god veiledning og gode diskusjoner. Vi ønsker også å takke Bo Vignes i Skatt Vest for bearbeiding av datasettet og for å ha besvart flere spørsmål i løpet av denne høsten.

Sammendrag

I denne oppgaven undersøker vi i hvilken grad termindata fra Skatteetaten kan brukes for å forbedre eksisterende regnskapsbaserte konkursprediksjonsmodeller. Konkursprediksjonsmodeller benyttes blant annet av banker, investorer, kredittratingbyråer og forvaltere for å kunne si noe om individuelle konkurssannsynligheter for enkelte selskap. Slike modeller benyttes også av Norges Bank og Finanstilsynet for å kunne si noe om bankenes kredittrisiko overfor foretakssektoren. Konkursprediksjonsmodeller bidrar til å bedre finansmarkedenes evne til å formidle kapital til lavest mulig kostnad.

Eksisterende modeller bruker i all hovedsak offentlig tilgjengelig regnskapsinformasjon for å predikere konkurser. Termindata fra Skatteetaten inneholder informasjon om blant annet selskaps merverdiavgift, forskuddstrekk, arbeidsgiveravgift og restskatt. I tillegg produserer Skatteetaten oppdaterte termindata flere ganger i året.

Vi estimerer først en referansem modell som skal tilsvare eksisterende regnskapsbaserte konkursprediksjonsmodeller, og er i stor grad inspirert av Norges Banks SEBRA-modell. Videre estimeres to modeller som inkluderer termindata fra Skatteetaten på ulike måter. Alle modellene estimeres med en ikke-parametrisk logistisk metode. Studien tar i bruk et datasett som består av termindata og næringsoppgaver for norske aksjeselskap i perioden 2008 til 2017.

Opgaven konkluderer med at informasjonen i termindata signifikant forbedrer prediksjonsevnen til eksisterende, regnskapsbaserte konkursprediksjonsmodeller. Dette indikerer at offentliggjøring av termindata kan bedre finansmarkedenes evne til å formidle kapital til sin beste anvendelse.

Innholdsfortegnelse

1	Innledning	1
1.1	Motivasjon	1
1.2	Problemstilling	3
1.3	Oppgavens struktur	3
2	Bakgrunn	4
2.1	Avgrensning	4
2.2	Definisjon av konkurs	4
2.3	Overordnet om konkurser	5
2.4	Termindata	6
2.5	Endringer i lovverket	7
3	Tidligere arbeid	9
3.1	Beavers univariate modell	9
3.2	Altman's Z-score	10
3.3	Ohlson's logistiske regresjonsmodell	12
3.4	Norges Banks SEBRA-modell	13
3.5	Maskinl�ring	16
3.6	Sammenligning mellom tidligere arbeider	16
4	Metode	19
4.1	Estimering	19
4.1.1	Parametrisk logistisk regresjon	19
4.1.2	Ikke-parametrisk logistisk regresjon	21
4.2	Validering	26

4.2.1	Out-of-time	26
4.2.2	Receiver Operating Characteristics (ROC)	26
5	Datasett	30
5.1	Beskrivelse av datasettet	30
5.1.1	Termindata	30
5.1.2	Næringsoppgave	30
5.1.3	Øvrige data	31
5.2	Avgrensning	31
5.3	Beskrivende statistikk	32
5.4	Avhengig variabel	34
5.5	Forklaringsvariabler	34
5.5.1	Forklaringsvariabler fra regnskapet	34
5.5.2	Forklaringsvariabler fra termindata	37
5.6	Datakvalitet	39
5.7	Paneldata	39
6	Empirisk analyse	41
6.1	Prediksjonsrammeverk	41
6.2	Modell 1 - Referansemodell	42
6.3	Modell 2 - Termindata fra samme år som årsregnskapet	44
6.4	Modell 3 - Termindata fra tidspunktet de er tilgjengelig	46
6.5	Sammenligning av de ulike modellene	47
6.6	Konkurshorisont	49
6.7	Validering og robusthet	51
6.8	Parametrisk modell	52
6.9	Regresjoner for alle modeller	54
7	Drøfting	55
7.1	Tolkning av resultater	55
7.2	Implikasjoner	57
7.3	Utfordringer og begrensninger	58

8	Konklusjon	60
	Referanser	61
	Appendiks	63
A1	Konkurs etter alder	64
A2	Konkurs per næring	64
A3	Modell 1 - Ikke-parametrisk	65
A4	Modell 2 - Ikke-parametrisk	69
A5	Modell 3 - Ikke paramterisk	73
A6	Modell 1 - Parametrisk	77
A7	Modell 2 - Parametrisk	78
A8	Modell 3 - Parametrisk	79

Figurliste

1.1	Bankenes utlånstap, gjeld i konkursrammede foretak og estimert konkurs- utsatt gjeld	2
2.1	Antall konkurser per kvartal	5
2.2	Skatteetatens frister	7
4.1	Lineær regresjon sammenlignet med logistisk regresjon	21
4.2	Den additive formen til GAM	21
4.3	En parametrisk logistisk modell sammenlignet med en ikke-parametrisk logistisk modell	23
4.4	Avveining mellom ”overfitting” og ”underfitting”	24
4.5	Eksempel på en ROC-kurve	28
5.1	Utestående mva-gjeld	34
6.1	Tidfesting i modell 1	44
6.2	Tidfesting i modell 2	46
6.3	Tidfesting i modell 3	47
6.4	ROC-kurver for alle modeller	48
6.5	ROC-kurver for modell 3 med ulike konkurshorisonter	50
6.6	ROC-kurver for modell 3 mot ulike valideringssett	51

Tabelliste

3.1	Variablene som inngår i Norges Banks SEBRA-modell	15
4.1	Type I og type II feil	27
5.1	Fordeling av antall konkurser	33
5.2	Regresjon som viser sammenhengen mellom konkurs og alder	33
5.3	Logistisk regresjon med terminvariabler	39
6.1	Variabler i modell 1	43
6.2	Variabler i modell 2	45
6.3	Sammenligning av modeller	47
6.4	Totale treffprosjenter	49
6.5	Sammenligning av AUC-verdier for ulike konkurshorisonter med modell 3	49
6.6	Sammenligning av AUC-verdier for modell 1 og modell 3 med ulike konkurshorisonter	50
6.7	Sammenligning av prediksjonsevnen til modell 3 i 2012 og 2013 for ulike konkurshorisonter	52
6.8	Sammenligning av modell 1 og modell 3 ved å benytte en parametriske logistisk metode	53
6.9	Oversikt over modellene	54

1. Innledning

1.1. Motivasjon

Konkurs kan ansees som en mekanisme som naturlig forekommer i et kapitalistisk system. Selskap som ikke presterer forsvinner, mens levedyktige selskap blir værende i markedet. Det kan dermed argumenteres for at konkurser er nødvendig for at markedet skal fungere optimalt. På en annen side er konkurser kostbart for samfunnet og berører flere parter. Eiere kan tape investeringer, kreditorer kan tape utestående gjeld, ansatte mister jobben og det offentlige kan gå glipp av skatter og avgifter. Kunder og leverandører kan også lide tap ved en konkurs. Å bedre kunne predikere hvilke selskap som går konkurs og hvilke som overlever vil dermed ha høy nytteverdi for flere aktører.

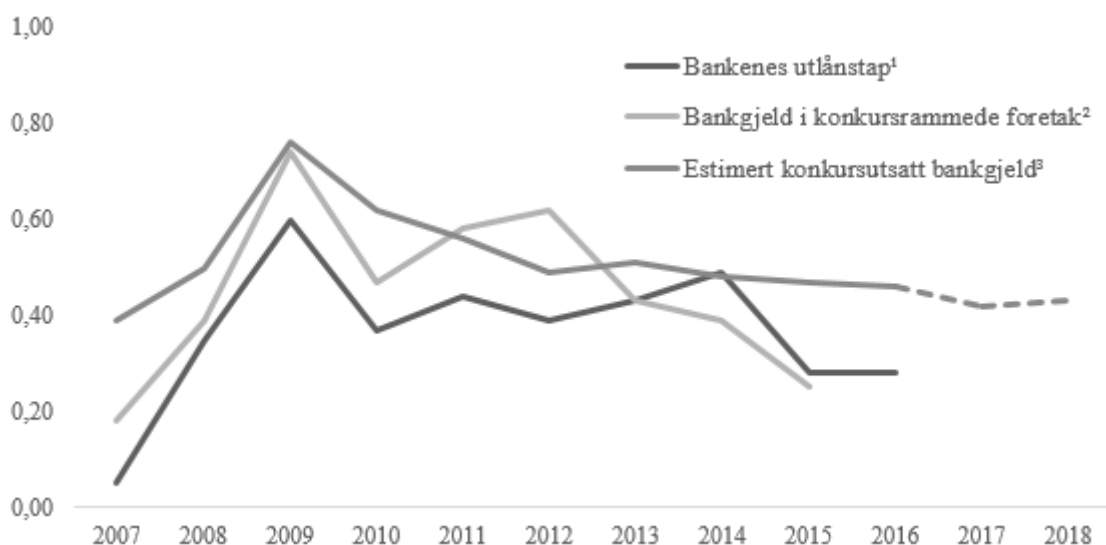
Å troverdig spå fremtiden til et selskap er vanskelig ettersom det alltid vil være faktorer som man umulig kan forutse. Litteraturen foreslår flere faktorer som påvirker den finansielle situasjonen til et selskap, blant annet interne prosesser, makroøkonomiske forhold, konkurranse i markedet og finansielle nøkkeltall. Eksisterende konkursprediksjonsmodeller benytter i all hovedsak finansielle nøkkeltall for å si noe om sannsynligheten for at et selskap går konkurs (Maricica og Georgeta, 2012). En av grunnene til at slike modeller er utbredt er at finansielle nøkkeltall er offentlig tilgjengelig, og det er relativt kostnadseffektivt å sammenligne finansielle nøkkeltall på tvers av selskap.

Konkursprediksjon bedrer finansmarkedenes evne til å formidle kapital til lavest mulig kostnad. Konkursprediksjonsmodeller regnes som et viktig hjelpemiddel for blant annet banker, investorer, forvaltere, kredittratingbyråer og kriserammede selskap (Altman et al., 2014). Disse institusjonene bruker konkursprediksjonsmodeller for å finne konkurssannsynligheten for enkelt-selskap. Modeller som predikerer konkurser blir også benyttet av myndighetene til å si noe om aggregerte konkurssannsynligheter og bankenes forventede utlånstap.

Norges Bank bruker aktivt konkursprediksjonsmodeller i sitt arbeid med å skape finansiell

stabilitet (Norges Bank, 2017). Sentralbanken bruker estimerte konkurssansynligheter vektet med foretakenes regnskapsførte bankgjeld som et mål på andelen konkursutsatt bankgjeld. Estimatet på konkursutsatt bankgjeld blir brukt som en indikator på kredittrisikoen som kan knyttes til foretakene. Selv om det ikke er en direkte sammenheng mellom konkurser og utlånstap i bankene kan vi i figur 1.1 se at de har en tilnærmet lik utvikling.

Figur 1.1 – Bankenes utlånstap, gjeld i konkursrammede foretak og estimert konkursutsatt gjeld. Prosent. 2007-2018. (Norges Bank, 2017)



1) Utlånstap som andel av totale utlån til foretak.

2) Regnskapsført bankgjeld i foretak som er registrert konkurs 1-2 år etter siste leverte regnskap som andel av total bankgjeld.

3) Modellanslag for 2017 og 2018.

Konkursprediksjonsmodeller kan også brukes til å estimere konkurssansynlighet for et spesifikt selskap. Kreditratingbyråer bruker lignende modeller for å estimere sannsynligheten for at et selskap går konkurs, noe som er utslagsgivende for hvilken "karakter" selskapet får (Petersen, Plenborg og Kinserdal, 2017). Banker og andre finansielle institusjoner benytter disse ratingene i vurderingen av potensielle debitorer.

En måte å utvide og potensielt forbedre eksisterende konkursprediksjonsmodeller på er å inkludere data fra andre kilder enn årsregnskapet. Skatteetaten er en slik kilde. Skatteetaten fungerer som en kreditor for svært mange aksjeselskap i Norge, og har informasjon om etterlevelse av blant annet merverdiavgift, arbeidsgiveravgift, forskuddstrekk og restskatt for norske selskap. Såkalte termindata fra Skatteetaten inneholder informasjon om

selskapet som ikke finnes i regnskapet, og som genereres hyppigere enn foretakenes årsregnskap. På bakgrunn av dette tror vi termindata fra Skatteetaten har potensiale til å bedre eksisterende konkursprediksjonsmodeller.

1.2. Problemstilling

I denne oppgaven skal vi undersøke hvordan termindata påvirker konkursprediksjon. Vi estimerer ulike konkursprediksjonsmodeller ved hjelp av etablerte statistiske metoder. Modellene vil ha en avhengig variabel som forteller om et selskap går konkurs eller ikke, mens forklaringsvariablene er ulike nøkkeltall. Studien vår skiller seg fra andre studier ved at nøkkeltallene ikke kun er basert på foretakenes årsregnskap, men også er opprettet ved hjelp av termindata fra Skatteetaten.

Hovedproblemstillingen vår for denne oppgaven er: *I hvilken grad kan termindata fra Skatteetaten brukes for å forbedre dagens regnskapsbaserte konkursprediksjonsmodeller?*

1.3. Oppgavens struktur

Oppgaven er delt inn i 8 kapitler. I kapittel 2 presenteres bakgrunn for oppgaven. I kapittel 3 omtales tidligere arbeid som er gjort innenfor konkursprediksjon. I kapittel 4 utdyper vi hvilke metoder vi har valgt å bruke for å estimere og validere modellene. I kapittel 5 følger en beskrivelse av datasettet, og hvordan vi har valgt avhengige variabler og forklaringsvariabler. I kapittel 6 presenteres analyser og resultater, mens vi i kapittel 7 tolker resultatene og deres implikasjoner. Til slutt vil vi i kapittel 8 konkludere og svare på problemstillingen.

2. Bakgrunn

2.1. Avgrensning

Denne masteroppgaven tar for seg konkurser på selskapsnivå, og er avgrenset til ikke-finansielle aksjeselskap (AS). Selskapsformer som allmennaksjeselskap (ASA), ansvarlig selskap (ANS) og enkeltpersonsforetak (ENK) blir ikke diskutert i denne oppgaven. Da vi blant annet tar i bruk mva-data, vil kun mva-pliktige selskap inngå i våre analyser. Vi analyserer ikke aksjeselskap som har mindre enn 500 000 i totale eiendeler. En mer detaljert beskrivelse av hvordan og hvorfor vi har avgrenset datasettet finnes i kapittel 5.

2.2. Definisjon av konkurs

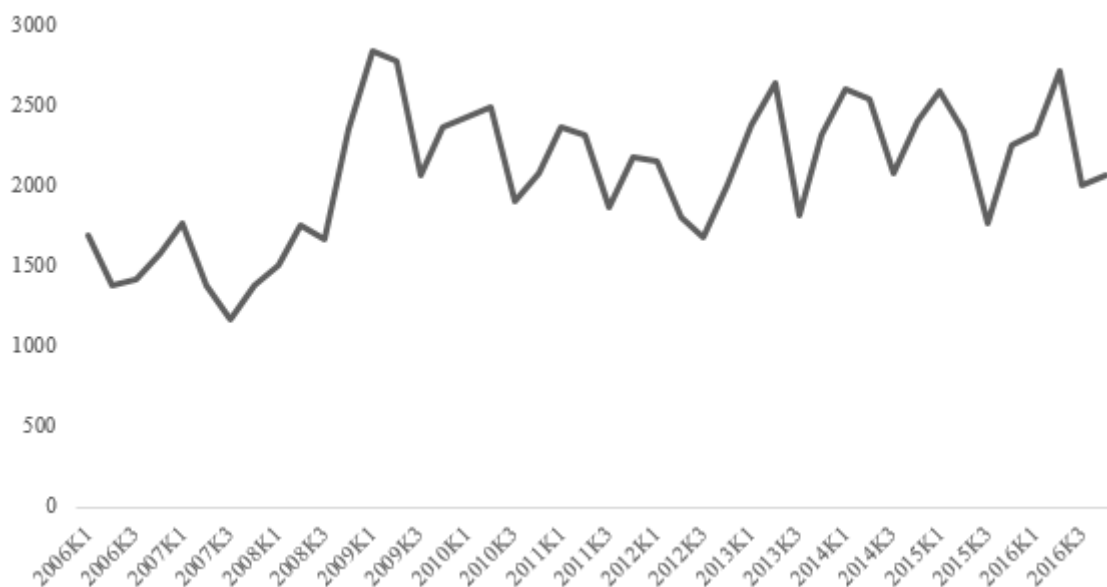
En konkurs oppstår når debitor ikke klarer å betale sine finansielle forpliktelser til sine långivere (Brækhus, 2015). Et selskap som går konkurs kalles en skyldner. Kreditorer som har utestående penger hos en låntaker, og i tillegg anser kunden å være insolvent, kan be retten om å åpne konkurs (Oslo Byfogdembete, 2017). I tillegg kan et aksjeselskap begjære seg selv konkurs.

I Konkursloven (1997) §61 står det at debtors bo skal tas under konkursbehandling dersom vedkommende er insolvent. I §61 defineres skyldneren som insolvent når denne ikke kan oppfylle sine forpliktelser når de forfaller. En skyldner kan kun begjæres konkurs dersom betalingsudyktigheten ikke er forbigående. En skyldner anses ikke som konkurs dersom en kombinasjon av skyldnerens inntekter og eiendeler tilsammen dekker de utestående forpliktelsene. En konkursprosess kan være tidkrevende og kostbart, og derfor er ofte konkursdatoen tingretten setter en stund etter at selskapet de facto er konkurs (Oslo Byfogdembete, 2017).

2.3. Overordnet om konkurser

Det finnes flere grunner til at virksomheter går konkurs. Det er i seg selv vanskelig å overleve i et marked som er preget av hard konkurranse. Selskap må ha priser som er konkurransedyktige, men som samtidig dekker nødvendige kostnader. I nedgangstider er det observert at flere selskap opplever finansielle vanskeligheter. I løpet av det siste tiåret har virksomheter i Norge blitt preget av blant annet finanskrisen som hadde store ringvirkninger. Utviklingen i olje- og gassektoren siden 2014 har også hatt negative konsekvenser for flere norske selskap. Fra Statistisk Sentralbyrå kan vi i figur 2.1 se følgende utvikling rundt antall konkurser:

Figur 2.1 – Antall konkurser per kvartal. (Statistisk Sentralbyrå, 2017a)



Figur 2.1 viser at det er flere konkurser i 2009. Det kan være en konsekvens av finanskrisen som brøt ut i 2008. I tillegg viser grafen at det er sesongvariasjoner. Det ser ut til at mange selskap blir slått konkurs i første og andre kvartal. Dette kan være et naturlig tidspunkt for konkurs da man i forbindelse med årsoppgjøret kan finne at fortsatt drift ikke er mulig.

Alle økonomier vil fluktuere ved å gå gjennom konjunktursykler. En konjunktursyklus er definert i antall perioder med vekst (ekspansjon) eller avtakende vekst (lavkonjunktur) (Investopedia, 2017). I en høykonjunktur er konsumentene villige til å investere og bruke penger. I en lavkonjunktur faller inntektene til konsumentene reelt sett, og det fører til

en lavere investeringsvilje. Dette har en negativ effekt på virksomheter, da kundene har mindre å bruke på deres produkter eller tjenester. Det fører til at selskapenes inntekt faller.

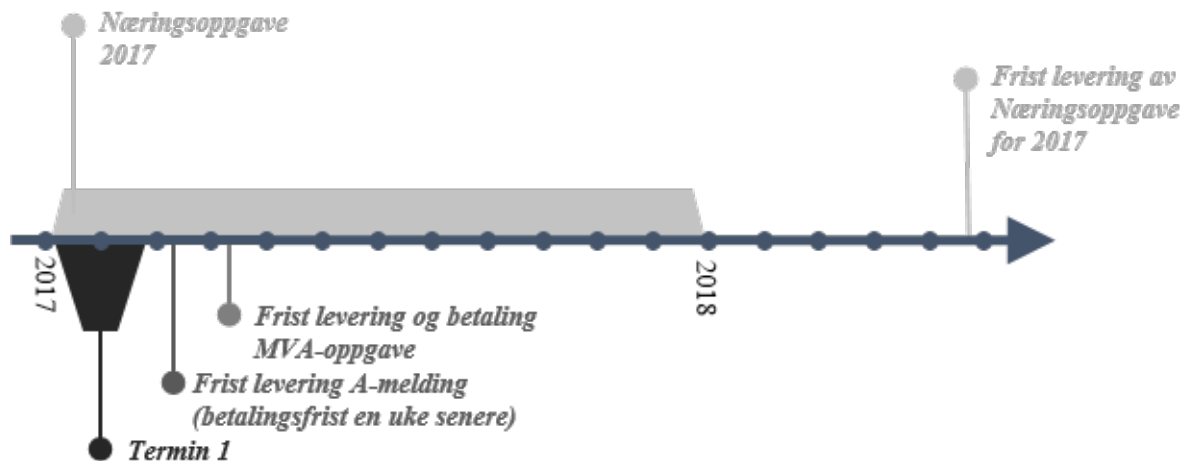
I 2010 ble det etablert cirka 46 000 foretak. Fem år senere, i 2015, var rundt 12 500 av disse fortsatt aktive (Statistisk Sentralbyrå, 2017b). Det vil si at rundt 73 prosent av selskapene som startet gikk konkurs, opphørte eller gjennomførte andre typer selskapsendringer i løpet av en fem års periode. Her skilte aksjeselskapene seg ut som de mest overlevelsesdyktige med en overlevelsesgrad på tilnærmet 49 prosent, mens enkeltmannsforetak var på 21 prosent. Denne statistikken viser tydelig at det er vanskelig å overleve de første årene etter man har startet et selskap.

Det faktum at det er mange involverte parter i enhver konkurs er noe av grunnen til at man stadig hører om konkurser og konkursbo i nyhetene. Under en konkurs skal verdiene som er igjen i selskapet fordeles mellom kreditorer og eiere, hvor kreditorer med pant- eller annen sikkerhetsrett har fortrinnsrett til dekning i de verdiene de har pant- eller sikkerhetsrett i (Konkursrådet, 2017). I tillegg koster konkurser samfunnet penger ved at man binder opp store ressurser i å avvikle et selskap på best mulig måte.

2.4. Termindata

Skatteetaten er et forvaltningsorgan ansvarlig for å innkreve skatter og avgifter på vegne av staten (Store Norske Leksikon, 2017). Termindata er i denne oppgaven definert som en samlebetegnelse på selskapsdata som omhandler merverdiavgift, forskuddstrekk, arbeidsgiveravgift og restskatt. Mva-pliktige aksjeselskap skal levere mva-oppgave og a-melding etter hver termin. De fleste selskap har seks pliktige terminer i året. Mva-oppgaven dokumenterer hvor mye merverdiavgift som skal betales til Skatteetaten, mens a-meldingen viser blant annet arbeidsgiveravgift og forskuddstrekk. Mva-oppgave skal leveres og betales ca halvannen måned etter avsluttet termin. A-melding har leveringsfrist ca en uke etter avsluttet termin, og forskuddstrekk og arbeidsgiveravgift skal betales ca en uke etter levering. Til sammenligning er fristen for å levere regnskapstall gjennom næringssoppgaven 31. mai året etter det aktuelle regnskapsåret. Figur 2.2 viser Skatteetatens innleveringsfrister for regnskapsåret 2017.

Figur 2.2 – Skatteetatens frister



Hver sirkel representerer starten på en måned.

2.5. Endringer i lovverket

I 2011 ble det vedtatt to endringer i Aksjeloven vedrørende revisorplikten og minstekravet til aksjekapital. Fra og med 01.01.2012 ble minstekravet redusert fra 100.000,- til 30.000,- NOK (Kobro, 2012). Det ble også mulig å belaste aksjekapitalen for stiftelsesomkostningene som før måtte holdes utenfor aksjekapitalen. I tillegg ble det gjort endringer i revisjonsplikten som gjør at aksjeselskap i noen tilfeller kan velge bort revisor. Totalt sett ble det enklere å starte aksjeselskap etter at disse endringene trådte i kraft. Det kan tenkes at det derfor er blitt relativt flere nyoppstartede aksjeselskap enn enkeltmannsforetak. Statistisk Sentralbyrå publiserer hvert femte år hvor mange selskap som har overlevd i den foregående fem års perioden. Når Statistisk Sentralbyrå publiserer konkurstallene for 2017 vil man få et inntrykk av om endringene i lovverket har ført til en reduksjon i overlevelsesgraden til aksjeselskap.

Fra og med 2017 kan Skatteetaten ilegge en daglig løpende tvangsmulkt dersom en virksomhet ikke leverer mva-meldingen, skattemeldingen eller andre pliktige opplysninger i tide (Skatteetaten, 2017). På sikt kan dette føre til at termindata kan ha en større effekt på konkursprediksjon. Seriøse og levedyktige aktører vil dermed ha et sterkere insentiv til å levere meldingene tidsnok. Kun de virksomhetene som virkelig sliter – og kanskje da vil gå mot konkurs – vil ikke kunne gjøre dette. Disse selskapene vil også raskere bygge

opp gjeld til Skatteetaten. Etersom ordningen er såpass ny vil det ikke være mulig å se på effekter av endringen i denne oppgaven, men er noe som kan være interessant å se på i fremtidige oppgaver.

3. Tidligere arbeid

I dette kapitlet presenteres ulike teorier, metoder og modeller knyttet til konkursprediksjon. I litteraturen finner man hovedsakelig to typer modeller for konkursprediksjon, regnskapsbaserte modeller og markedsbaserte modeller (Berg, 2007). Moody's EDF-modell (Expected Default Frequency) er et eksempel på en markedsbasert modell (Nazeran og Dwyer, 2015). Denne typen modeller tar utgangspunkt i selskapets markedsverdi hvor aksjekurs som regel blir brukt som en tilnærming. Modeller som baserer seg på markedsverdier krever dermed at selskapene er børsnoterte. Regnskapsbaserte modeller bruker i motsetning informasjon fra regnskapet for å predikere konkurs. Da de fleste norske aksjeselskap ikke er børsnoterte vil vi fokusere på regnskapsbaserte modeller (Eklund, Larsen og Bernhardsen, 2001).

Før man fikk kvantitative størrelser på hvordan foretak presterte, etablerte man byråer som hadde i oppgave å gi kvalitativ informasjon vedrørende kredittverdigheten til bestemte aktører (Altman, 1968). Formelle studier rundt selskapskonkurser begynte rundt 1930-tallet, og siden har flere studier konkludert med at bedrifter som går konkurs har signifikant forskjellige nøkkeltallsmålinger fra de bedriftene som fortsetter å operere. William H. Beaver (1966), Edward I. Altman (1968) og James Ohlson (1980) står bak noen av de mest kjente internasjonale studiene på konkursprediksjon ved bruk av finansielle nøkkeltall. I Norge har Norges Bank utviklet en regnskapsbasert konkursprediksjonsmodell kalt SEBRA-modellen.

3.1. Beavers univariate modell

William H. Beavers univariate modell fra 1966 regnes som en av de første studiene innenfor nøkkeltallbasert konkursprediksjon. Univariate analyser ser på hvert enkelt nøkkeltall individuelt og studien antar dermed at en enkelt forklaringsvariabel kan benyttes for å predikere konkurs. Beaver (1966) brukte et parret utvalg hvor han sammenlignet gjennomsnittsverdier til nøkkeltallene for 79 konkursbedrifter og 79 ikke-konkursbedrifter,

hvor et solvent selskap ble paret med et insolvent selskap. Selskapene ble paret slik at hvert par bestod av selskap som tilhørte samme bransje, og var av tilnærmet lik størrelse. På denne måten ble bransje- og størrelseseffekter minimert. Beaver startet med 30 nøkkeltall som senere ble redusert til følgende seks nøkkeltall:

$$i = \frac{\textit{Kontantstrm}}{\textit{Samlet gjeld}}$$

$$ii = \frac{\textit{Netto resultat}}{\textit{Totalkapital}}$$

$$iii = \frac{\textit{Samlet gjeld}}{\textit{Totalkapital}}$$

$$iv = \frac{\textit{Arbeidskapital}}{\textit{Totalkapital}}$$

$$v = \frac{\textit{Kortsiktige eiendeler}}{\textit{Kortsiktig gjeld}}$$

$$vi = \frac{\textit{Likvide eiendeler}}{\textit{Kortsiktig gjeld}}$$

Svakheten med univariat metode er at man kan få ulike konklusjoner for ulike nøkkeltall for samme selskap avhengig av hvor tungt nøkkeltallene vektet (Altman, 1968). Et nøkkeltall kan predikere konkurs, mens et annet nøkkeltall predikerer ikke-konkurs for samme selskap. Dette betyr at modellen ikke tar hensyn til sammenhengen mellom nøkkeltallene for en og samme bedrift.

3.2. Altmans Z-score

Edward I. Altman (1968) utviklet en konkursprediksjonsmodell som tar i bruk multivariat lineær diskriminantanalyse. Multivariat diskriminantanalyse (MDA) er en statistisk metode som egner seg for studier hvor den avhengige variabelen er dikotom (Hair et al., 1998). Metoden klassifiserer en observasjon i en av flere a priori grupperinger avhengig av observasjonens egne egenskaper. Etter at gruppene er etablert og data innhentet, forsøker MDA å finne en lineær kombinasjon av de ulike egenskapene som "best" skiller de uli-

ke gruppene. En fordel ved bruk av multivariat diskriminantanalyse sammenlignet med univariate metoder er at metoden vurderer en mengde kjennetegn som er felles for de relevante firmaene, i tillegg til interaksjonen mellom disse egenskapene.

Resultatet fra en MDA er på formen $Z = v_1x_1 + v_2x_2 + \dots + v_nx_n$, og transformerer individuelle variabelverdier til en diskriminant score eller Z-verdi som videre blir bruk til å klassifisere objektet. Hvor

$v_1, v_2, \dots, v_n =$ Diskriminante koeffisienter

$x_1, x_2, \dots, x_n =$ Uavhengige variabler

Altman (1968) samlet et utvalg av 66 selskap som var delt likt i to grupper – konkurs og ikke-konkurs. Konkurs-selskapene i utvalget var blitt begjært konkurs i perioden 1946 – 1965, mens selskapene med fortsatt drift var aktive i 1966. Modellen baserer seg på 22 finansielle nøkkeltall valgt med bakgrunn i popularitet i tidligere studier og relevans for Altmans studie, i tillegg til noen nye nøkkeltall. Fem av disse 22 ble valgt på bakgrunn av at den kombinasjonen best predikerte selskapskonkurser. Prosessen for å komme frem til de valgte måltallene var iterativ, hvor bidraget fra hele variabelprofilen ble evaluert. Det vil si at den endelig valgte gruppen av variabler ikke består av de mest signifikante variablene målt uavhengig, men av de som samlet gir best prediksjonsverdi.

De fem nøkkeltallene blir multiplisert med koeffisienter fra diskriminantanalysen for å få den såkalte Z-scoren. Ved bruk av et selskaps finansielle informasjon vil man få en Z-skår som impliserer selskapets sannsynlighet for konkurs. Hvis Z-skåren er under 1.81, har selskapet en høy sannsynlighet for å gå konkurs, mens en skår i intervallet 1.81 – 3.0 tilsier at selskapet er i en gråson, og en score over 3.0 tilsier solid drift.

Altmans modell kan utledes på følgende form:

$$Z = 1.2X_1 + 1.4X_2 + 3.3X_3 + 0.6X_4 + 1.0X_5$$

hvor

$$X_1 = \frac{\textit{Arbeidskapital}}{\textit{Totalkapital}}$$

$$X2 = \frac{\textit{Tilbakeholdt overskudd}}{\textit{Totalkapital}}$$

$$X3 = \frac{\textit{EBIT}}{\textit{Totalkapital}}$$

$$X4 = \frac{\textit{Markedsverdi EK}}{\textit{Bokfrt verdi EK}}$$

$$X5 = \frac{\textit{Omsetning}}{\textit{Totalkapital}}$$

$Z = \textit{Diskriminant score}$

3.3. Ohlsons logistiske regresjonsmodell

James Ohlson (1980) valgte å bruke betinget logistisk regresjon med en "maximum likelihood" estimator til å predikere selskapskonkurser. Metoden er en alternativ fremgangsmåte til multippel diskriminantanalyse, og er en generalisert lineær modell (GLM). Ohlson argumenterer for at logistisk regresjon er bedre enn MDA blant annet på grunn av at resultatet fra MDA er en skår som i seg selv har liten intuitiv tolkning. MDA er i utgangspunktet kun et ordinale rangeringsverktøy, mens logistisk regresjon gir tolkbare koeffisienter.

Logistisk regresjon er en statistisk metode for å analysere datasett som har en eller flere forklaringsvariabler som avgjør utfallet. Utfallet er målt med en binær variabel som kun har to mulige utfall, «1» for sann/suksess eller «0» for falsk/nederlag. Målet med logistisk regresjon er å finne den modellen som best beskriver sammenhengen mellom den binære variabelen (den avhengige variabelen), og de uavhengige forklaringsvariablene.

Logistisk regresjon definerer sannsynligheten for konkurs slik:

$$= \frac{1}{1 + e^{(-y)}}$$

der $y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$.

Datagrunnlaget i Ohlsons (1980) studie var betydelig større enn det Altman benyttet. Ohlson tok utgangspunkt i et utvalg av 105 konkursselskap, og 2058 ikke-konkursselskap. Modellen har ni variabler, som ble valgt på bakgrunn av enkelhet. Variablene inneholder

informasjon om den finansielle statusen til selskapet og lønnsomhetsvariabler.

Ohlsons (1980) konkursprediksjonsmodell kan utledes på følgende form:

$$y = 1.32 - 0.407X1 + 6.03X2 - 1.43X3 + 0.076X4 - 2.37X5 - 1.33X6 \\ + 0.285X7 - 1.72X8 - 0.521X9$$

hvor

$$X1 = \textit{Størrelse}$$

$$X2 = \frac{\textit{Total gjeld}}{\textit{Totale eiendeler}}$$

$$X3 = \frac{\textit{Kortsiktige eiendeler} - \textit{Kortsiktig gjeld}}{\textit{Totalkapital}}$$

$$X4 = \frac{\textit{Kortsiktig gjeld}}{\textit{Kortsiktige eiendeler}}$$

$$X5 = \frac{\textit{Resultat}}{\textit{Totalkapital}}$$

$$X6 = \frac{\textit{Driftsresultat}}{\textit{Total gjeld}}$$

$$X7 = \textit{Dummy1}$$

$$X8 = \textit{Dummy2}$$

$$X9 = \frac{\textit{Endring i resultat}}{(|\textit{Resultat}_t| + |\textit{Resultat}_{(t-1)}|)}$$

hvor Dummy1 er lik «1» hvis selskapet hadde negativt resultat de siste to årene og «0» ellers, og Dummy2 er lik «1» hvis total gjeld overstiger totale eiendeler og «0» ellers.

3.4. Norges Banks SEBRA-modell

Norges Bank opprettet i 1987 en database med regnskapsdata for norske selskap (Smogeli, 1987). SEBRA-databasen (System for EDB-Basert Regnskapsanalyse) inneholder årsregnskap fra og med 1981. Eivind Bernhardsen utviklet i 2001 en konkursprediksjons-

modell kalt SEBRA-modellen. Modellen er utviklet for Norges Bank og bruker regnskapsvariabler fra SEBRA-databasen for å predikere konkurser. Sentralbanken bruker SEBRA-modellen i sine analyser av bankenes kredittrisiko ovenfor foretakssektoren (Eklund, Larsen og Bernhardsen, 2001). Modellen predikerer individuelle konkurssansynligheter som kan aggregeres for å få et bilde av den totale risikoen i foretakssektoren. Modellen er i senere tid blitt videreutviklet flere ganger med tanke på hvilke forklaringsvariabler som inngår i modellen. I 2007 ble SEBRA-modellen revidert og videreutviklet til to modeller kalt SEBRA-basis og SEBRA-utvidet (Bernhardsen og Larsen, 2007).

Datagrunnlaget til SEBRA-basis og SEBRA-utvidet består av årsregnskap og informasjon om selskapets alder, størrelse og bransjetilhørighet (Bernhardsen og Larsen, 2007). Datasettet inneholder norske aksjeselskap for perioden 1990 til 2002. Finansielle selskap og selskap med bokførte verdier på under 500 000 NOK blir ekskludert fra datasettet. Det resterende datasettet består av om lag en million årsregnskap, hvorav 20 000 representerer konkurser.

Modellen er utformet som en variant av logistisk regresjon, slik at den gir individuelle konkurssansynlighetsestimater for hvert enkelt selskap (Eklund, Larsen og Bernhardsen, 2001). Fordelen med denne tilnærmingen er at individuelle estimater kan knyttes direkte til selskapets gjeld og dermed predikere det enkelte selskaps gjeldsrisiko. Metoden er klassifisert som en generalisert additiv modell (GAM), og er en utvidelse av generalisert lineær modell (GLM). GAM blir nærmere forklart i kapittel 4.

SEBRA-modellen predikerer konkurssansynligheter ved hjelp av nøkkeltall fra norske aksjeselskaps årsregnskap, samt et par andre selskapskarakteristikker (Eklund, Larsen og Bernhardsen, 2001). Det er gjennomført en omfattende søkeprosess for å finne forklaringsvariabler som er både bedriftsøkonomisk relevante og statistisk signifikante. Tabell 3.1 viser variablene som inngår i SEBRA-basis og SEBRA-utvidet.

Forklaringsvariablene i SEBRA-modellen inkluderer nøkkeltall og/eller dummyvariabler fra følgende kategorier: inntjening, soliditet, likviditet, alder og størrelse. Modellene inneholder også variabler for gjennomsnittlig inntjening og soliditet i bransjen selskapet tilhører, samt standardavviket for inntjening innad i bransjen.

Tabell 3.1 – Variablene som inngår i Norges Banks SEBRA-modell. (Bernhardsen og Larsen, 2007)

Variabeldefinisjon	Variabeltype	Varierer over
Ordinært resultat før av- og nedskrivninger i prosent	Nøkkeltall	Foretak/år
	Gjennomsnitt	Foretak/år
	Standardavvik	Bransje/år
	Korrelasjon med Norgesporteføljen	Bransje
Egenkapital i prosent av totalkapital	Nøkkeltall	Foretak/år
	Gjennomsnitt	Bransje/år
Innskutt egenkapital mindre enn bokført egenkapital	Indikator	Foretak/år
Likvider minus kortsiktig Alder (år) = 1,2,3,...,8	Nøkkeltall	Foretak/år
	Indikatorer	Foretak/år
Sum eiendeler	Nøkkeltall	Foretak/år
Leverandørgjeld i prosent av totalkapital	Nøkkeltall	Foretak/år
Skyldige offentlige avgifter i prosent av totalkapital	Nøkkeltall	Foretak/år

Variabler merket i grått inngår i SEBRA-basis, mens SEBRA-utvidet innholder samtlige variabler i tabellen.

SEBRA-modellene brukes i dag av Norges Bank og Finanstilsynet for å analysere bankenes kredittrisiko overfor foretakssektoren (Norges Bank, 2017; Finanstilsynet, 2017). For Norges Bank og Finanstilsynet er hovedformålet med modellen å følge utviklingen i kredittrisikoen i foretakssektoren på aggregert nivå, men modellen kan også predikere individuelle konkurssansynligheter. De fleste norske banker bruker lignende modeller for å prise lån, velge satsingsområder og prioritere ressursinnsats i engasjementsarbeidet (Eklund, Larsen og Bernhardsen, 2001). I 2016 ble det også utarbeidet en versjon av SEBRA-modellen som inkluderer bransjespesifikke makroøkonomiske faktorer, samt kredittratinger (Hjelseth og Raknerud, 2016).

3.5. Maskinlæring

Kumar og Ravi (2007) går i sin artikkel gjennom forskning som er gjort på konkursprediksjon i perioden 1968 til 2005. De gir en oversikt over ulike metoder benyttet i perioden, og skiller mellom to ulike teknikker for å løse konkursprediksjonsproblemet, statistiske teknikker og intelligente teknikker. Den brede kategorien statistiske teknikker inkluderer flere av metodene diskutert i dette kapittelet, blant annet lineær diskriminant analyse, multivariat diskriminant analyse og logistisk regresjon. Av intelligente teknikker forklares diverse maskinlæringsteknikker, inkludert nevrale nettverk, support vector machine, k-nærmeste naboer og klassifiseringstrær.

Arthur Samuel (2000) definerer maskinlæring som en samlebetegnelse for metoder som har muligheten til å lære uten at de eksplisitt er blitt programmert. Metodene har som motiv å generalisere basert på erfaring (Alpaydin, 2014). I vår kontekst vil generalisering bety å prestere på nye og usette oppgaver etter å ha bygget opp erfaringer på et læringsdatasett.

James et al. (2014) definerer begrepet statistisk læring som metoder for å modellere og forstå komplekse datasett. Begrepet overlapper sterkt med maskinlæring. Generalisert additiv modell (GAM), som vi tar i bruk i denne oppgaven, representerer et slikt grensetilfelle mellom statistisk læring og maskinlæring. GAM lærer ved å lete etter ikke-lineære sammenhenger i datasettet, og en slipper dermed å manuelt prøve forskjellige transformasjoner for hver variabel individuelt. Denne egenskapen ved statistisk læring gjør det mulig å finne skjulte mønstre i datagrunnlaget. GAM blir nærmere forklart i kapittel 4.

3.6. Sammenligning mellom tidligere arbeider

I dette kapittelet har vi sett på tidligere arbeid innen regnskapsbaserte konkursprediksjonsmodeller. De statistiske metodene som benyttes varierer og vi har diskutert univariat analyse, lineær diskriminantanalyse (LDA), generalisert lineær modell (GLM), generalisert additiv modell (GAM), samt maskinlæringsteknikker. Det er interessant å se at selv om metodene varierer mellom modellene, er forklaringsvariablene som blir benyttet relativt like. Samtlige av forklaringsvariablene som blir benyttet i de ulike modellene kan kategoriseres slik: inntjening, soliditet, likviditet, størrelse, alder og bransje. Måten de

ulike modellene velger å representere de ulike kategoriene på varierer i noen grad. Vi avslutter dette kapittelet med å oppsummere de seks kategoriene av forklaringsvariabler.

Inntjening

Over tid må inntjening dekke foretakets betalingsforpliktelser; hvis ikke vil likviditeten til selskapet svekkes (Eklund, Larsen og Bernhardsen, 2001). Inntjening er også viktig for å kunne hente fersk kapital som gjeld og ny egenkapital. Eksempelvis bruker Altman EBIT som andel av totale eiendeler for å representere inntjening, mens SEBRA-modellen bruker ordinært resultat før av- og nedskrivninger i prosent av total gjeld.

Soliditet

Foretak som har høy egenkapitalandel har større evne til å tåle tap (Eklund, Larsen og Bernhardsen, 2001). I dårlige tider fungerer egenkapital som en buffer mot konkurs. Høy egenkapitalandel gjør det også mulig for selskap å skaffe midler gjennom salg av eiendeler uten heftelser, samt at det generelt betyr lavere løpende utgifter til renter og avdrag. Soliditet kan også gi et akkumulert bilde av hvordan den historiske inntjeningen for foretaket har vært. I SEBRA-modellen er soliditet representert ved to variabler. Den første variabelen er egenkapital i prosent av total kapitalen. Den andre er en dummy-variabel for svekket egenkapital, om innskutt egenkapital er mindre enn bokført egenkapital. Ohlson bruker total gjeld over totale eiendeler.

Likviditet

I mange tilfeller er manglende likviditet en utløsende årsak til konkurs (Eklund, Larsen og Bernhardsen, 2001). Det er derfor viktig at likviditet inngår som en eller flere forklaringsvariabler i en konkursprediksjonsmodell. Her benytter de ulike modellene flere nøkkeltall med små variasjoner. For eksempel brukes betalingsmidler minus kortsiktig gjeld som andel av driftsinntekter, skyldige offentlige avgifter som andel av total kapital, leverandørgjeld som andel av total kapital og kortsiktig gjeld over kortsiktige eiendeler.

Alder

Eklund, Larsen og Bernhardsen (2001) viser til at både deres testresultater og studier i andre land tilsier at konkurshyppigheten er større blant nyetablerte foretak. De forklarer at årsakene kan være at nyetablerte selskap har mindre kompetanse, mindre tilgang på

kapital og færre gunstige forretningsforbindelser. Alder inngår som dummyvariabler i Norges Banks SEBRA-modell.

Størrelse

Små selskap er generelt mer konkursutsatt enn store selskap (Eklund, Larsen og Bernhardsen, 2001). Små selskap er typisk mer sårbare for enkelthendelser da de normalt har en smal produktportefølje og opererer i et begrenset geografisk område. Mindre selskap er også ofte nyetablerte og er derfor utsatt for de samme problemene som unge selskap er. Både Ohlsons modell og Norges Banks SEBRA-modell inkluderer størrelse som en forklaringsvariabel. I SEBRA-modellen inngår størrelse som logaritmen av totale eiendeler (Eklund, Larsen og Bernhardsen, 2001).

Bransje

I SEBRA-modellen inngår følgende tre bransjevariabler: gjennomsnittlig egenkapitalandel, gjennomsnittlig inntjening og standardavvik for inntjening.

Bransjer med lav gjennomsnittlig egenkapitalandel har normalt høyere konkurshyppighet (Eklund, Larsen og Bernhardsen, 2001). En forklaring kan være at slike bransjer preges av høyere konkurranse og dermed lav profitt. Det kan også være slik at færre velger å etablere seg i bransjer hvor bankene krever høy egenkapitalandel, og på den måten siler man ut useriøse og svake foretak i slike bransjer. Bransjegjennomsnitt av egenkapitalandel inngår som forklaringsvariabel i SEBRA-modellen.

Det er grunn til å anta at selskap som operer i bransjer med lav avkastning har større konkurssannsynlighet. I SEBRA-modellen inngår bransjegjennomsnitt av ordinært resultat før av- og nedskrivninger i prosent av total gjeld som forklaringsvariabel (Bernhardsen og Larsen, 2007).

I bransjer med stor risiko vil avkastning normalt ha større svingninger enn i bransjer med mindre risiko (Eklund, Larsen og Bernhardsen, 2001). Et selskap som opererer i en bransje med høyt standardavvik i avkastningen vil dermed være mer konkursutsatt. I SEBRA-modellen inngår bransjestandardavvik for variabelen ordinært resultat før av- og nedskrivninger i prosent av total gjeld.

4. Metode

Konkursprediksjon er et klassifiseringsproblem som gjennomføres i to trinn, estimering og validering (Kumar og Ravi, 2007). I det første trinnet benyttes et utvalg av datasettet for å estimere en modell. I det andre trinnet valideres modellen ved at den testes på et annet utvalg av datasettet. En god konkursprediksjonsmodell vil korrekt klassifisere selskap som konkurs og ikke-konkurs med høy treffprosent. I dette kapitlet går vi gjennom metodene vi bruker for å estimere og validere modellene våre.

4.1. Estimering

Vi bruker en ikke-parametrisk logistisk regresjonsmodell for å predikere konkurser. Modellen er en variant av en generalisert additiv modell (GAM), utviklet av Hastie og Tibshirani (1986). GAM brukes i SEBRA-modellen som anvendes av både Norges Bank og Finanstilsynet (Norges Bank, 2017; Finanstilsynet, 2017). Daniel Berg (2007) påviser dessuten at GAM er signifikant bedre enn lineær diskriminerings analyse, tradisjonell logistisk regresjon og nevralt nettverk til å predikere konkurser. Ranik Wahlstrøm og Fredrik Helland (2016) finner i sin masteroppgave om konkursprediksjon at GAM presterer bedre enn samtlige maskinlæringsteknikker i studien deres.

GAM er en ikke-parametrisk metode, og skiller seg dermed fra de tradisjonelle parametriske metodene som lineær regresjon og logistisk regresjon. Vi starter med å forklare en parametrisk logistisk regresjon for deretter å forklare hvordan den kan utvides til en ikke-parametrisk logistisk regresjonen.

4.1.1. Parametrisk logistisk regresjon

En logistisk regresjon er en regresjonsmodell hvor den avhengige variabelen er kategorisk (James et al., 2014). I stedet for å modellere den avhengige variabelen direkte, modellerer en logistisk regresjon sannsynligheten for at den avhengige variabelen tilhører en bestemt kategori. I konkurssammenheng vil den avhengige variabelen enten representere konkurs

eller ikke-konkurs, og en logistisk regresjon vil dermed gi konkurssannsynligheten til et selskap. For eksempel kan konkurssannsynligheten til et selskap gitt forklaringsvariabelen gjeld skrives på følgende form:

$$P(\text{Konkurs} = 1 \mid \text{gjeld}) \quad (4.1)$$

Sannsynligheten $P(\text{Konkurs} = 1 \mid \text{gjeld})$ vil være mellom 0 og 1. For hver verdi av gjeld kan modellen gi en konkurssannsynlighet for selskapet. Fordelen med logistisk regresjon fremfor lineær regresjon er nettopp det at den gir sannsynligheter mellom 0 og 1. Problemet med lineær regresjon blir tydelig i figur 4.1. Til venstre ser vi forholdet mellom $P(\text{Konkurs} = 1 \mid \text{gjeld})$ og forklaringsvariabelen gjeld ved bruk av en lineær regresjon. Når gjeld er under 500, er konkurssannsynligheten negativ, noe som ikke gir mening. For å unngå dette problemet bruker logistisk regresjon en link-funksjon som sikrer sannsynligheter mellom 0 og 1 og gir S-formen til høyre i figur 4.1. En logistisk regresjon kan da uttrykkes på følgende form:

$$P(\text{Konkurs} = 1 \mid X_1, \dots, X_n) = \frac{1}{(1 + e^{-(\alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n)})} \quad (4.2)$$

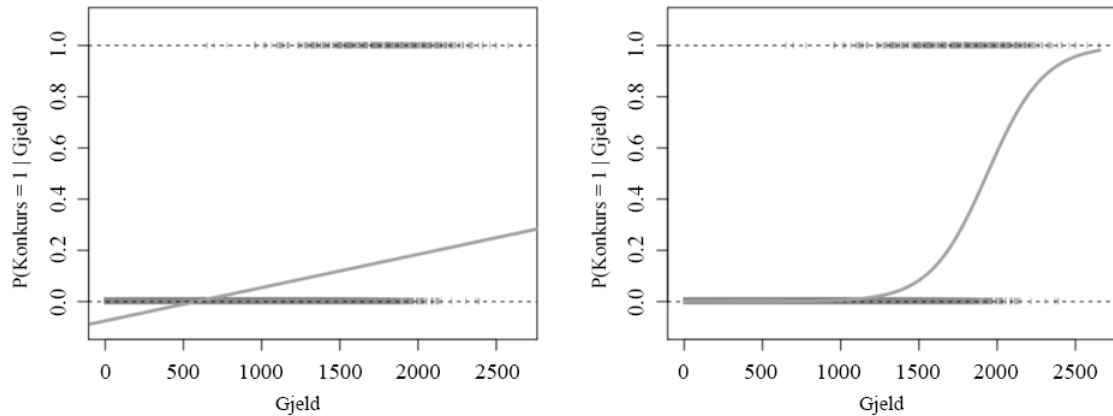
Her er X_1, X_2, \dots, X_n forklaringsvariablene, $\beta_1, \beta_2, \dots, \beta_n$ representerer koeffisientene til forklaringsvariablene og α er konstantleddet. Funksjonen kan transponeres, og skrives da på følgende form:

$$\log \frac{\mu}{1 - \mu} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n \quad (4.3)$$

hvor μ representerer sannsynligheten $P(\text{Konkurs} = 1 \mid X_1, \dots, X_n)$.

Logistisk regresjon tar som regel i bruk "maximum likelihood" for å finne verdiene på koeffisientene og konstantsleddet (Nelder og Wedderburn, 1972). "Maximum likelihood" finner parameterverdiene som maksimerer sannsynligheten for den faktisk observerte stikkprøven. Metoden setter først en tentativ løsning for deretter å starte en itereringsprosess som ser etter bedre løsninger.

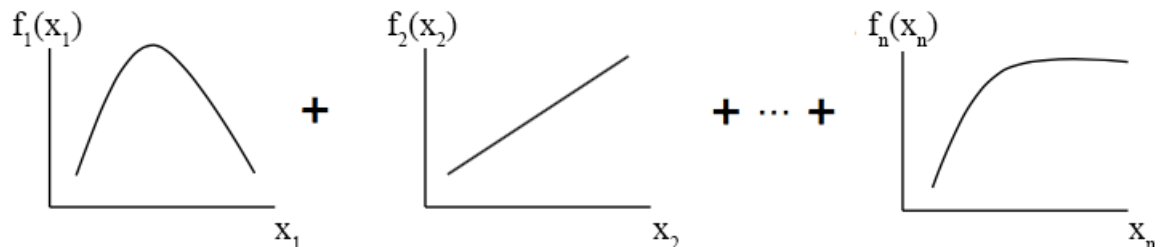
Figur 4.1 – Lineær regresjon sammenlignet med logistisk regresjon (James et al., 2014)



Den logistiske regresjonen vi har beskrevet er parametrisk, noe som tilsier at modellen blir estimert med et forhåndsbestemt antall parametere (James et al., 2014). Dette innebærer at funksjonsformen er gitt og at estimeringen av modellen er redusert til problemet med å finne et sett koeffisienter. En logistisk regresjon har en forhåndsbestemt lineær form ved at det kun er en koeffisient per forklaringsvariabel. Man antar dermed at $\log \frac{\mu}{1-\mu}$ har et lineært forhold til forklaringsvariablene. Forklaringsvariablene er da antatt enten positivt korrelert med den avhengige variabelen for alle intervaller, eller negativt korrelert med den avhengige variabelen for alle intervaller. Ved å velge en logistisk modell gjør man dermed en antakelse om forholdet mellom hver forklaringsvariabel og konkurssannsynligheten. Hvis denne antakelsen ikke stemmer og det sanne forholdet er mer komplekst, vil en ikke-parametrisk modell prestere bedre.

4.1.2. Ikke-parametrisk logistisk regresjon

Figur 4.2 – Den additive formen til GAM (Larsen, 2016)



Trevor Hastie og Robert Tibshirani (1986) utviklet GAM for å løse utfordringene knyttet til antakelsene ved parametriske modeller. Målet med modellen er å ta høyde for

ikke-lineære effekter i forklaringsvariablene, og på denne måten avsløre skjulte mønstre i datagrunnlaget. Modellen erstatter den lineære formen til en parametrisk logistisk regresjon $\sum \beta_j X_j$ med en additiv form som består av en sum av uspesifiserte funksjoner $\sum f_j(X_j)$, visualisert i figur 4.2. GAM kan på lik linje med logistisk regresjon gi sannsynligheter mellom 0 og 1, og kalles da GAM-logit. GAM-logit uttrykkes på følgende form:

$$P(\text{Konkurs} = 1 \mid X_1, \dots, X_n) = \frac{1}{(1 + e^{-(\alpha + f_1(x_1) + f_2(x_2) + \dots + f_n(x_n)))}} \quad (4.4)$$

Her representerer $f(x_1), f(x_2), \dots, f(x_n)$ de uspesifiserte funksjonene. Funksjonen kan transponeres, og skrives da på følgende form:

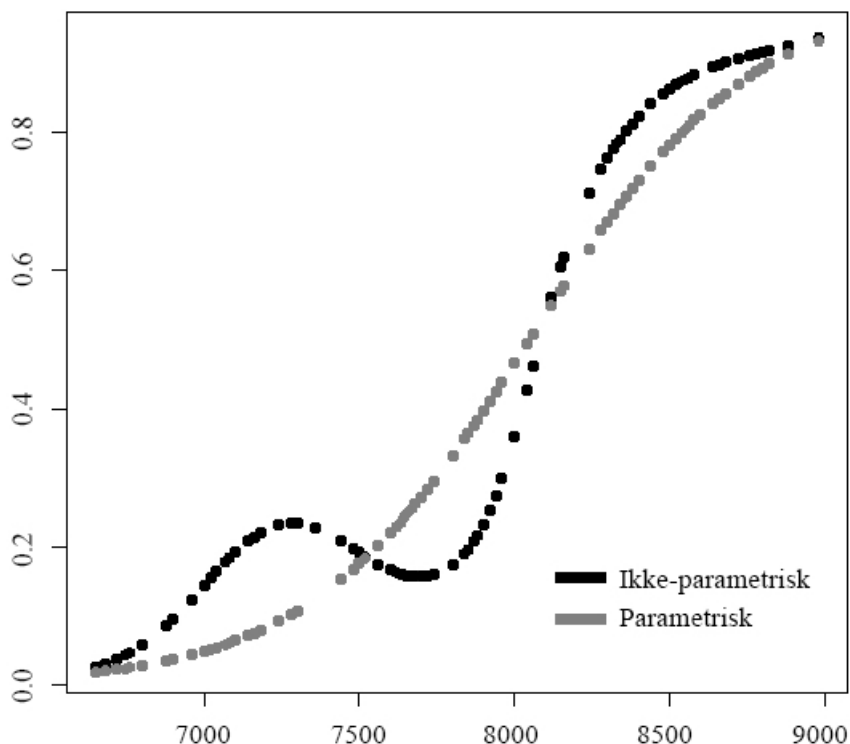
$$\log \frac{\mu}{1 - \mu} = \alpha + f_1(x_1) + f_2(x_2) + \dots + f_n(x_n) \quad (4.5)$$

hvor μ representerer sannsynligheten $P(\text{Konkurs} = 1 \mid X_1, \dots, X_n)$.

Funksjonsformene til hver forklaringsvariabel $f_j(X_j)$ er uspesifisert, og blir automatisk konstruert under modellestimering (Hastie og Tibshirani, 1986). Dette gjør GAM til en mer fleksibel modell da den på forhånd ikke har sterke antakelser om forholdet mellom den avhengige variabelen og forklaringsvariabelen. I motsetning vil en parametrisk modell, som logistisk regresjon, kunne tvinge fram en tilpasning som ikke representerer virkeligheten på kritiske punkter (Berg, 2007). Figur 4.3 viser hvordan en parametrisk modell gir et upresist bilde på den virkelige sammenhengen mellom den avhengige variabelen og forklaringsvariabelen, mens en ikke-parametrisk modell gir et bedre bilde. GAM beregner funksjoner for variablene individuelt, og unngår dermed problemet med at variansen øker drastisk når man inkluderer flere forklaringsvariabler i modellen, også kalt "curse of dimensionality" (Berg, 2007).

Den statistiske utfordringen ved GAM er å estimere parameterne til de uspesifiserte funksjonene $f_j(X_j)$ i modellen (Larsen, 2016). Hvis alle funksjoner var tillatt i modellestimeringen ville "maximun likelihood" estimere en svært kompleks modell som sannsynligvis ville lidd av "overfitting", hvor modellen forklarer støy istedenfor underliggende sammen-

Figur 4.3 – En parametrisk logistisk modell sammenlignet med en ikke-parametrisk logistisk modell



Figuren viser hvordan en ikke-parametrisk modell fanger opp et mønster som en parametrisk modell overser og istedet tvinger en tilpasning. Den aktuelle figuren viser at i intervallet 7300 til 8000 er den uavhengige variabelen negativt korrelert med den avhengige variabelen.

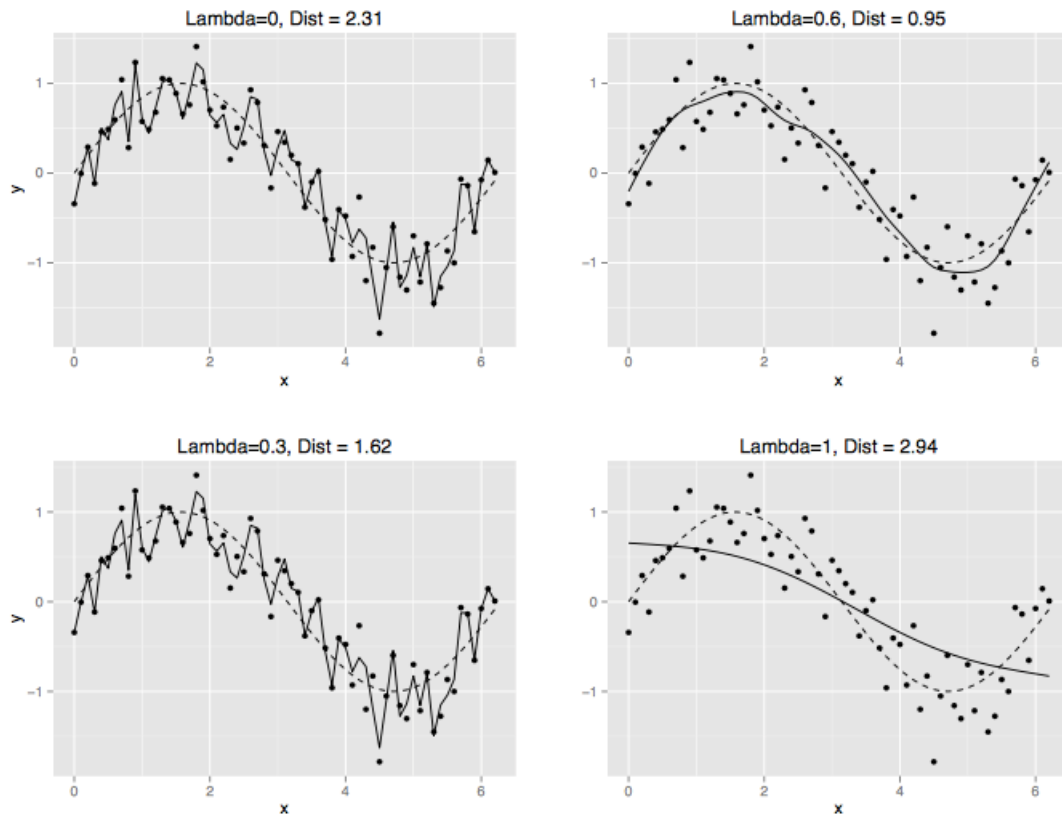
henger. GAM benytter derfor en prosess kalt ”smoothing” for å unngå dette problemet. I GAM bruker man en modifisert ”maximum likelihood” som inkluderer en straff når funksjonen $f_j(X_j)$ øker i kompleksitet, og straffer dermed ”wiggliness” i funksjonen. Å finne et kompromiss mellom å straffe ”wiggliness” og å straffe ”badness of fit” er selve kjernen i GAM. Figur 4.4 illustrerer fire ulike løsninger for dette kompromisset.

For å finne parameterne til de uspesifiserte funksjonene $f_j(X_j)$ benytter vi penalized regression splines og generalized cross validation (GCV). Penalized regression splines er en algoritme som finner funksjonene $f_j(X_j)$ som minimerer funksjon 4.6.

$$\sum_{i=1}^n \{y_i - f(x_i)\}^2 + \lambda \int \{f''(u)\}^2 du \quad (4.6)$$

Det første leddet i minimeringsproblemet, $\sum_{i=1}^n \{y_i - f(x_i)\}^2$, oppfordrer de uspesifiserte funksjonene $f_j(X_j)$ til å best mulig forklarer y_i (James et al., 2014). I vår GAM-logit

Figur 4.4 – Avveining mellom ”overfitting” og ”underfitting” (Larsen, 2016)



Viser avveiningen mellom å straffe ”wigglines” og å straffe ”badness of fit”. Med λ lik 0 straffes ”badness of fit” hardest, mens med λ lik 1 straffes ”wiggliness” hardest. Den optimale løsningen er typisk et sted i mellom.

modell vil y_i representere hele logit-funksjonen $\frac{\mu}{1-\mu}$. Det første leddet straffer dermed ”badness of fit”. Det andre leddet i minimeringsfunksjonen finner den andre deriverte av de uspesifiserte funksjonene som tilsvarende hvor mye kurven i funksjonen endrer seg. For eksempel vil den andre deriverte av en rett linje være lik 0, og øke hvis kurven blir mer kompleks. Hele integralet, $\lambda \int \{f''(u)\}^2 du$, kan tolkes som et mål på den totale endringen i kurven til funksjonen over hele intervallet til forklaringsvariabelen. Det andre leddet i minimeringsfunksjonen straffer dermed ”wiggliness”. Hvor hardt ”wiggliness” blir straffet avhenger av størrelsen på λ i minimeringsfunksjonen. Setter vi λ lik 0, blir straffen oversett og resulterer dermed i ”overfitting”. Settes λ for høyt, ender vi opp med en modell som lider av ”underfitting”. Figur 4.4 illustrerer hvordan ulike λ påvirker løsningen. Vi finner optimal λ ved å benytte generalized cross validation (GCV). GCV setter $\lambda \geq 0$ som

minimerer funksjon 4.7.

$$\frac{nD}{(n - DoF)^2} \tag{4.7}$$

Her representerer n antall observasjoner, D representerer modellavvik og DoF representerer modellens effektive frihetsgrader (Wood, 2000). Vi ser at den foretrukne λ er den som reduserer modellavvik og reduserer modellens effektive frihetsgrader. Å minimere funksjon 4.7 representer dermed avveiningen mellom ”badness of fit” og ”wiggliness”. Vi benytter R-pakken ”mgcv” for å estimere våre GAM-modeller utviklet av Simon Wood (2011).

GAMs mulighet til å ta i bruk uspesifiserte funksjoner gjør modellen fleksibel ettersom den ikke gjør antakelser om forholdet mellom forklaringsvariablene og den avhengige variabelen (Larsen, 2016). Fleksibiliteten i GAM går på bekostning av at modellen blir noe vanskeligere å tolke. En forklaringsvariabel, x , kan for eksempel være positivt korrelert med den avhengige variabelen for et intervall av x og negativt korrelert for et annet intervall. GAM-logit vil derfor ikke kunne produsere entydige koeffisienter for hver variabel. I motsetning vil en parametriske logistisk regresjon gi koeffisienter hvor fortegn og størrelse kan tolkes for å si noe om hvilken retning og hvor stor grad forklaringsvariabelen påvirker den avhengige variabelen. Valget mellom en parametriske og en ikke-parametriske modell er derfor ofte en avveining mellom presise prediksjoner på den ene siden og lett tolkbare koeffisienter på den andre siden. Multikollinearitet kan by på utfordringer hvis man prøver å tolke forklaringsvariablenes individuelle effekter på den avhengige variabelen (Berg, 2007). På lik linje med Berg fjerner vi ikke høyt korrelerte forklaringsvariabler, og man bør dermed uansett være forsiktig med å bruke våre modeller til å si noe presist om den enkelte forklaringsvariabelens effekt på konkurransen. Når formålet med modellen i all hovedsak er å predikere, vil verken tolkningsproblemer eller multikollinearitet være et problem (Berg, 2007).

4.2. Validering

Etter at man har estimert en modell, er det naturlig å teste modellens prediksjonsevne. For å validere og måle modellens prediksjonsevne bruker vi et ”out-of-time”-datautvalg og illustrerer resultatet ved hjelp av receiver operating characteristic (ROC) og area under the curve (AUC).

4.2.1. Out-of-time

Prediksjonsevnen til en modell kan være svært sensitivt for hvilke utvalg av datasettet man bruker for validering (Sobehart, Keenan og Stein, 2000). Det er viktig å ikke gjøre modellen avhengig av datasettet modellen ble konstruert med, og vi skiller derfor ut et treningssett og et testsett. Modellen blir estimert ved hjelp av et utdrag av datasettet kalt treningssettet, for deretter å bli validert mot av et annet utdrag kalt testsettet. Testsettet kan enten være ”out-of-sample” eller ”out-of-time”. ”Out-of-sample” vil i vårt datasett bety at vi estimerer modellen med data fra en gruppe selskap for en periode, mens vi validerer modellen mot en annen gruppe selskap for samme periode. ”Out-of-time” bruker derimot et testsett fra en annen periode enn treningssettet. Konkursprediksjonsmodeller brukes for å si noe om sannsynligheten for at et selskap går konkurs i fremtiden, og man vil derfor ønske å teste modellen for selskap i en senere periode. En ”out-of-time” validering vil dermed være av størst interesse i utviklingen av konkursprediksjonsmodellen, hvor vi lager et treningssett for en periode og et testsett for en senere periode.

4.2.2. Receiver Operating Characteristics (ROC)

I delkapittelet om estimering har vi forklart hvordan GAM-logit produserer konkurssannsynligheter for hvert selskap. For å kunne måle prestasjonsevnen til modellen er vi avhengig av å sette en terskelverdi for hvor høy konkurssannsynligheten skal være før vi klassifiserer selskapet som konkurs. Man kan for eksempel si at hvis konkurssannsynligheten er større enn 0.5, skal selskapet klassifiseres som et konkursselskap. Alternativt kan man sette terskelverdien slik at man får et akseptabelt nivå for feilklassifiseringer av konkursselskap. Ønsker man å predikere korrekt i for eksempel 95 % av konkurstilfel-

lene, setter man terskelverdien lavt nok for å få dette utfallet. Dette går naturligvis på bekostning av at man også klassifiserer flere ikke-konkursselskap som konkurs.

Receiver Operating Characteristics (ROC) er en metode for å visualisere prestasjonen til en rekke klassifiseringer (Fawcett, 2006). Når man predikerer en binær hendelse slik som i konkursprediksjon, vil hver prediksjon kunne plasseres i en av fire klassifiseringsutfall; ekte positiv, ekte negativ, falsk positiv og falsk negativ. Et selskap som er predikert konkurs og faktisk går konkurs, blir klassifisert som ekte positiv. Tilsvarende blir et selskap som predikeres ikke-konkurs og faktisk ikke går konkurs, klassifisert som ekte negativ. Et selskap som derimot blir predikert konkurs men ikke går konkurs, blir klassifisert som falsk positiv, såkalt type 1 feil. Falsk negativ er når et selskap predikeres ikke-konkurs men faktisk går konkurs, en type-2 feil. En konkursprediksjonsmodell vil sette en konkurssannsynlighet for hvert enkelt selskap i testsettet. En bruker kan velge hvor mange type-1 feil eller type-2 feil som tolereres ved å sette en terskel for hva konkurssannsynligheten skal være for at et selskap skal kategoriseres som konkurs. En lav terskel vil føre til flere type-1 feil, mens en høy terskel vil føre til flere type-2 feil.

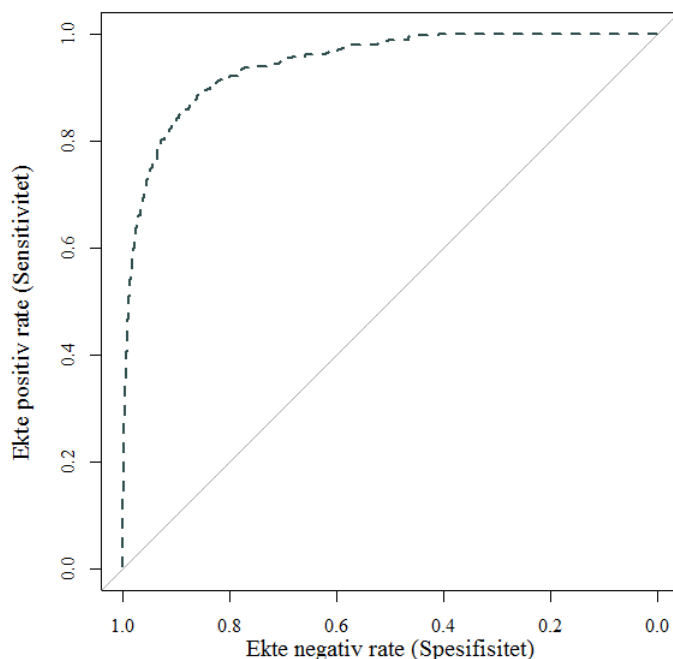
Tabell 4.1 – Type I og type II feil

		Predikert verdi	
		Konkurs	Ikke konkurs
Ekte verdi	Konkurs	Ekte positiv	Falsk negativ
	Ikke konkurs	Falsk positiv	Ekte negativ

En ROC-graf brukes for å visualisere avveiningen mellom type 1 og type 2 feil (Fawcett, 2006). X-aksen i en ROC-graf representerer spesifisitet, mens y-aksen representerer sensitivitet. Med spesifisitet menes sannsynligheten for en ekte negativ klassifisering gitt faktisk negativ hendelse. Sensitivitet er sannsynligheten for en ekte positiv klassifisering gitt faktisk positiv hendelse.

I en prediksjon hvor konkurs er den positive hendelsen vil x-aksen representere andel riktig klassifiserte ikke-konkursselskap. Y-aksen vil representere andel riktig klassifiserte konkursselskap. Merk at x-aksen er fallende, mens y-aksen er stigende. Av kurven i figur

Figur 4.5 – Eksempel på en ROC-kurve



4.5 ser vi at hvis man øker sensitiviteten, reduseres spesifisiteten. Dette tilsier at hvis man ønsker å ha høy treffsikkerhet på konkursselskap, vil man få lavere treffsikkerhet på ikke-konkursselskap. Den stiplete kurven representerer her en modell med god prediksjonsevne, mens den diagonale linjen representerer en modell som tilfeldig klassifiserer hvert selskap.

$$\text{Sensitivitet} = \frac{\text{Ekte positive klassifiseringer}}{\text{Totalt antall positive hendelser}} \quad (4.8)$$

$$\text{Spesifisitet} = \frac{\text{Ekte negative klassifiseringer}}{\text{Totalt antall negative hendelser}} \quad (4.9)$$

I konkursprediksjon kan det antas at kostnadene er ulik for type-1 feil og type-2 feil. For eksempel vil det å innvilge lån til et selskap som går konkurs typisk være dyrere enn å ikke innvilge lån til et selskap som ikke går konkurs. Konkursprediksjon har vanligvis også svært få positive hendelser, konkurser. Å måle prediksjonsevnen ved å kun telle antall riktige klassifiseringer vil dermed være problematisk. Hvis en ser for seg at av 100 selskap

er det kun ett selskap som går konkurs, vil man oppnå en treffsikkerhet på 99 % ved å predikere alle selskapene ikke-konkurs. Et slikt treffsikkerhetsmål er missvisende og vi benytter et alternativt mål på modellens prediksjonsevne, area under the curve (AUC). Prediksjonsevnen til en modell kan tallfestes ved å beregne arealet under ROC-grafen (Fawcett, 2006). AUC vil ha en verdi på mellom 0 og 1. En modell som predikerer perfekt, vil ha en AUC lik 1, mens en modell som tilfeldig klassifiserer selskap, den diagonale linjen i figur 4.5, vil ha en AUC lik 0.5. AUC har følgende statistiske tolkning: AUC er lik sannsynligheten for at en tilfeldig valgt positiv hendelse vil bli klassifisert høyere enn en tilfeldig valgt negativ hendelse. Vi tar i bruk AUC for å måle prediksjonsevnen til våre modeller. For å beregne og visualisere ROC og finne AUC benytter vi pakken "proc" utviklet av Robin et al (2011).

5. Datasett

Datagrunnlaget vi bruker i denne oppgaven er utsendt av Skatteetaten, og er satt sammen av Bo Vignes i Skatt Vest. Selskapene som inngår i datasettet er anonymisert, ettersom termindata ikke er offentlig tilgjengelig og derfor anses som sensitiv informasjon.

5.1. Beskrivelse av datasettet

Datasettet fra Skatteetaten består av samtlige selskap som har vært oppført med mva-plikt fra første termin 2008 til og med andre termin 2017. Datasettet har en panelstruktur hvor hvert selskap blir observert hver termin opptil seks ganger per år. Variablene som inngår i datasettet er i hovedsak hentet fra to kilder, termindata og næringsoppgaven. I tillegg er det koblet på enkelte selskapsopplysninger fra Enhetsregisteret.

5.1.1. Termindata

Termindata i denne oppgaven er definert som en samlebetegnelse på selskapsdata som omhandler merverdiavgift, forskuddstrekk, arbeidsgiveravgift og restskatt. Datasettet skiller mellom merverdiavgift og annen skatt, hvor annen skatt består av forskuddstrekk, arbeidsgiveravgift og restskatt. Annen skatt oppgis kun summert og blir heretter kalt SOFIE, etter navnet på innkrevingsystemet. Variablene som inngår i termindata inkluderer blant annet hvor mye merverdiavgift selskapet har utestående, hvor mye SOFIE selskapet har utestående, hvorvidt mva-oppgaven er levert og betalbar merverdiavgift. Uttreksdato for hver termin er 20 dager etter forfallsdato for mva-innbetaling for den aktuelle termin. Dette tilsvarer to måneder etter avsluttet termin.

5.1.2. Næringsoppgave

I tillegg til termindata har vi uttrekk fra næringsoppgaven for tilhørende selskap. Næringsoppgaven må leveres av alle som driver næringsvirksomhet, og alle inntekter, kostnader

og balanseposter skal fremgå (Visma, 2017). Næringsoppgaven inneholder dermed samme regnskapskonti som årsregnskapet. Vårt datasett inkluderer enkelte konti på aggregert nivå og vil i enkelte tilfeller være mindre detaljert enn årsregnskapet.

Opplysningene fra næringsoppgaven er koblet på med utgangspunkt i uttrekksdato for termindata og når vi kan anta at opplysningene var tilgjengelige. Det antas at næringsoppgaven ikke er tilgjengelig før i oktober, noe som tilsvarer uttrekksdatoen for fjerde termin. I månedene fra oktober til årsslutt hentes opplysninger fra næringsoppgaven som gjelder for året før, mens i månedene før oktober hentes opplysningene fra næringsoppgaven for to år tidligere. Informasjonen i næringsoppgaven er således siste kjente tilstand, og tilskrives til virksomhetene med utgangspunkt i når vi antar at de var tilgjengelige.

5.1.3. Øvrige data

I tillegg til termindata og opplysninger fra næringsoppgaven inkluderer datasettet informasjon fra Enhetsregisteret. Variablene fra Enhetsregisteret inkluderer blant annet eventuell konkursdato og diverse selskapsdata. Selskapsdata inneholder blant annet informasjon om selskapsform, stiftelsesdato, hvilken bransje selskap opererer i og hvilket skattekontor de forholder seg til.

5.2. Avgrensning

Siden vi undersøker om termindata kan forbedre eksisterende konkursprediksjonsmodeller er det ønskelig å gjøre tilnærmet samme avgrensninger i datasettet som dagens modeller gjør. Våre avgrensninger er derfor i stor grad inspirert av SEBRA-modellen.

Vi har i likhet med SEBRA-modellen valgt å gjøre analysene på ikke-finansielle aksjeselskap. Vi fjerner derfor selskap med næringskode 64, *finansieringsvirksomhet*, 65, *forsikringsvirksomhet og pensjonskasser, unntatt trygdeordninger underlagt offentlig forvaltning* og 66, *tjenester tilknyttet finansierings- og forsikringsvirksomhet*. I likhet med SEBRA-modellen fjerner vi også selskap som har mindre enn 500 000 NOK i totale eiendeler. Datasettet er avgrenset til kun å inkludere selskap som har seks pliktige mva-terminer i året. Dette gjelder de aller fleste aksjeselskap, og vi vil da få like mange observasjoner per år, per selskap.

Gjennom denne oppgaven ønsker vi å bygge modeller som skal predikere konkurs opptil tre år i forveien, noe som betyr at vi må ha informasjon om hvorvidt selskapet går konkurs innen tre år eller ikke. Vi har ikke fullstendig informasjon om hvilke selskap som går konkurs i løpet av 2017 før i 2018. Vi vet dermed ikke om selskap observert i perioden 2014 til 2017 går konkurs om tre år. Datagrunnlaget analysene er basert på er derfor begrenset til perioden 2008 til 2013, mens vi henter konkursdatoer til og med 2016.

Det ble utført en test på data fra næringsoppgaven for å se hvorvidt summen av egenkapital, kortsiktig og langsiktig gjeld tilsvarte totale eiendeler i selskapet. I de få tilfellene hvor det var mer enn 5% avvik ble disse fjernet. Dette ble gjort ettersom balansen til et selskap skal være i balanse, og et for stort avvik fra dette antas å vise til feil inntasting eller andre feilkilder som vi ikke har mulighet å avdekke årsaken til. Noen observasjoner faller ut på grunn av manglende næringsoppgaver. Enkelte selskap har aldri levert næringsoppgaven, og vi fjerner hele selskapet, mens andre mister årsobservasjoner på grunn av manglende data. På grunn av dette forsvinner flere observasjoner fra selskap i dets første leveår, selv om de da har tilgjengelig termindata, men naturlig nok ikke har levert næringsoppgave for tidligere år.

5.3. Beskrivende statistikk

Totalt i perioden 2008 til 2017 inneholder datasettet 229 195 selskap hvorav 17 093 går konkurs en gang i løpet av denne ti års perioden. Tabell 5.1 viser antall selskap i datasettet for hvert år, og hvor mange selskap som går konkurs innen ett, to eller tre års konkurshorisont. Konkursraten over år for ett års konkurshorisont har vært avtakende i perioden fra 2008 frem til 2012, før det i 2013 ser ut til å øke i absolutt forstand.

Ofte er det mindre og yngre selskap som går konkurs, noe som ble understreket i kapittel 2 med den femårige overlevelsesstatistikken fra Statistisk Sentralbyrå. Dette blir bekreftet i vårt datasett. Resultatet i tabell 5.2 viser at selskapets alder har signifikant betydning for sannsynligheten for å gå konkurs. Koeffisientene viser at økende alder bidrar til lavere konkurssannsynlighet relativt til den yngste gruppen av selskap. I tillegg ser vi av fordelingen i appendiks A1 at antall konkurser er sterkt fallende med alder. Det er også slik at noen næringer er mer konkursutsatt enn andre, noe som vises i appendiks A2, hvor

Tabell 5.1 – Fordeling av antall konkurser

År	Antall selskap	1 år	2 år	3 år
2008	83 714	422	2 063	3 276
2009	88 348	411	1 732	2 990
2010	90 166	377	1 702	2 694
2011	92 438	375	1 437	2 626
2012	94 995	329	1 593	2 714
2013	99 076	413	1 735	2 833

Antall selskap som går konkurs innen ett, to eller tre år i perioden 2008 til 2013.

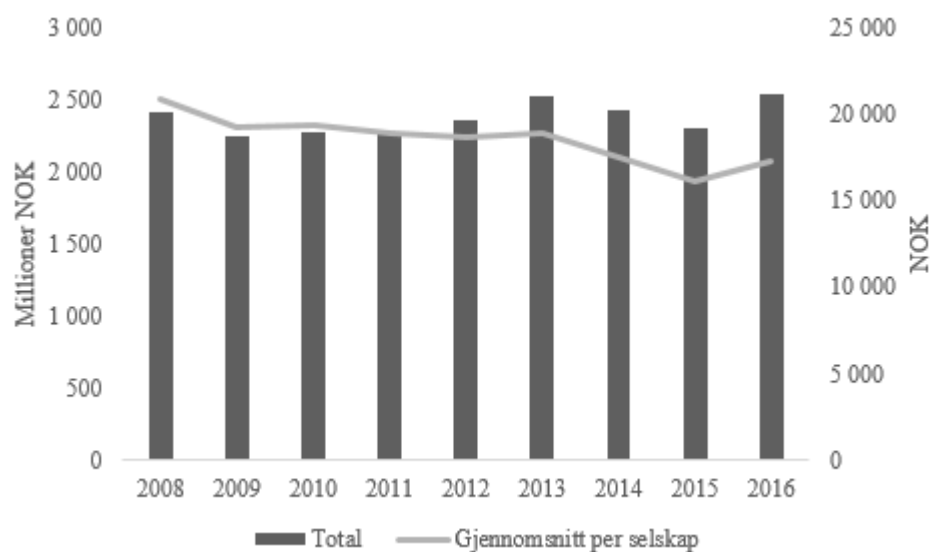
enkelte næringer skiller seg ut hva gjelder antall konkurser.

Tabell 5.2 – Regresjon som viser sammenhengen mellom konkurs og alder

Variabler	(1) konk
Alder 5-9	-0.489*** (0.0190)
Alder 10-25	-1.084*** (0.0199)
Alder > 26	-1.614*** (0.0424)
Konstantledd	-5.508*** (0.0106)
Observasjoner	7,141,463
Standardavvik i parantes	
*** p<0.01, ** p<0.05, * p<0.1	

Det som skiller vår studie fra andre studier er at vi inkluderer termindata. Figur 5.1 viser Skatteetatens utestående mva-gjeld for ulike år. Vi ser at størrelsen på den totale mva-restansen er relativt stabil i perioden 2008 til 2016, men at gjennomsnittlig gjeld per selskap er noe redusert i senere år. Selskapene i vårt datasett hadde i 2016 totalt utestående mva-gjeld på over 2,5 milliarder kroner. Figur 5.1 gir et bilde på størrelsen av Skatteetaten som kreditor for norske selskap.

Figur 5.1 – Utestående mva-gjeld



5.4. Avhengig variabel

I denne oppgaven defineres tre ulike avhengige variabler som representerer ulike konkurshorisonter, konk1, konk2 og konk3. Konk1 representer om selskapet går konkurs innen ett år etter det aktuelle regnskapsåret. Konk2 representer om selskapet går konkurs innen to år. Konk3 representerer om selskapet går konkurs innen tre år. De avhengige variablene tar utgangspunkt i konkursdatoen hentet fra Enhetsregisteret. Variablene er binære og tar kun verdien «1» eller «0», hvor «1» representerer konkurs og «0» ikke-konkurs.

5.5. Forklaringsvariabler

Forklaringsvariablene som inngår i våre modeller kan deles inn i to grupper. For det første inkluderer vi regnskapsbaserte forklaringsvariabler som skal representere eksisterende konkursprediksjonsmodeller. Disse variablene er hentet fra næringsoppgaven. Vi utvider så modellen ved å inkludere flere forklaringsvariabler fra termindata.

5.5.1. Forklaringsvariabler fra regnskapet

Denne oppgaven har ikke som mål å avdekke hvilke regnskapsbaserte forklaringsvariabler som best predikerer konkurs, men å avdekke om termindata kan inkluderes for å bedre eksisterende regnskapsbaserte konkursprediksjonsmodeller. For å lage vår modell har vi

derfor tatt utgangspunkt i de regnskapsbaserte forklaringsvariablene som benyttes i tidligere arbeid, inkludert Beaver (1966), Altman (1968), Ohlson (1980), og i Norges Banks SEBRA-modell 2007. Det går igjen i tidligere studier at man ønsker en kombinasjon av nøkkeltall som dekker inntjening, soliditet, likviditet og størrelse, i tillegg til at man vil ha variabler for bransje og alder på selskapet. Vi har derfor valgt å inkludere tilsvarende nøkkeltall for å representere alle kategoriene. I slutten av kapittel 3 begrunner vi hvorfor variabler som tilhører de ulike kategoriene antas å ha prediksjonsevne.

Inntjening

En variabel som tilsvarende totalkapitalrentabilitet, og er avkastningen på selskapets aktiva i forhold til totalkapitalen.

$$\frac{\textit{Driftsresultat} + \textit{Finansinntekter}}{\textit{Totalkapital}}$$

Soliditet

En variabel som måler hvor stor andel av balansen som består av egenkapital.

$$\frac{\textit{Egenkapital}}{\textit{Totalkapital}}$$

Likviditet

En proxy på variabelen «leverandørgjeld i prosent av totalkapital» fra Norges Banks SEBRA-modell. Viser hvor mye av den totale balansesummen som består av kortsiktig gjeld.

$$\frac{\textit{Kortsiktig gjeld}}{\textit{Totalkapital}}$$

Størrelse

Størrelsen på eiendeler er et mål på hvor stort selskapet er, og vi inkluderer variabelen på logaritmisk form.

$$\ln(\textit{SumEiendeler})$$

Arbeidskapital

Arbeidskapital er hvor mye et selskap trenger for å finansiere løpende utgifter, og viser hvor stor andel av balansen som kreves av driften.

$$\frac{\text{Omlpsmidler} - \text{Kortsiktig gjeld}}{\text{Totalkapital}}$$

Bransjevariabler

For å si noe om bransjekarakteristikker har påvirkning på konkurssannsynlighet, inkluderes følgende bransjevariabler:

Gjennomsnitt soliditet

Variabelen er gjennomsnittsverdien av egenkapitalprosenten i hver bransje hvert år.

Gjennomsnitt inntjening

Variabelen er gjennomsnittsverdien av totalkapitalrentabiliteten i bransjen hvert år.

Standardavvik inntjening per bransje

Variabelen er standardavviket til inntjening i hver bransje over år.

Aldersdummyer

For å kunne si noe om alderen til selskapet påvirker konkurssannsynligheten er det inkludert aldersdummyer. Variabelen stiftelsesdato er satt som oppstartstidspunkt, og er regnet frem til terminåret som den siste observasjon er fra. Vi har valgt å lage aldersdummyer for 0 til 8 år.

- Aldersdummy 1 er lik «1» hvis selskapet er 0 eller 1 år gammelt
- Aldersdummy 2 er lik «1» når selskapet er 2 år gammelt
- Aldersdummy 3 er lik «1» når selskapet er 3 år gammelt
- Aldersdummy 4 er lik «1» når selskapet er 4 år gammelt
- Aldersdummy 5 er lik «1» når selskapet er 5 år gammelt
- Aldersdummy 6 er lik «1» når selskapet er 6 år gammelt
- Aldersdummy 7 er lik «1» når selskapet er 7 år gammelt
- Aldersdummy 8 er lik «1» når selskapet er 8 år gammelt

5.5.2. Forklaringsvariabler fra termindata

Vi inkluderer totalt fire nye variabler fra termindata. Disse er valgt på bakgrunn av økonomisk intuisjon, samt en iterativ prosess hvor vi stod igjen med den kombinasjonen av variabler som maksimerer prediksjonsevnen. Variablene er ikke offentlig tilgjengelig, og er derfor ikke inkludert i eksisterende modeller. Tre av variablene er oppgitt som en andel av selskapets totale eiendeler, og er dermed generert ved å bruke data fra både termindata og fra næringsoppgaven. Under følger terminvariablene vi har valgt å inkludere.

Total mva-restanse og total SOFIE-restanse

Mva total restanse er en variabel som viser hvor mye merverdiavgift selskapet har utestående per statusdato. SOFIE total restanse viser hvor mye andre skatter og avgifter selskapet har utestående per statusdato.

$$\frac{Mva\ total\ restanse}{Sum\ Eiendeler}$$

$$\frac{SOFIE\ total\ restanse}{Sum\ Eiendeler}$$

Restansevariablene gir en oversikt over hvor mye utestående gjeld selskap har til Skatteetaten. Staten er å anse som en streng kreditor, og vi antar at selskap vil prioritere gjeld til Skatteetaten. Når selskap likevel ikke betaler denne gjelden, tyder dette på finansielle vanskeligheter. Disse variablene er mer presise enn for eksempel «Skyldige offentlige avgifter i prosent av totalkapital» som benyttes i SEBRA modellen. Restansevariablene har informasjon direkte fra kreditor, i tillegg til at det er splittet på ulike typer offentlige skatter og avgifter.

Andel mangler

Variabelen andel manglende terminoppgaver beskriver hvor stor andel av de siste seks

pliktige mva-oppgaver som ikke er levert ved uttrekksdato.

Betalbar merverdiavgift

Betalbar merverdiavgift er differansen mellom utgående og inngående merverdiavgift for hver termin. Variabelen benyttes som en proxy på terminresultat. Utgående merverdiavgift er avgift som et selskap har fakturert og krevd inn på vegne av staten, og tilsvarer dermed et mål på omsetning. Inngående merverdiavgift er avgift som et selskap har betalt ved for eksempel varekjøp, og tilsvarer dermed et mål på kostnader.

$$\frac{\textit{Utgående mva} - \textit{inngående mva}}{\textit{Sum Eiendeler}}$$

For å gi et oversikt over hvordan de fire nye terminvariablene påvirker konkurs, har vi inkludert resultatet fra en logistisk regresjon i tabell 5.3. Av regresjonen ser vi at samtlige av de fire inkluderte terminvariablene er signifikante. Koeffesientene til mva-restanse, SOFIE-restanse og andel mangler er positive, noe som tilsier at høye verdier øker sannsynligheten for konkurs. Variabelen betalbar merverdiavgift har en negativ koeffesient, noe som tilsier at høye verdier reduserer konkurssannsynligheten. Det virker intuitivt at restanse og andel manglende oppgaver er positivt korrelert med konkurs, og at betalbar avgift, som er en proxy på terminresultat, er negativt korrelert med konkurs.

Tabell 5.3 – Logistisk regresjon med terminvariabler

	<i>Avhengig variabel:</i>
	konk1
Andel mangler	1.682*** (0.148)
Total mva-restanse	6.863*** (0.255)
Total SOFIE-restanse	3.814*** (0.400)
Betalbar merverdiavgift	−1.785*** (0.297)
Konstantledd	−5.522*** (0.028)
Observasjoner	354,666
Log Likelihood	−9,581.507
<i>Note:</i>	*p<0.1; **p<0.05; ***p<0.01

5.6. Datakvalitet

En utfordring ved vårt datasett er at når selskap er nær konkurs, reduseres datakvaliteten. Flere konkursselskap mangler næringsoppgave i årene før konkurs. Tabell 5.1 viser at det i vårt datasett er relativt få observasjoner av selskap som går konkurs innen ett år. Dette er et resultat av at flere selskapsobservasjoner fjernes fra datasettet rett før konkurs på grunn av manglende næringsoppgave. Denne egenskapen ved datasettet, kombinert med at konkurs er en svært sjelden hendelse, resulterer i at vi har svært få observasjoner av hvordan et selskap ser ut like før konkurs.

5.7. Paneldata

Vårt datasett er på panelform, hvor vi har flere observasjoner av samme selskap over tid. Når samme selskap blir observert flere ganger, kan det tenkes at forutsetningen om uavhengige observasjoner blir brutt. Å behandle hver observasjon separat i en ”pooled” regresjon vil da for det første gi feil inferens og vil for det andre ikke utnytte all informasjon som ligger i datasettet. Bernhardsen (2001) viser i sitt arbeid med SEBRA-datasettet at

man ikke kan forkaste hypotesen om at individuell heterogenitet er fraværende. Dette tilsier at en uavhengig pooled modell ikke er en ukorrekt tilnærming. Vi følger Bernhardsen (2001) og antar uavhengighet.¹

Bernhardsen (2001) undersøkte også om en dynamisk spesifisering av SEBRA-datasettet kunne være hensiktsmessig. Han prøvde blant annet å inkludere lag-variabler for å se om utviklingen for et selskap fra en periode til neste periode hadde prediksjonsevne. Slike variabler viste seg å være lite signifikante. Robusthetsanalyser viser at det heller ikke i vårt datasett er vesentlige dynamiske effekter. Vi tar ikke i bruk lag-variabler, og ser dermed bort fra dynamiske effekter.

¹Bernhardsen (2001) tester dette formelt i sitt datasett og finner støtte for antagelsen.

6. Empirisk analyse

I dette kapitlet gjennomfører vi våre analyser på datasettet presentert i kapittel 5 ved hjelp av metodene presentert i kapittel 4. Først vil vi forklare prediksjonsrammeverket og hvordan de ulike modellene i analysen er bygget opp. Deretter sammenligner vi de ulike modellenes prediksjonsevne, samt at vi analyserer ulike konkurshorisonter og modellenes robusthet.

Det er to egenskaper ved Skatteetatens termindata som potensielt kan bedre eksisterende regnskapsbaserte konkursprediksjonsmodeller. Den første egenskapen er at det kan ligge informasjon i Skatteetatens data som er korrelert med konkurser, men som ikke fanges opp av regnskapsinformasjon. Den andre egenskapen er at termindata rapporteres hyppigere og tidligere enn årsregnskapet. Årsregnskapet for år n blir typisk ikke lagt frem før i midten av år $n+1$. Dette resulterer i at eksisterende regnskapsbaserte modeller ikke har tilgjengelig informasjon om årets første termin før halvannet år senere. I motsetning til det vil Skatteetatens termindata ha informasjon allerede to måneder etter første termin. På bakgrunn av disse to egenskapene ved termindata velger vi å utarbeide tre konkursprediksjonsmodeller:

1. Referansemodell uten termindata basert på eksisterende regnskapsbaserte konkursprediksjonsmodeller.
2. En modell som inkluderer termindata fra samme år som årsregnskapet.
3. En modell som inkluderer termindata fra tidspunktet de er tilgjengelig.

Samtlige modeller er estimert ved bruk av generalisert additiv metode, GAM-logit, forklart i kapittel 4. Ved å sammenligne disse tre modellene undersøkes prediksjonsevnen til de to nevnte egenskapene ved termindata isolert.

6.1. Prediksjonsrammeverk

Prestasjonen til en prediksjonsmodell kan være svært sensitiv for hvilke datasett som blir brukt for validering (Berg, 2007). Det er derfor viktig at observasjoner som blir brukt i

modellestimeringen ikke inngår i datasettet som blir brukt for å validere modellen. Vi estimerer modellen med observasjoner fra en periode og validerer modellen mot observasjoner fra en senere periode, såkalt out-of-time (Sobehart, Keenan og Stein, 2000).

Hvis vi eksempelvis utvikler en modell basert på observasjoner fra regnskapsåret 2008 med en to års konkurshorisont, estimerer vi sannsynligheter for om selskapene går konkurs i perioden 2009 til 2010. Vi kan altså ikke estimere denne modellen før i 2011 når konkursdata for 2010 er tilgjengelig. Deretter kan vi bruke modellen på observasjoner fra 2009 for å predikere om selskapene går konkurs i perioden 2010 til 2011. Vi kan altså ikke validere en slik modell før i 2012 når konkursdata for 2011 er tilgjengelig.

En naturlig tanke vil være å kun benytte ferske data når man bygger en konkursprediksjonsmodell, da de er mest relevante for å si noe om konkurs i fremtiden. Berg (2007) finner derimot i sine studier at konkursprediksjonsmodeller som er estimert med data fra flere år predikerer bedre enn modeller som er estimert kun basert på data fra det siste året. Han argumenterer for at dette er tilfelle fordi man må anta at det er de samme faktorene som fører til at selskap går konkurs hvert år. Hvis man ikke antar dette, vil modellen uansett bare kunne brukes på data fra samme periode den ble estimert på og dermed ikke ha noe prediktiv evne. Vi konstruerer derfor et treningssett som inneholder alle selskap i perioden 2008 til 2011 som brukes for å estimere alle modellene våre. Data fra 2012 og 2013 brukes som to separate valideringssett.

6.2. Modell 1 - Referansemmodell

Vår problemstilling er om termindata fra Skatteetaten kan bedre eksisterende konkursprediksjonsmodeller. Det er derfor ønskelig å estimere en referansemmodell som representerer dagens regnskapsbaserte konkursprediksjonsmodeller. En slik modell kan brukes som sammenligningsgrunnlag for våre nye modeller som inkluderer termindata.

Modell 1 har som mål å ligne mest mulig på eksisterende regnskapsbaserte modeller, og vi bruker derfor tilsvarende forklaringsvariabler som er forklart nærmere i kapittel 5. Listen over forklaringsvariabler som inngår i modell 1 vises i tabell 6.1.

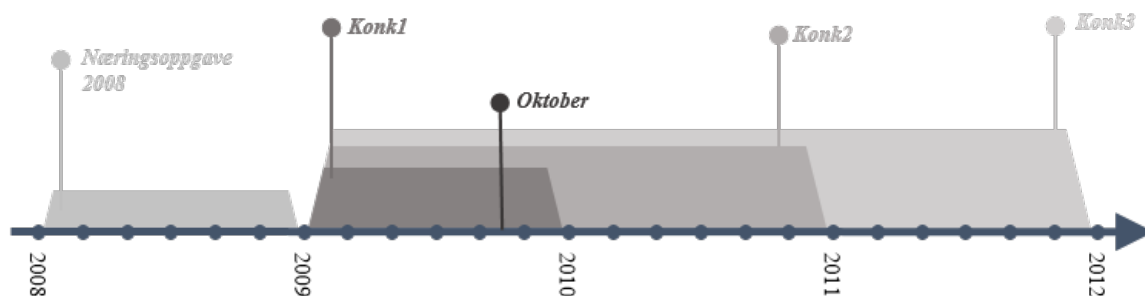
Næringsoppgaven for år n vil være tilgjengelig i oktober år $n+1$. Vi ser dermed for oss at

Tabell 6.1 – Variabler i modell 1

Variabeldefinisjon	Kategori	Variabeltype
Driftsresultat pluss renteinntekter i prosent av totalkapital	Inntjening	Nøkkeltall
	Inntjening i bransjen	Gjennomsnitt
	Inntjening i bransjen	Standardavvik
Egenkapital i prosent av totalkapital	Soliditet	Nøkkeltall
	Soliditet i bransjen	Gjennomsnitt
Kortsiktig gjeld i prosent av totalkapital	Likviditet	Nøkkeltall
Arbeidskapital i prosent av totalkapital	Likviditet	Nøkkeltall
Logaritmen av sum eiendeler	Størrelse	Nøkkeltall
Alder(år) = 1,2,3,4,5,6,7,8	Alder	Indikatorer

vi står med en ny næringsoppgave i oktober og prøver å predikere om foretaket kommer til å gå konkurs. En konkurshorisont på ett år vil da predikere om selskapet går konkurs i løpet av år $n+1$, altså samme år som næringsoppgaven blir publisert. En konkurshorisont på 2 år vil predikere om selskapet går konkurs i løpet av perioden $n+1$ til $n+2$. Til slutt vil en konkurshorisont på 3 år predikere om selskapet går konkurs i løpet av perioden $n+1$ til $n+3$.

Figur 6.1 – Tidfesting i modell 1



Figur 6.1 viser hvordan tidfestingen fungerer i modell 1 med utgangspunkt i 2008 som år n .

6.3. Modell 2 - Termindata fra samme år som årsregnskapet

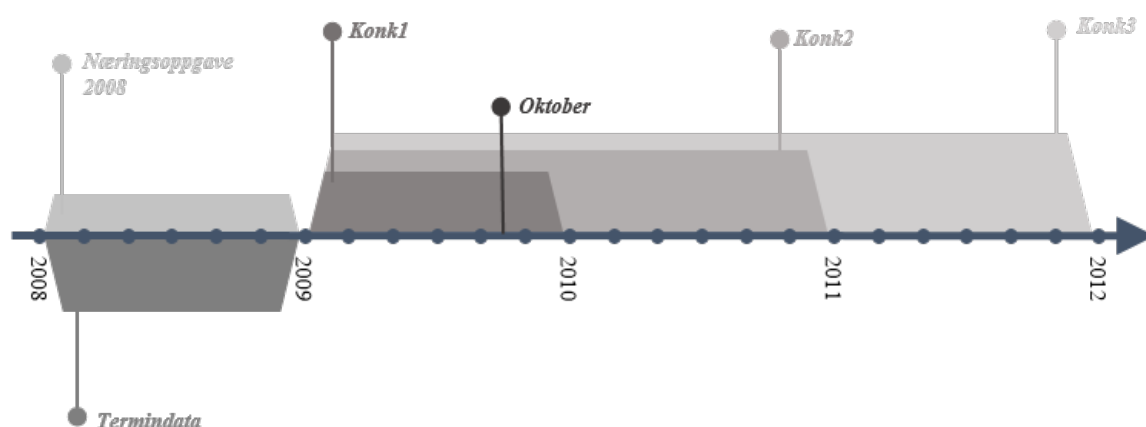
For å undersøke om termindata fra Skatteetaten kan bedre prediksjonsevnen til referansemodellen bygger vi ut modell 1 ved å inkludere nye nøkkeltall. I modell 2 beholder vi samtlige forklaringsvariabler fra modell 1 og inkluderer fire nye nøkkeltall fra termindata. Tabell 6.2 viser alle forklaringsvariablene som inngår i modell 2, hvor de fire nye er markert i grått.

Som i modell 1, ser vi for oss i modell 2 at vi står med en fersk næringsoppgave i oktober, og forsøker å predikere om foretaket kommer til å gå konkurs. Forskjellen er at vi nå også har nøkkeltall fra termindata i modellen. Disse nøkkeltallene er hentet fra samme år som næringsoppgaven. Vi ser dermed for oss at næringsoppgave og termindata for år n blir gjort tilgjengelig samtidig i oktober i år $n+1$. Termindata er i denne modellen omgjort fra seks terminer til å representere selskapet ved årsslutt, slik et årsregnskap gjør. Med en slik tilnærming kan vi uproblematisk kombinere tall fra årsregnskapet med tall fra termindata for å opprette nye nøkkeltall. Total mva-restanse i denne modellen er total utestående mva-gjeld ved årsslutt. Tilsvarende gjelder for total SOFIE-restanse, som representerer arbeidsgiveravgift, forskuddstrekk og restskattekrav. Andel ikke leverte mva-oppgaver beskriver her andelen pliktige terminoppgaver i det aktuelle året som ikke er levert ved årsslutt. Betalbar avgift representerer mva-avgift som er generert i årets siste termin.

Tabell 6.2 – Variabler i modell 2

Variabeldefinisjon	Kategori	Variabeltype
Driftsresultat pluss renteinntekter i prosent av total kapital	Inntjening	Nøkkeltall
	Inntjening i bransjen	Gjennomsnitt
	Inntjening i bransjen	Standardavvik
Egenkapital i prosent av total kapital	Soliditet	Nøkkeltall
	Soliditet i bransjen	Gjennomsnitt
Kortsiktig gjeld i prosent av total kapital	Likviditet	Nøkkeltall
Arbeidskapital i prosent av total kapital	Likviditet	Nøkkeltall
Logaritmen av sum eiendeler	Størrelse	Nøkkeltall
Alder(år) = 1,2,3,4,5,6,7,8	Alder	Indikatorer
Total mva-restanse i prosent av total kapital	Merverdiavgift	Nøkkeltall
Total SOFIE-restanse i prosent av total kapital	SOFIE	Nøkkeltall
Andel ikke leverte mva-oppgaver siste seks terminer	Merverdiavgift	Nøkkeltall
Betalbar merverdiavgift i prosent av total kapital	Merverdiavgift	Nøkkeltall

Figur 6.2 – Tidfesting i modell 2



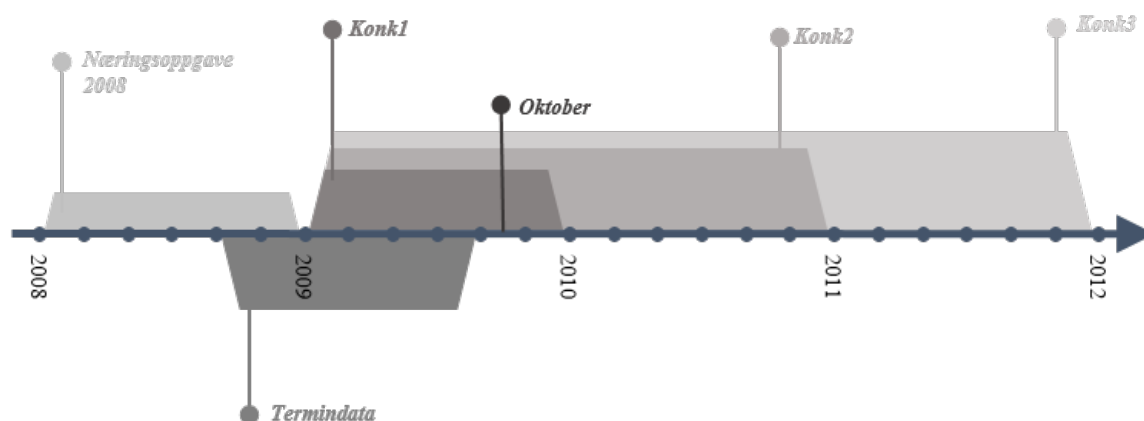
Figur 6.2 viser at modell 2 inneholder termindata fra samme periode som næringsoppgaven.

6.4. Modell 3 - Termindata fra tidspunktet de er tilgjengelig

I modell 3 bruker vi de samme forklaringsvariablene som i modell 2. Som i modell 1 og 2, ser vi også for oss i modell 3 at vi står med en ny næringsoppgave i oktober og prøver å predikere om foretaket kommer til å gå konkurs. Forskjellen fra modell 2 er at vi i modell 3 benytter ferske termindata som er tilgjengelig i oktober. Vi går her bort fra å sammenstille næringsoppgave og termindata for samme periode. Skatteetaten har termindata tilgjengelig allerede to måneder etter den aktuelle terminen, som tilsier at vi i slutten av oktober har tilgjengelig data fra fjerde termin i år $n+1$. Dette er fire terminer senere enn termindata benyttet i modell 2, og vi utnytter dermed at termindata er tilgjengelig tidligere enn næringsoppgaven. Nøkkeltallene i modell 3 er som resultat av dette konstruert av næringsoppgaven for år n og termindata fra fjerde termin år $n+1$.

Figur 6.3 viser at modell 3 inneholder ferske termindata som er tilgjengelig når næringsoppgaven blir tilgjengelig i oktober.

Figur 6.3 – Tidfesting i modell 3



6.5. Sammenligning av de ulike modellene

Når vi sammenligner modell 1 til 3, benytter vi en ett års konkurshorisont og validerer modellene mot observasjoner fra 2013.

Tabell 6.3 – Sammenligning av modeller

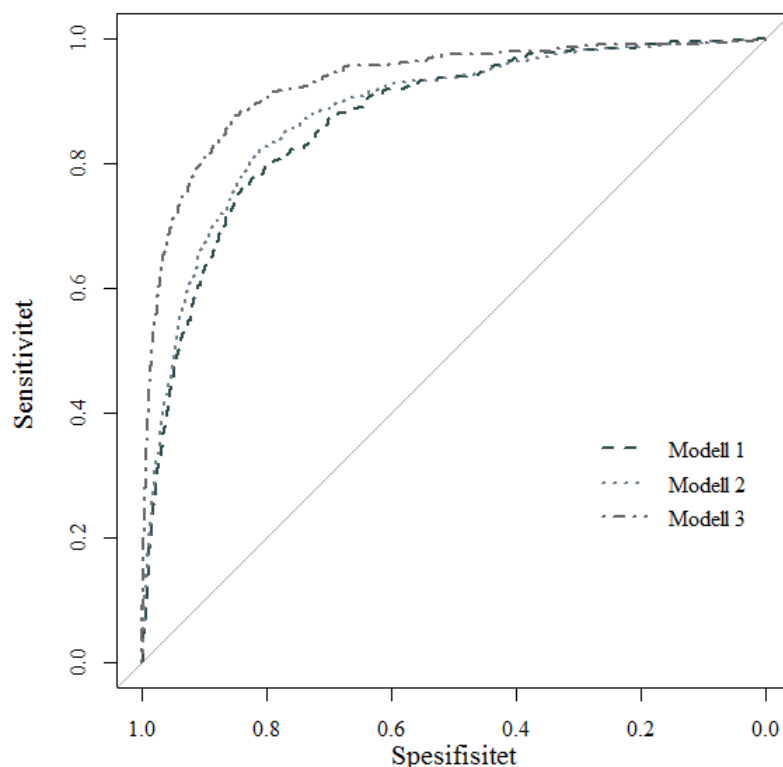
Modell	AUC	95% Konfidensintervall	Signifikans
1	0.8687	[0.852 - 0.8853]	-
2	0.8796	[0.8629 - 0.8963]	T
3	0.9284	[0.9144 - 0.9424]	TT

Signifikans representerer om AUC er signifikant forskjellig fra modellen ovenfor seg. T representer forskjellig på 5% signifikansnivå.

Tabell 6.3 viser AUC-verdiene til de ulike modellene med tilhørende konfidensintervall. Vi ser at både modell 2 og modell 3 er signifikant bedre enn referansemodellen, modell 1. Vi ser også at modell 3, som benytter ferske termindata, oppnår en AUC på hele 0.9284, som er signifikant bedre enn modell 2. Figur 6.4 viser de tilhørende ROC-kurvene, og vi får bekreftet at modell 3 presterer bedre enn de andre modellene. Alle AUC-verdiene er signifikant forskjellige fra hverandre.

For å illustrere hva forbedringen i AUC betyr i praksis, sammenligner vi modell 1 og modell 3 sine totale treffprosent ved ulike terskelverdier. Ulike brukere har ulike preferanser for hvor mange feilklassifiserte konkursselskap de tolererer, og vil derfor kategorisere sel-

Figur 6.4 – ROC-kurver for alle modeller



skap konkurs ved ulike konkurssannsynligheter. Tabell 6.4 viser totale treffprosent ved ulike preferanser. La oss si at en ønsker å sette en terskelverdi som er slik at man korrekt klassifiserer 95 % av konkursselskapene. Som vi ser av tabell 6.4 vil modell 1 oppnå en total treffprosent på 44.97 %, mens modell 3 oppnår en total treffprosent på 67.84 %. Når terskelverdien er satt slik at modell 1 og modell 3 korrekt klassifiserer 95 % av konkursselskapene, vil modell 3 korrekt klassifisere vesentlig flere ikke-konkursselskap.

En annen måte å sammenligne modellenes treffprosent er å sette terskelverdier som resulterer i at andel korrekt klassifiserte konkursselskap er lik andel korrekt klassifiserte ikke-konkursselskap i hver modell (sensitivitet = spesifisitet). Hvis vi setter spesifisitet lik sensitivitet i modell 1, får vi en total treffprosent på 79.2 %. Setter vi sensitivitet lik spesifisitet i modell 3, får vi en total treffprosent på 85.98 %.

Tabell 6.4 – Totale treffprosjenter

Andel korrekt predikerte konkurser	Modell 1	Modell 3
99 %	20.15 %	27.23 %
95 %	44.97 %	67.84 %
90 %	64.10 %	81.02 %
85 %	72.04 %	86.54 %
80 %	79.06 %	90.43 %
50 %	93.92 %	98.17 %

Tabellen viser totale treffprosjenter for modell 1 og modell 3 ved ulike toleranser for feilklassifiseringer av konkursselskap.

6.6. Konkurshorisont

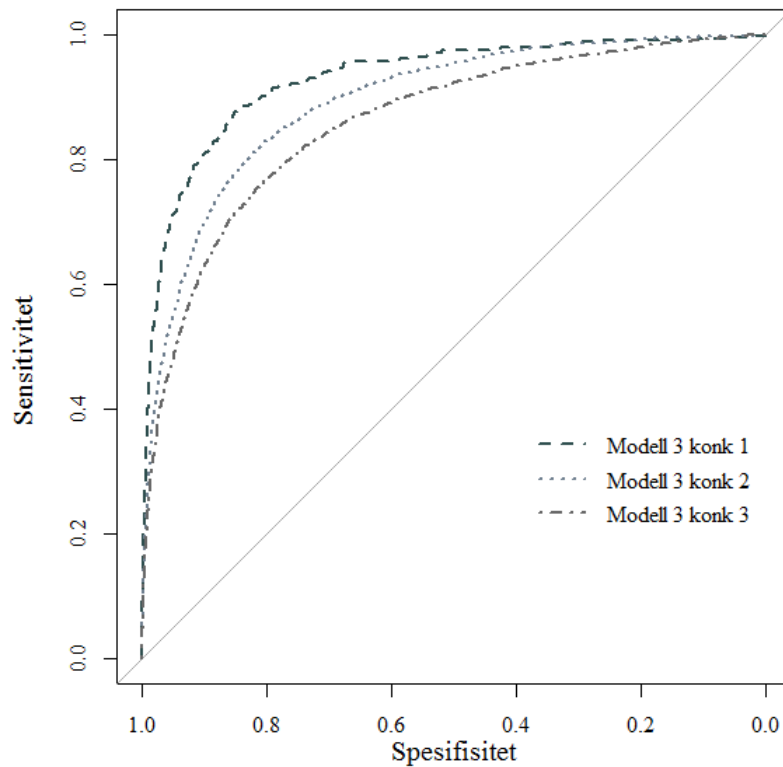
Ettersom modell 3 viser seg å være signifikant bedre enn de andre modellene, velger vi å analysere hvordan modellen presterer ved ulike konkurshorisonter. Begrepet konkurshorisont referer til tidshorizonten modellen prøver å predikere. En ett års konkurshorisont brukes for å si noe om sannsynligheten for at et selskap går konkurs innen ett år etter regnskapsåret. En modell med to års konkurshorisont vil tilsvarende si noe om sannsynligheten for at et selskap går konkurs i innen to år etter regnskapsåret. En modell med kort konkurshorisont vil typisk ha best prediksjonsevne, men vil ikke fange opp selskap som vil gå konkurs på lengre sikt. En modell med lang konkurshorisont vil derimot være bedre på å fange opp tidlige farer, men dette går på bekostning av treffsikkerheten. Hvilken konkurshorisont som er aktuell vil variere for ulike formål, og vi velger derfor å estimere modell 3 med flere konkurshorisonter. Vi estimerer modellen med ett, to og tre års konkurshorisont.

Tabell 6.5 – Sammenligning av AUC-verdier for ulike konkurshorisonter med modell 3

Konkurshorisont	AUC	95% Konfidensintervall	Signifikans
3 år	0.8589	[0.8515 - 0.8664]	-
2 år	0.8928	[0.885 - 0.9006]	T
1 år	0.9284	[0.9144 - 0.9424]	TT

Signifikans representerer om AUC er signifikant forskjellig fra modellen ovenfor seg. T representer forskjellig på 5% signifikansnivå.

Figur 6.5 – ROC-kurver for modell 3 med ulike konkurshorisonter



Som forventet ser vi i tabell 6.5 at kortere konkurshorisonter har signifikant bedre prediksjonsevne. Figur 6.5 viser ROC-kurvene til modell 3 med ulike konkurshorisonter.

Tabell 6.6 – Sammenligning av AUC-verdier for modell 1 og modell 3 med ulike konkurshorisonter

Konkurshorizont	Modell 1	Modell 3	p-verdi
1 år	0.8687	0.9283	7.587e-07
2 år	0.8563	0.8928	1.409e-09
3 år	0.8336	0.8589	2.967e-06

P-verdien er fra DeLongs test for to ROC-kurver, hvor alternativhypotesen er at den sanne differansen ikke er lik 0.

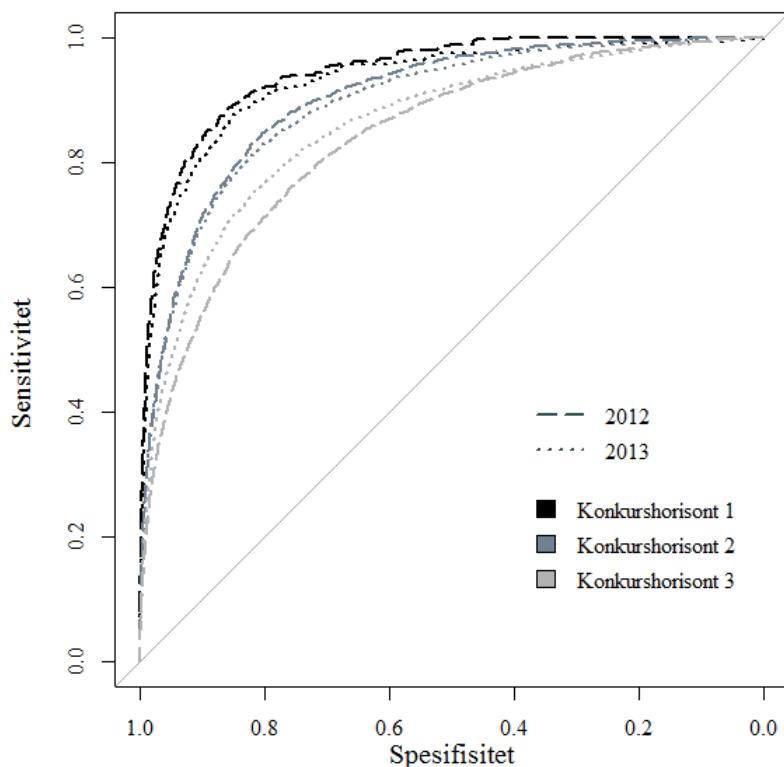
Vi ønsker å analysere om termindata også forbedrer eksisterende modeller med lengre konkurshorisonter. Tabell 6.6 sammenligner AUC-verdiene til modell 1 og modell 3 for samtlige konkurshorisonter. Vi ser at modell 3 er signifikant bedre enn modell 1 for samtlige konkurshorisonter. Modell 3 er betydelig bedre enn modell 1 med en konkurshorizont på ett år, men AUC-verdien avtar raskere for modell 3 enn for modell 1 når konkurshori-

sonten øker. For eksempel faller modell 3 med cirka 3.5 prosentpoeng fra konk1 til konk2, mens modell 1 kun faller med 1.2 prosentpoeng. Dette indikerer at termindatas relative prediksjonsevne avtar når konkurshorizonten øker.

6.7. Validering og robusthet

Vi ønsker å undersøke om modellen er robust over tid. En prediksjonsmodells prestasjon kan være svært sensitiv for hvilke datasett som blir brukt for validering (Berg, 2007). For å bedre sikre oss mot å ikke validere modellen mot et ikke-representativt år, validerer vi modellene mot to ulike out-of-time testsett, et for alle selskap i 2012 og et for alle selskap i 2013. Hvis man ønsker å beholde modellene over en periode, vil det være av interesse å se om prediksjonsevnen faller over tid. Spørsmålet blir da om modellene presterer dårligere når de blir validert mot selskapsdata fra 2013 enn mot selskapsdata fra 2012.

Figur 6.6 – ROC-kurver for modell 3 mot valideringssett fra 2012 og 2013



Figur 6.6 viser ROC-kurvene for modell 3 med konkurshorizont på ett, to og tre år når den valideres mot 2012 og 2013. Det ser ut til at prediksjonsevnen er noe bedre når modellen

valideres mot 2012 både med ett og to års konkurshorisont. Med en konkurshorisont på tre år ser modellen ut til å prestere bedre når den valideres mot 2013. I tabell 6.7 ser vi at det kun er med en konkurshorisont på tre år at modell 3 presterer signifikant forskjellig når den valideres mot 2012 og 2013. P-verdiene for ett og to års konkurshorisont viser at disse ikke er forskjellige på 5% signifikansnivå. Resultatene indikerer at modellen er robust for konkurshorisonter på ett og to år, mens den er mindre robust med en konkurshorisont på tre år.

Tabell 6.7 – Sammenligning av prediksjonsevnen til modell 3 i 2012 og 2013 for ulike konkurshorisonter

Konkurshorisont	2012	2013	p-verdi
1 år	0.9440 [0.9326 - 0.9553]	0.9284 [0.9144 - 0.9424]	0.08958
2 år	0.9008 [0.8934 - 0.9082]	0.8928 [0.885 - 0.9006]	0.1419
3 år	0.8381 [0.8304 - 0.8459]	0.8589 [0.8511 - 0.8661]	0.000147

95% konfidensintervall i klammeparanteser. P-verdien er fra DeLongs test for to ROC-kurver, hvor alternativhypotesen er at den sanne differansen ikke er lik 0.

6.8. Parametrisk modell

Samtlige modeller i denne analysen er estimert ved bruk av den ikke-parametriske metoden GAM-logit. I kapittel 4 diskuterer vi hvordan ikke-parametriske modeller typisk gir mer presise prediksjoner på bekostning av at de mangler tolkbare koeffisienter for forklaringsvariablene. Å estimere en modell ved bruk av GAM-logit krever dessuten mye mer datakraft enn estimering ved bruk av en parametrisk metode. På bakgrunn av dette ønsker vi å undersøke om de nye terminvariablene også bedrer konkursprediksjonsmodeller som tar i bruk parametrisk logistisk metode.

Sammenligner man resultatene fra den ikke-parametriske metoden i tabell 6.6 med resultatene fra den parametriske metoden i tabell 6.8, får vi bekreftet at den ikke-parametriske metoden, GAM-logit, presterer vesentlig bedre. I likhet med GAM-logit har parametrisk logistisk metode signifikant bedre prediksjonsevne med modell 3 enn med modell 1 for samtlige konkurshorisonter. Dette tilsier at de nye terminvariablene forbedrer modellens

Tabell 6.8 – Sammenligning av modell 1 og modell 3 ved å benytte en parametrisk logistisk metode

Konkurshorizont	Modell 1	Modell 3	p-verdi
1 år	0.7514	0.8501	2.2e-16
2 år	0.7747	0.8234	2.2e-16
3 år	0.7683	0.8057	2.2e-16

P-verdien er fra DeLongs test for to ROC-kurver, hvor alternativhypotesen er at den sanne differansen ikke er lik 0.

prediksjonsevne uavhengig av hvilken metode man anvender.

6.9. Regresjoner for alle modeller

En fullstendig oversikt over samtlige modeller finnes i appendiks. Der vises både ikke-parametriske regresjoner og parametriske regresjoner for modell 1 til 3 med de tre utprøvde konkurshorisonter. ROC-kurver og AUC-verdier for samtlige modeller validert mot både 2012 og 2013 er også oppgitt i appendiks. En oppsummering av hver modell med tilhørende appendiks finnes i tabell 6.9.

Tabell 6.9 – Oversikt over modellene

Modellnavn	Metode	Appendiks
Modell 1	Ikke-parametrisk	A3
Modell 2	Ikke-parametrisk	A4
Modell 3	Ikke-parametrisk	A5
Modell 1	Parametrisk	A6
Modell 2	Parametrisk	A7
Modell 3	Parametrisk	A8

7. Drøfting

Resultatene fra våre analyser viser at modeller som tar i bruk termindata har signifikant bedre prediksjonsevne enn referansemodellen som kun tar i bruk regnskapsdata. Modellen som tar i bruk ferske termindata predikerer signifikant bedre en modellen som tar i bruk termindata fra samme periode som regnskapet. Ved å ta i bruk termindata bedres prediksjonsevnen for samtlige tre konkurshorisonter, men fordelene med termindata er størst ved kortere konkurshorisonter. Vi ser også at med en konkurshorisonnt på ett eller to år er modellene våre robuste over tid. Med en konkurshorisonnt på tre år er modellene noe sensitive for hvilke år de valideres mot. I dette kapitlet drøfter vi hvorfor resultatene er som de er og hvilke implikasjoner dette har.

7.1. Tolkning av resultater

For å predikere konkurser er man avhengige av å oppdage faktorer som leder til konkurs så tidlig som mulig. Regnskapstall som brukes i dagens modeller lider av sjelden og sen rapportering, og blir som regel ikke tilgjengelig før sent året etter det aktuelle regnskapsåret. Nøkkeltall fra regnskapet måler dermed fortiden til et selskap, noe som kan være mindre informativt når man ønsker å predikere fremtiden. Termindata derimot kan potensielt gjøres tilgjengelig seks ganger i året allerede to måneder etter aktuell termin. Hvis man benytter ferske termindata, vil man i større grad få et bilde av hvordan selskapet presterer i dag. Modell 3, som tar i bruk ferske termindata, presterer bedre enn modell 2 som tar i bruk termindata fra året før. Dette indikerer at termindatas hyppige rapportering bidrar til å øke prediksjonsevnen.

Et potensielt problem med eksisterende regnskapsbaserte konkursprediksjonsmodeller er at de kan være offer for regnskapsfleksibilitet. Konservativ eller aggressiv regnskapsføring kan føre til at bokverdier i balansen ikke representerer de sanne underliggende verdiene. Dette er spesielt relevant for oss da man ofte observerer aggressiv regnskapsføring for selskap som er nær konkurs (Ronen og Yaari, 2008). Fleksibilitet i regnskapsregler gjør

det dermed mulig for ledelsen å mislede regnskapsbrukerne om den underliggende økonomiske situasjonen til foretaket. Dette er ikke nødvendigvis ulovlig, men et resultat av de skjønnsmessige vurderingene som inngår i utarbeidelsen av et regnskap. Ulik regnskapspraksis mellom selskap og bransjer kan derfor bidra til å svekke nøkkeltallenes generelle evne til å predikere konkurser. Hillegeist et al. (2004) er også kritisk til regnskapsdatas effektivitet når det gjelder konkursprediksjon. De trekker frem at regnskapet er formulert under et fortsatt-drift-prinsipp, som antar at selskapet ikke går konkurs. Regnskapets evne til å predikere konkurs er derfor begrenset. Forklaringsvariablene fra termindata som mva-restanse, SOFIE-restanse, manglende leveringer og betalbar avgift er i mindre grad preget av subjektive vurderinger og antakelser om fortsatt drift. Fra et slikt perspektiv kan termindata antas å være mer et mer nøytralt og objektivt mål på den underliggende økonomiske situasjonen i selskapet.

Det kan argumenteres for at termindata kan fange opp et foretaks profesjonalitet og seriøsitet i større grad enn hva finansielle nøkkeltall gjør. Skatteetaten er å anse som en vesentlig kreditor for alle aksjeselskap i Norge, og hvordan selskap forholder seg til etaten kan gi et bilde av holdningene i selskapet. Et profesjonelt og seriøst foretak kan antas å ha kontroll på innlevering og betaling av terminoppgaver. De nye tvangsmulktreglene nevnt i kapittel 2 vil i enda større grad være disiplinerende overfor selskap. Bøtene som Skatteetaten nå kan utstede vil antageligvis føre til at termindata vil få høyere prediksjonsevne. Bøter ved ikke levert terminoppgave eller manglende betaling vil føre til at de selskapene som har mulighet til å levere og å betale vil ha et større insentiv til å gjøre dette i tide. Derfor vil det kun være selskap som virkelig sliter som ikke leverer terminoppgaver eller betaler gjelden sin. Dette fører til at informasjon om utestående merverdiavgift og SOFIE, samt andel manglende leveringer antageligvis vil være mer særegent for selskaper som sliter.

Et selskap som er nær konkurs, vil typisk vise tydeligere konkurssignaler enn et selskap som er lengre unna konkurs. Som forventet presterer derfor samtlige modeller bedre desto kortere konkurshorizonten er. Resultatene fra våre analyser viser at modellen som tar i bruk ferske termindata har signifikant bedre prediksjonsevne for samtlige tre konkurshorisonter. Det er likevel interessant å se at termindatas relative evne til å bedre referanse-

modellen er avtakende når konkurshorizonten øker. Dette tilsier at den største fordelene med termindata sammenlignet med regnskapsdata er å fange opp de kritiske faktorene like før konkurs. Det virker naturlig at utestående gjeld til Skatteetaten er en mer kritisk indikator på konkurs innen kort tid enn regnskapsbaserte nøkkeltall som inntjening, soliditet, og størrelse.

Hvor godt en modell presterer kan være sensitivt for hvilken periode modellen blir estimert over og hvilken periode modellen blir validert mot. Modellene våre er estimert med data fra 2008 til 2011 og validert mot data fra 2012 og 2013. Datagrunnlaget inneholder dermed år fra finanskrisen. Enkelte av disse årene kan dermed antas å være atypiske år for konkurser. Makroøkonomiske forhold som ikke er inkludert i modellene vil i slike tilfeller i større grad kunne være forklarende for konkurser enn regnskapstall og termindata. Slike effekter som er spesifikke for år ville kunne redusere prediksjonsmodellenes robusthet over tid ettersom effektene ikke nødvendigvis er tilstede i flere år. I et forsøk på å kompensere for dette har vi estimert modellene med data fra fire år slik at modellene i mindre grad fanger opp støy fra spesifikke år. Av samme grunn velger vi å se på hvordan modellene presterer på et datasett fra to ulike år, 2012 og 2013. Av resultatene våre ser vi at modellene er relativt stabil, og man oppnår relativt lik prediksjonsevne når man tester den mot 2012 og 2013. Det er likevel enkelte tilfeller hvor modellene presterer signifikant forskjellig etter hvilke av de to årene modellene blir testet mot. Dette gjelder spesielt med konkurshorizont på tre år noe som tilsier mindre robuste modeller ved lengre konkurshorisonter.

7.2. Implikasjoner

SEBRA-modellene som Norges Bank og Finanstilsynet benytter i dag oppnår en AUC på mellom 0.86 og 0.88, avhengig av hvilke bransjer modellen valideres mot (Hjelseth og Rakerud, 2016). Vår referansemodell, som er estimert med regnskapsdata, oppnår en AUC på mellom 0.84 og 0.88 avhengig av hvilken konkurshorizont som settes. SEBRA-modellene er ikke direkte sammenlignbare med våre modeller da de definerer konkursdato på en annen måte. Vår referansemodell har som funksjon å være et sammenligningsgrunnlag for våre nye modeller. Det er likevel interessant å se at referansemodellens prediksjonsevne

til en viss grad sammenfaller med SEBRA-modellene. Det kan antas at hvis vår regnskapsbaserte referansemodell bedres ved å inkludere termindata, vil de regnskapsbaserte SEBRA-modellene også bedres ved å inkludere termindata.

En forbedret konkursprediksjonsmodell vil gi Norges Bank og Finanstilsynet mer presise anslag på bankenes kredittrisiko mot foretakssektoren. Ved å aggregere de individuelle konkurssannsynlighetene kan man predikere bankenes fremtidige utlånstap med større presisjon. Bedre konkursprediksjonsmodeller kan også brukes til å estimere konkurssannsynligheter for hvert enkelt selskap. Dette er av interesse for kreditorer, men også for kredittvurderingselskap, eiere, ansatte, kunder og leverandører. Totalt sett vil en mer treffsikker konkursprediksjonsmodell bedre finansmarkedenes evne til å formidle kapital til lavest mulig kostnad.

Når vi skriver denne oppgaven er termindata fra Skatteetaten unntatt offentlighet. Så lenge dette står seg vil det ikke være mulig å benytte konkursprediksjonsmodellene vi har utviklet i denne oppgaven. Det er tydelig fra våre resultater at det ligger verdifull informasjon i termindata som markedet ikke har tilgang til. Funnene i denne oppgaven tilsier at offentliggjøring av termindata vil kunne bedre eksisterende konkursprediksjonsmodeller.

7.3. utfordringer og begrensninger

En utfordring med konkursprediksjon er at faktorene som leder til konkurs kan antas å være de samme faktorene som leder til oppløsning, oppkjøp, sammenslåing eller andre former for omorganisering av selskap. Såkalte distress-oppkjøp er oppkjøp av selskap med finansielle problemer. Selskap som har finansielle vanskeligheter kan også få restrukturert sine gjeldsforpliktelser eller eventuelt oppløses. Dermed vil enkelte selskap som ser ut til å ha høy konkurssannsynlighet unngå konkurs gjennom en form for ekstern hjelp. Dette kan virke som et forstyrrende element i våre modeller og redusere prediksjonsevnen.

I datasettet har vi en variabel for konkursdato som er hentet fra Enhetsregisteret. I vår oppgave definerer vi konkurstidspunktet ved bruk av denne datoen. Enkelte studier velger å sette denne konkursdatoen tidligere på grunn av en antakelse om at selskapet de facto går konkurs før denne datoen. For eksempel velger Bernhardsen (2001) å definere et selskap som konkurs det siste året selskapet har levert årsregnskap, gitt at det faktisk

går konkurs innen tre år. Vårt datasett tar i bruk næringsoppgaven som regnskapsinformasjon, og inneholder dermed ikke informasjon om hvorvidt selskapets årsregnskap er levert. Vi kan derfor ikke benytte oss av samme konkursdefinisjon som SEBRA-modellen.

En utfordring med konkursprediksjon er at når selskap nærmer seg konkurs, leverer selskapet mindre regnskapsinformasjon. Denne praksisen fører til at enkelte selskapsobservasjoner forsvinner fra datasettet like før selskapet går konkurs. Dette, kombinert med at konkurs er en svært sjelden hendelse, resulterer i at vi har vesentlig færre observasjoner av hvordan selskap ser ut like før konkurs enn observasjoner av selskap som ikke går konkurs. Bernhardsen (2001) bekrefter denne problematikken, og vi antar at redusert datakvalitet for selskap som er nær konkurs er et generelt problem knyttet til konkursprediksjon og ikke særegent for vårt datasett.

Modellene våre ser ut til å være svake ved ekstreme verdier på forklaringsvariablene. Dette er en naturlig konsekvens av at vi har få observasjoner med ekstreme verdier. Prediksjonsevnen blir ikke vesentlig påvirket da dette gjelder svært få selskap.

8. Konklusjon

Hovedproblemstillingen for denne oppgaven har vært i hvilken grad termindata fra Skatteetaten kan brukes for å bedre eksisterende regnskapsbaserte konkursprediksjonsmodeller. For det første ønsket vi å undersøke om termindata bedrer eksisterende regnskapsbaserte konkursprediksjonsmodeller hvis den brukes for samme periode som regnskapet. For det andre ønsket vi å utnytte at termindata er tidligere tilgjengelig enn regnskapsdata, og undersøke om dette vil forbedre eksisterende modeller ytterligere.

Resultatene fra våre analyser viser at modellen med termindata fra samme år som regnskapet har signifikant bedre prediksjonsevne enn referansemodellen. Våre analyser viser at modellen med ferske termindata har signifikant bedre prediksjonsevne enn både referansemodellen og modellen med termindata fra samme år som regnskapet.

Våre analyser viser at modellen som inkluderer termindata er signifikant bedre enn referansemodellen for samtlige utprøvde konkurshorisonter. Når konkurshorisonten øker, kan fordelene ved termindata likevel se ut til å avta. Dette indikerer at termindatas evne til å bedre dagens regnskapsbaserte konkursprediksjonsmodeller er størst ved korte konkurshorisonter.

Bibliografi

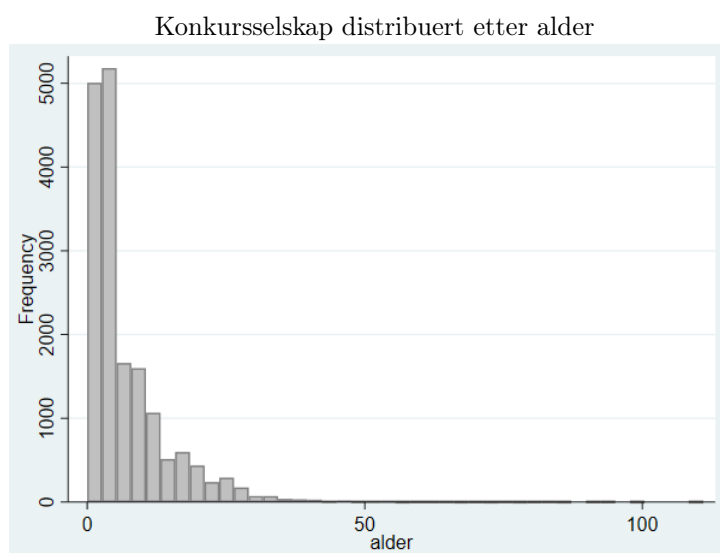
- Alpaydin, E. (2014). *Introduction to Machine Learning*. MIT press.
- Altman, E. (1968). «Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy». I: *The Journal of Finance* 23.4, s. 589–609.
- Altman, E., Iwanicz-Drozdowska, M., Laitinen, E. og Suvas, A. (2014). «Distressed Firm and Bankruptcy Prediction in an International Context: A Review and Empirical Analysis of Altman's Z-Score Model». I: *Journal of International Financial Management Accounting*.
- Beaver, W. H. (1966). «Financial Ratios As Predictors of Failure». I: *Journal of Accounting Research*, s. 71–111.
- Berg, D. (2007). «Bankruptcy Prediction by Generalized Additive Models». I: *Applied Stochastic Models in Business and Industry* 23.2, s. 129–143.
- Bernhardsen, E. (2001). «Working Paper, A Model of Bankruptcy Prediction». I: *Norges Bank Papers*.
- Bernhardsen, E. og Larsen, K. (2007). «Modellering av kredittrisiko i foretakssektoren - videreutvikling av SEBRA-modellen». I: *Penger og Kreditt*, s. 60–66.
- Brækhus, S. (2015). *Konkurs, Store Norske Leksikon*. "<https://snl.no/konkurs>". (Sjekknet 06.11.2017).
- Eklund, T., Larsen, K. og Bernhardsen, E. (2001). «Modell for analyse av kredittrisiko i foretakssektoren». I: *Penger og Kreditt*, s. 109–116.
- Fawcett, T. (2006). «An introduction to ROC analysis». I: *Pattern Recognition Letters* 27.8, s. 861–874.
- Finanstilsynet (2017). *Finansielt Utsyn 2017*.
- Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., Tatham, R. L. et al. (1998). *Multivariate data analysis*. Bd. 5. 3. Prentice Hall Upper Saddle River, NJ.
- Hastie, T. og Tibshirani, R. (1986). «Generalized Additive Models». I: *Statistical Science*, s. 297–318.

- Hillegeist, S. A, Keating, E. K, Cram, D. P og Lundstedt, K. G (2004). «Assessing the Probability of Bankruptcy». I: *Review of accounting studies* 9.1, s. 5–34.
- Hjelseth, I. og Raknerud, A. (2016). «A model of credit risk in the corporate sector based on bankruptcy prediction». I: *Norges Bank*.
- Investopedia (2017). *Business Cycles*. "<http://www.investopedia.com/terms/b/businesscycle.asp>".
- James, G., Witten, D., Hastie, T. og Tibshirani, R. (2014). *An Introduction to Statistical Learning: With Applications in R*.
- Kobro, J. (2012). *Nå kreves mindre aksjekapital og revisor kan velges bort i aksjeselskaper - vilkår og konsekvenser*. <https://www.raeder.no/Kompetanse/Skatt/Artikler/Naa-kreves-mindre-aksjekapital-og-revisor-kan-velges-bort-i-aksjeselskaper-Vilkaar-og-konsekvenser>.
- Konkursloven (1997). *Lov 08. juni 1984 nr. 58 om konkurs*.
- Konkursrådet (2017). *Pengekravs stilling i konkurs*. <http://www.konkursradet.no/pengekravs-stilling-i-konkurs.304777.no.html>.
- Kumar, P. R. og Ravi, V. (2007). «Bankruptcy prediction in banks and firms via statistical and intelligent techniques – A review». I: *European Journal of Operational Research*.
- Larsen, K. (2016). «GAM: The Predictive Modeling Silver Bullet». I: *Multithreaded. Stitch Fix* 30.
- Maricica, M. og Georgeta, V. (2012). «Business failure risk analysis using financial ratios». I: *Procedia-Social and Behavioral Sciences* 62, s. 728–732.
- Nazeran, P. og Dwyer, D. (2015). «Credit Risk Modeling of Public Firms: EDF9». I: *Moody's Analytics*.
- Nelder, J. og Wedderburn, R (1972). *Generalized linear models*. Wiley Online Library.
- Norges Bank (2017). *Finansiell stabilitet: Sårbarhet og Risiko*.
- Ohlson, J. A. (1980). «Financial Ratios and the Probabilistic Prediction of Bankruptcy». I: *Journal of Accounting Research*, s. 109–131.
- Oslo Byfogdembete (2017). *Om konkurs*. <https://www.domstol.no/no/Enkelt-domstol/Oslo-byfogdembete/Konkurs/Om-konkurs/>. (Sjekkset 21.09.2017).
- Petersen, C., Plenborg, T. og Kinserdal, F. (2017). *Financial Statement Analysis*. Fagbokforlaget, s. 371–412.

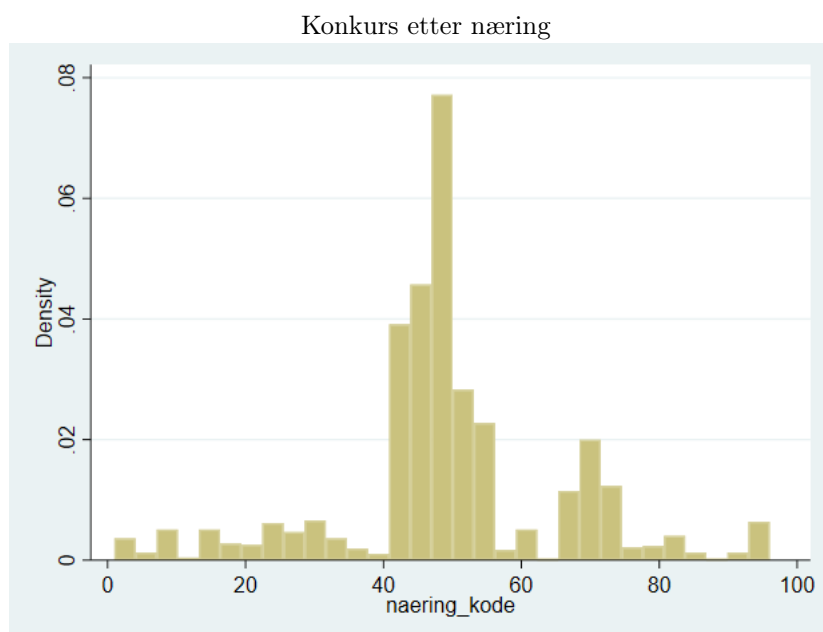
- Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J-C. og Müller, M. (2011). «pROC: an open-source package for R and S+ to analyze and compare ROC curves». I: *BMC Bioinformatics* 12, s. 77.
- Ronen, J. og Yaari, V. (2008). *Earnings Management*. Springer, s. 29.
- Samuel, A. L. (2000). «Some Studies in Machine Learning Using the Game of Checkers». I: *IBM Journal of Research and Development* 44.1.2, s. 206–226.
- Skatteetaten (2017). *Tvangsmulkt innføres for mva-meldingen og andre pliktige meldinger til Skatteetaten*. <http://www.skatteetaten.no/no/0m-skatteetaten/0m-oss/Prosjekter/ny-skatteforvaltningslov/tvangsmulkt/tvangsmulkt-innfores/>.
- Smogeli, P.O. (1987). *Dokumentasjonsnotat SEBRA*.
- Sobehart, J., Keenan, S. og Stein, R. (2000). «Benchmarking Quantitative Default Risk Models: A Validation Methodology». I: *Moody's Investors Service*.
- Statistisk Sentralbyrå (2017a). *Opna konkursar*. <https://www.ssb.no/konkurs>. (Sjekket 06.11.2017).
- (2017b). *Mer enn en fjerdedel av foretakene overlevde i fem år*. <https://www.ssb.no/virksomheter-foretak-og-regnskap/artikler-og-publikasjoner/mer-enn-en-fjerdedel-av-foretakene-overlevde-i-fem-ar>.
- Store Norske Leksikon (2017). *Skatteetaten*. <https://snl.no/Skatteetaten>. (Sjekket 16.11.2017).
- Visma (2017). *Næringsoppgave - Hva er en næringsoppgave?* <https://www.visma.no/eaccounting/regnskapsordbok/n/naeringsoppgave/>. (Sjekket 09.11.2017).
- Wahlstrom, R. og Helland, F. (2016). «Konkursprediksjon for norske selskaper – en analyse ved maskinlæringsteknikker og tradisjonelle statistiske metode». I: *Norges teknisk-naturvitenskapelige universitet - Handelshøyskolen i Trondheim*.
- Wood, S. N. (2000). «Modelling and Smoothing Parameter Estimation with Multiple Quadratic Penalties». I: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 62.2, s. 413–428.
- Wood, S N. (2011). «Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models». I: *Journal of the Royal Statistical Society (B)* 73.1, s. 3–36.

Appendiks

A1. Konkurs etter alder



A2. Konkurs per næring



A3. Modell 1 - Ikke-parametrisk

Modell 1 - Konkurshorison 1

	<i>Parametriske koeffisienter</i>		
	Koeffisient	z	Std.avvik
alder_1	0.427***	5.205	0.082
alder_2	0.502***	5.556	0.090
alder_3	0.289***	2.874	0.100
alder_4	0.453***	4.565	0.099
alder_5	0.228**	1.974	0.115
alder_6	0.367***	3.151	0.116
alder_7	0.272**	2.054	0.132
alder_8	-0.002	-0.013	0.156
Konstantledd	-6.863***	-99.442	0.069
<i>Tilnærmet signifikans av ikke-parametriske funksjoner</i>			
	Edf	Ref.df	Chi.sq
s(inntjening)	8.245***	8.780	379.063
s(soliditet)	8.119***	8.634	384.2
s(likviditet)	7.724***	8.413	144.237
s(størrelse)	2.125	2.712	5.393
s(arbeidskapital)	3.352*	4.420	10.384
s(bransjesnitt inntjening)	5.233***	5.983	36.921
s(bransje stdavvik inntjening)	3.542**	4.410	13.532
s(bransjesnitt soliditet)	1.05	1.105	1.491
Observasjoner		356,505	
Justert R ²		0.020	
Deviance explained		18.9%	
Log Likelihood		-7,843.298	
AUC 2012		0.8785	
AUC 2013		0.8687	

Note:

*p<0.1; **p<0.05; ***p<0.01

Edf: Estimerte frihetsgrader, høye verdier indikerer komplekse funksjoner.

Modell 1 - Konkurshorisont 2

	<i>Parametriske koeffisienter</i>		
	Koeffisient	z	Std.avvik
alder_1	0.7589***	19.084	0.04
alder_2	0.6529***	14.250	0.0458
alder_3	0.4408***	8.832	0.05
alder_4	0.3918***	7.385	0.053
alder_5	0.3526***	6.170	0.0571
alder_6	0.4001***	6.676	0.06
alder_7	0.2400***	3.474	0.07
alder_8	0.1537*	2.059	0.0746
Konstantledd	-5.2491***	-165.802	0.032
<i>Tilnærmet signifikans av ikke-parametriske funksjoner</i>			
	Edf	Ref.df	Chi.sq
s(inntjening)	8.694***	8.962	886.31
s(soliditet)	8.601***	8.913	1441.12
s(likviditet)	8.5724***	8.910	573.56
s(størrelse)	4.058***	5.062	20.71
s(arbeidskapital)	6.305***	7.551	44.88
s(bransjesnitt inntjening)	7.798***	8.267	155.87
s(bransje stdavvik inntjening)	7.720***	8.508	91.16
s(bransjesnitt soliditet)	8.301	8.516	38.42
Observasjoner		356,505	
Justert R ²		0.0572	
Deviance explained		19.6%	
Log Likelihood		-25,375.770	
AUC 2012		0.8486	
AUC 2013		0.8564	

Note:

*p<0.1; **p<0.05; ***p<0.01

Edf: Estimerte frihetsgrader, høye verdier indikerer komplekse funksjoner.

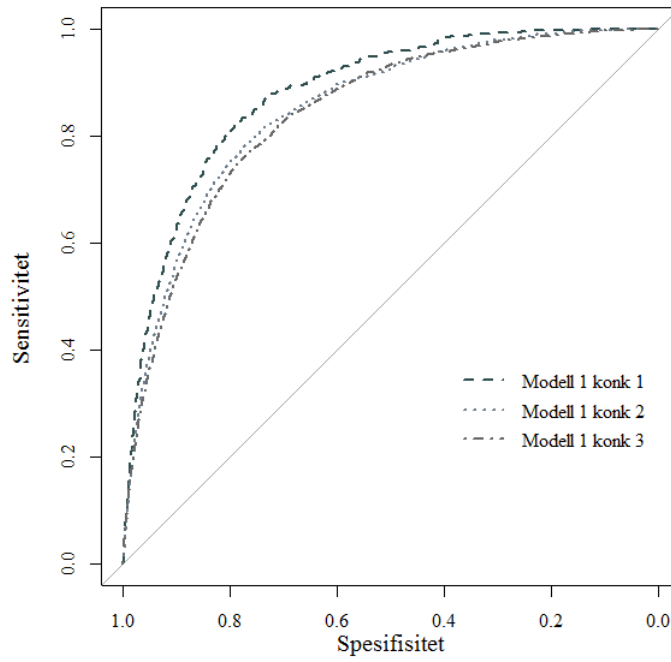
Modell 1 - Konkurshorisont 3

	<i>Parametriske koeffisienter</i>		
	Koeffisient	z	Std.avvik
alder_1	0.8224***	26.634	0.031
alder_2	0.6152***	16.838	0.0365
alder_3	0.4525***	11.637	0.0389
alder_4	0.3836***	9.260	0.0414
alder_5	0.3351**	7.509	0.0445
alder_6	0.3405***	7.167	0.0475
alder_7	0.2742**	5.202	0.0527
alder_8	0.1038**	1.773	0.0586
Konstantledd	-4.5625***	-197.050	0.0232
<i>Tilnærmet signifikans av ikke-parametriske funksjoner</i>			
	Edf	Ref.df	Chi.sq
s(inntjening)	8.637***	8.950	1222.63
s(soliditet)	8.526***	8.887	2085.56
s(likviditet)	8.496***	8.902	936.60
s(størrelse)	4.571***	5.649	73.32
s(arbeidskapital)	8.109***	7.784	85.75
s(bransjesnitt inntjening)	8.077***	8.523	283.24
s(bransje stdavvik inntjening)	7.370***	8.260	138.93
s(bransjesnitt soliditet)	8.319***	8.545	67.11
Observasjoner		356,505	
Justert R ²		0.075	
Deviance explained		18.8%	
Log Likelihood		-38,887.340	
AUC 2012		0.8407	
AUC 2013		0.8336	

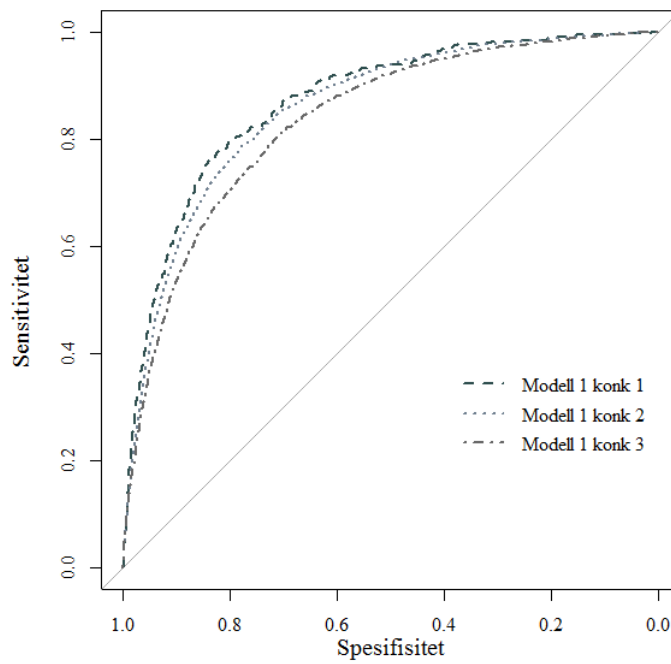
Note:

*p<0.1; **p<0.05; ***p<0.01

Edf: Estimerte frihetsgrader, høye verdier indikerer komplekse funksjoner.



Tre konkurshorisonter for modell 1 validert mot 2012



Tre konkurshorisonter for modell 1 validert mot 2013

A4. Modell 2 - Ikke-parametrisk

Modell 2 - Konkurshorison 1

	<i>Parametriske koeffisienter</i>		
	Koeffisient	z	Std.avvik
alder_1	0.4372***	5.277	0.0829
alder_2	0.5022***	5.502	0.0913
alder_3	0.3130**	3.097	0.1011
alder_4	0.4759***	4.767	0.9992
alder_5	0.2340*	2.016	0.1160
alder_6	0.3799**	3.242	0.1171
alder_7	0.2722*	2.044	0.1332
alder_8	-0.0058	-0.037	0.1569
Konstantledd	-6.9320***	-98.235	0.0706
	<i>Tilnærmet signifikans av ikke-parametriske funksjoner</i>		
	Edf	Ref.df	Chi.sq
s(inntjening)	8.375***	8.845	371.74
s(soliditet)	8.129***	8.643	313.35
s(likviditet)	8.705***	8.943	121.93
s(størrelse)	2.173**	2.776	11.26
s(arbeidskapital)	3.794*	4.984	12.75
s(bransjesnitt inntjening)	5.256***	6.020	45.16
s(bransje stdavvik inntjening)	3.446*	4.291	12.26
s(bransjesnitt soliditet)	1.001	1.003	1.31
s(andel mangler)	1.029	1.058	0.40
s(total mva-restanse)	7.418***	7.964	307.46
s(total SOFIE-restanse)	2.238**	2.754	15.39
s(betalbar avgift)	8.090	8.764	11.97
Observasjoner		356,054	
Justert R ²		0.0249	
Deviance explained		21.1%	
Log Likelihood		-7,645.560	
AUC 2012		0.8944	
AUC 2013		0.8796	

Note:

*p<0.1; **p<0.05; ***p<0.01

Edf: Estimerte frihetsgrader, høye verdier indikerer komplekse funksjoner.

Modell 2 - Konkurshorizont 2

<i>Parametriske koeffisienter</i>			
	Koeffisient	z	Std.avvik
alder_1	0.7856***	19.367	0.041
alder_2	0.6679***	14.298	0.047
alder_3	0.4576***	9.025	0.051
alder_4	0.4056***	7.525	0.054
alder_5	0.3531***	6.085	0.058
alder_6	0.4124***	6.777	0.061
alder_7	0.2409***	3.439	0.07
alder_8	0.1525*	2.017	0.076
Konstantledd	-5.3393***	-164.925	0.032
<i>Tilnærmet signifikans av ikke-parametriske funksjoner</i>			
	Edf	Ref.df	Chi.sq
s(inntjening)	8.782***	8.981	844.76
s(soliditet)	8.533***	8.887	1157.09
s(likviditet)	8.842***	8.977	442.37
s(størrelse)	5.553	6.683	9.34
s(arbeidskapital)	5.429***	6.747	53.53
s(bransjesnitt inntjening)	7.807***	8.282	159.79
s(bransje stdavvik inntjening)	5.294***	6.362	77.24
s(bransjesnitt soliditet)	4.646*	5.495	13.75
s(andel mangler)	1.607	1.960	1.72
s(total mva-restanse)	8.963***	8.999	1199.81
s(total SOFIE-restanse)	7.899***	8.528	152.65
s(betalbar avgift)	5.104	6.211	8.91
Observasjoner		356,054	
Justert R ²		0.074	
Deviance explained		22.6%	
Log Likelihood		-24,431.500	
AUC 2012		0.872	
AUC 2013		0.8746	

Note:

*p<0.1; **p<0.05; ***p<0.01

Edf: Estimerte frihetsgrader, høye verdier indikerer komplekse funksjoner.

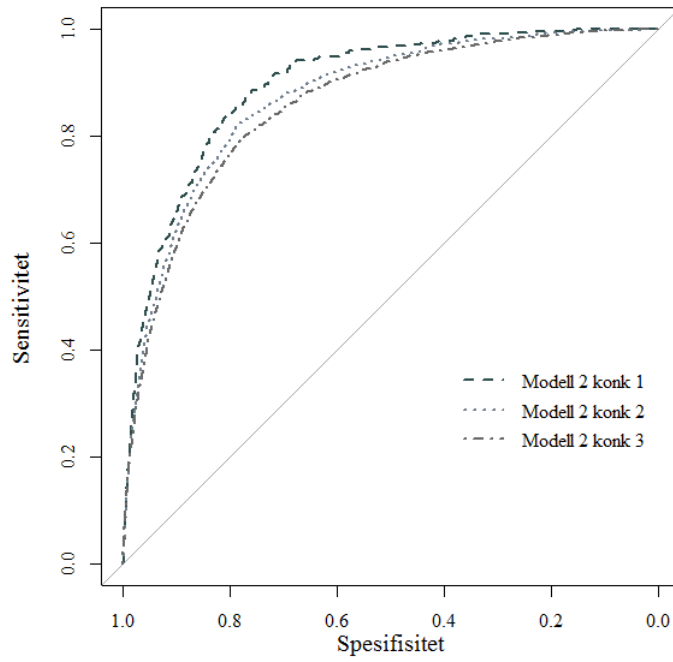
Modell 2 - Konkurshorizont 3

<i>Parametriske koeffisienter</i>			
	Koeffisient	z	Std.avvik
alder_1	0.8496***	26.916	0.0316
alder_2	0.6292***	16.855	0.0373
alder_3	0.4650**	11.729	0.0397
alder_4	0.3955***	9.381	0.0422
alder_5	0.3373***	7.420	0.3373
alder_6	0.3533***	7.310	0.0483
alder_7	0.2779***	5.183	0.0536
alder_8	0.0980*	1.649	0.0595
Konstantledd	-4619***	-193.813	0.0239
<i>Tilnærmet signifikans av ikke-parametriske funksjoner</i>			
	Edf	Ref.df	Chi.sq
s(inntjening)	8.705***	8.967	1182.24
s(soliditet)	8.392***	8.827	1665.523
s(likviditet)	8.764***	8.966	690.13
s(størrelse)	8.648**	8.932	21.26
s(arbeidskapital)	7.405***	8.409	93.38
s(bransjesnitt inntjening)	8.043***	8.451	293.38
s(bransje stdavvik inntjening)	6.631***	7.648	127.71
s(bransjesnitt soliditet)	7.919***	8.081	35.12
s(andel mangler)	1.243*	1.441	4.19
s(total mva-restanse)	8.943***	8.997	1780.16
s(total SOFIE-restanse)	8.824***	8.990	230.76
s(betalbar avgift)	7.259***	8.224	26.39
Observasjoner		356,054	
Justert R ²		0.0973	
Deviance explained		21.1%	
Log Likelihood		-37,471.850	
AUC 2012		0.8588	
AUC 2013		0.8557	

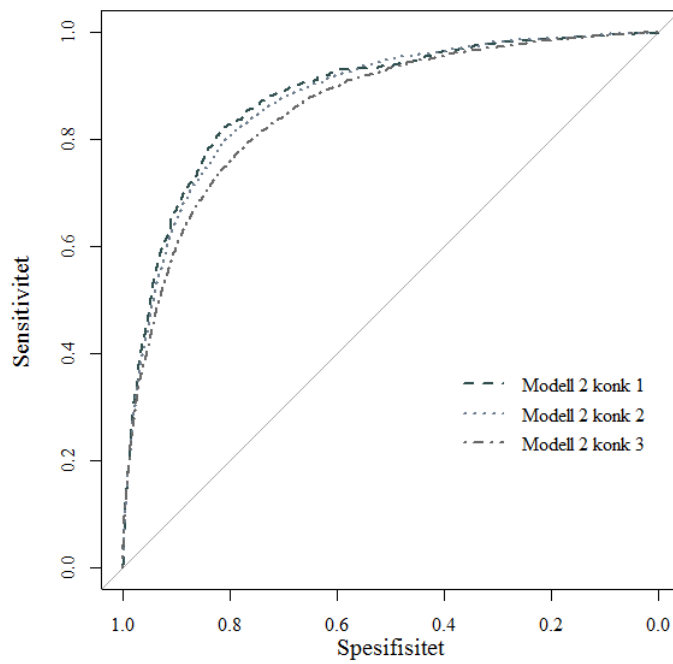
Note:

*p<0.1; **p<0.05; ***p<0.01

Edf: Estimerte frihetsgrader, høye verdier indikerer komplekse funksjoner.



Tre konkurshorisonter for modell 2 validert mot 2012



Tre konkurshorisonter for modell 2 validert mot 2013

A5. Modell 3 - Ikke paramterisk

Modell 3 - Konkurshorison 1

	<i>Parametriske koeffisienter</i>		
	Koeffisient	z	Std.avvik
alder_1	0.7794***	7.590	0.103
alder_2	0.3052***	3.215	0.095
alder_3	0.5767***	6.419	0.089
alder_4	0.4165***	4.205	0.099
alder_5	0.4836***	4.713	0.103
alder_6	0.3325***	2.861	0.116
alder_7	0.4517***	3.694	0.122
alder_8	0.2639*	1.928	0.137
Konstantledd	-7.26***	-57.417	0.129
<i>Tilnærmet signifikans av ikke-parametriske funksjoner</i>			
	Edf	Ref.df	Chi.sq
s(inntjening)	8.847***	8.977	426.47
s(soliditet)	8.339***	8.785	206.53
s(likviditet)	7.864***	8.527	147.52
s(størrelse)	4.684***	5.737	49.90
s(arbeidskapital)	5.263***	6.546	46.57
s(bransjesnitt inntjening)	2.214	2.776	2.05
s(bransje stdavvik inntjenig)	1.002**	1.005	3.89
s(bransjesnitt soliditet)	1.868***	2.203	11.42
s(andel mangler)	5.580	6.238	212.78
s(total mva-restanse)	8.967***	8.999	1415.26
s(total SOFIE-restanse)	7.816***	8.218	65.87
s(betalbar avigft)	7.897	8.646	265.21
Observasjoner		354,666	
Justert R ²		0.072	
Deviance explained		30.1%	
Log Likelihood		-7,325.086	
AUC 2012		0.944	
AUC 2013		0.9284	

Note:

*p<0.1; **p<0.05; ***p<0.01

Edf: Estimerte frihetsgrader, høye verdier indikerer komplekse funksjoner.

Modell 3 - Konkurshorisont 2

<i>Parametriske koeffisienter</i>			
	Koeffisient	z	Std.avvik
alder_1	1.0363***	21.152	0.048
alder_2	0.6498***	14.408	0.045
alder_3	0.6528***	14.407	0.045
alder_4	0.5545***	11.436	0.048
alder_5	0.4314***	8.083	0.053
alder_6	0.3755***	6.460	0.058
alder_7	0.4358***	7.014	0.062
alder_8	0.2729***	3.984	0.068
Konstantledd	-5.258***	-177.709	0.029
<i>Tilnærmet signifikans av ikke-parametriske funksjoner</i>			
	Edf	Ref.df	Chi.sq
s(inntjening)	8.627***	8.945	965.23
s(soliditet)	8.607***	8.924	1067.26
s(likviditet)	8.662***	8.943	494.67
s(størrelse)	4.695**	5.769	19.20
s(arbeidskapital)	6.163***	7.430	120.56
s(bransjesnitt inntjening)	1.001	1.002	0.21
s(bransje stdavvik inntjening)	8.703***	8.967	175.43
s(bransjesnitt soliditet)	2.209***	2.679	20.55
s(andel mangler)	4.260***	4.905	76.18
s(total mva-restanse)	8.925***	8.997	3217.84
s(total SOFIE-restanse)	8.724***	8.969	294.95
s(betalbar avgift)	7.917***	8.667	332.48
Observasjoner		354,666	
Justert R ²		0.118	
Deviance explained		26.5%	
Log Likelihood		-25,173.600	
AUC 2012		0.9008	
AUC 2013		0.8928	

Note:

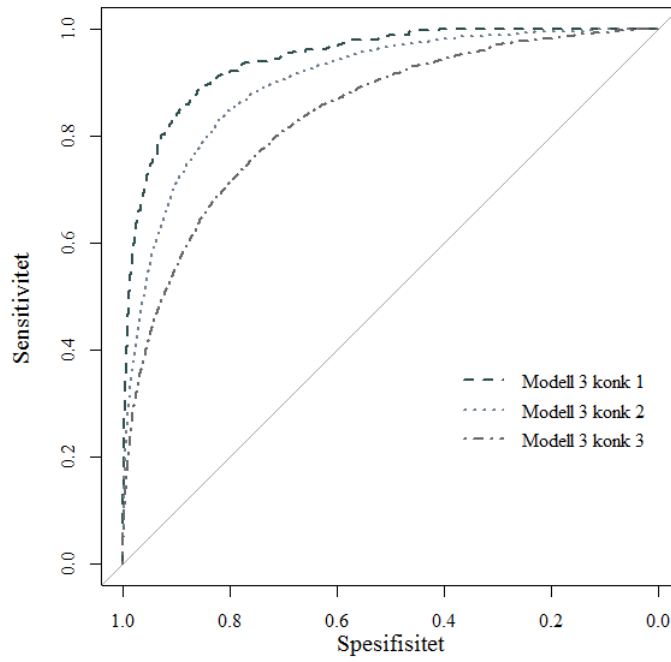
*p<0.1; **p<0.05; ***p<0.01

Edf: Estimerte frihetsgrader, høye verdier indikerer komplekse funksjoner.

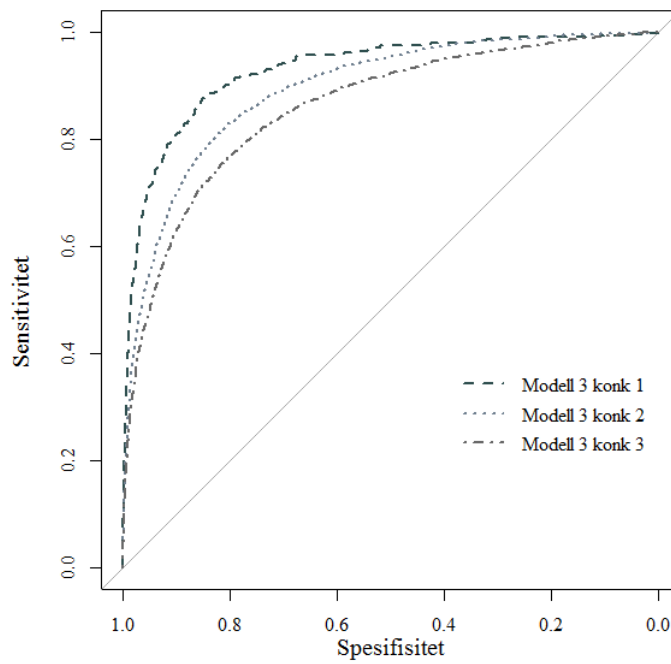
Modell 3 - Konkurshorizont 3

<i>Parametriske koeffisienter</i>			
	Koeffisient	z	Std.avvik
alder_1	1.0498***	27.087	0.039
alder_2	0.7235***	20.451	0.035
alder_3	0.6471***	17.989	0.036
alder_4	0.5282***	13.718	0.039
alder_5	0.4525***	10.859	0.042
alder_6	0.3888***	8.582	0.045
alder_7	0.3769***	7.604	0.050
alder_8	0.2554***	4.741	0.054
Konstantledd	-4.5593***	-209.759	0.022
<i>Tilnærmet signifikans av ikke-parametriske funksjoner</i>			
	Edf	Ref.df	Chi.sq
s(inntjening)	8.715***	8.972	1363.13
s(soliditet)	8.702***	8.957	1582.58
s(likviditet)	8.759***	8.969	714.50
s(størrelse)	4.689**	5.773	18.42
s(arbeidskapital)	6.560***	7.796	145.32
s(bransjesnitt inntjening)	1.024	1.047	0.78
s(bransje stdavvik inntjening)	8.905***	8.996	268.15
s(bransjesnitt soliditet)	8.840***	8.986	95.66
s(andel mangler)	3.924***	4.572	124.77
s(total mva-restanse)	8.938***	8.998	4080.15
s(total SOFIE-restanse)	8.647***	8.953	419.49
s(betalbar avgift)	6.467***	7.545	250.48
Observasjoner		354,666	
Justert R ²		0.134	
Deviance explained		24.6%	
Log Likelihood		-38,578.290	
AUC 2012		0.8381	
AUC 2013		0.8589	

Note: *p<0.1; **p<0.05; ***p<0.01
 Edf: Estimerte frihetsgrader, høye verdier indikerer komplekse funksjoner.



Tre konkurshorisonter for modell 3 validert mot 2012



Tre konkurshorisonter for modell 3 validert mot 2013

A6. Modell 1 - Parametrisk

Modell 1 - alle konkurshorisonter

	<i>Avhengig variabel:</i>		
	konk1	konk2	konk3
	(1)	(2)	(3)
inntjening	-0.344*** (0.039)	-0.579*** (0.026)	-0.643*** (0.024)
soliditet	-0.016* (0.009)	-0.010 (0.008)	-0.042*** (0.015)
likviditet	0.337*** (0.077)	0.579*** (0.042)	0.653*** (0.035)
størrelse	-0.197*** (0.023)	-0.199*** (0.011)	-0.188*** (0.009)
arbeidskapital	0.230*** (0.076)	0.140*** (0.042)	0.132*** (0.034)
bransjesnitt inntjening	0.0003 (0.004)	-0.002 (0.001)	-0.001 (0.001)
bransje stdavvik inntjening	-0.0003** (0.0001)	-0.0003*** (0.0001)	-0.0003*** (0.00004)
bransjesnitt soliditet	-0.003*** (0.001)	-0.002*** (0.0004)	-0.002*** (0.0003)
alder_1	1.217*** (0.095)	1.370*** (0.044)	1.352*** (0.035)
alder_2	0.932*** (0.089)	1.056*** (0.041)	1.039*** (0.032)
alder_3	0.989*** (0.085)	0.936*** (0.042)	0.871*** (0.033)
alder_4	0.773*** (0.094)	0.789*** (0.045)	0.715*** (0.036)
alder_5	0.798*** (0.097)	0.623*** (0.051)	0.600*** (0.039)
alder_6	0.581*** (0.111)	0.533*** (0.055)	0.511*** (0.043)
alder_7	0.695*** (0.116)	0.609*** (0.059)	0.522*** (0.047)
alder_8	0.490*** (0.131)	0.427*** (0.065)	0.379*** (0.051)
Konstantledd	-3.110*** (0.363)	-1.718*** (0.182)	-1.351*** (0.142)
Observasjoner	354,666	354,666	354,666
Log Likelihood	-9,730.240	-31,615.020	-46,952.300
AUC 2012	0.7257	0.7606	0.7683
AUC 2013	0.7257	0.7747	0.7625

Note:

*p<0.1; **p<0.05; ***p<0.01

A7. Modell 2 - Parametrisk

Modell 2 - alle konkurshorisonter

	<i>Dependent variable:</i>		
	konk1 (1)	konk2 (2)	konk3 (3)
inntjening	-0.336*** (0.032)	-0.671*** (0.025)	-0.706*** (0.025)
soliditet	-0.011 (0.012)	0.028** (0.013)	0.0004 (0.011)
likviditet	0.220*** (0.078)	0.417*** (0.047)	0.531*** (0.037)
størrelse	-0.235*** (0.024)	-0.192*** (0.012)	-0.174*** (0.009)
arbeidskapital	0.227*** (0.079)	0.314*** (0.046)	0.327*** (0.037)
bransjesnitt inntjening	-3.774*** (0.558)	-3.803*** (0.300)	-4.204*** (0.248)
bransje stdavv inntjening	-0.824*** (0.186)	-1.028*** (0.097)	-0.954*** (0.074)
bransjesnitt soliditet	-0.469 (0.364)	-0.838*** (0.173)	-0.768*** (0.137)
alder_1	0.963*** (0.081)	1.147*** (0.039)	1.138*** (0.030)
alder_2	0.976*** (0.090)	1.000*** (0.045)	0.899*** (0.036)
alder_3	0.658*** (0.100)	0.716*** (0.049)	0.676*** (0.038)
alder_4	0.752*** (0.099)	0.599*** (0.053)	0.543*** (0.041)
alder_5	0.482*** (0.115)	0.526*** (0.057)	0.466*** (0.044)
alder_6	0.611*** (0.115)	0.556*** (0.059)	0.461*** (0.047)
alder_7	0.502*** (0.132)	0.405*** (0.069)	0.404*** (0.052)
alder_8	0.203 (0.155)	0.300*** (0.074)	0.221*** (0.058)
andel mangler	0.493** (0.218)	0.608*** (0.104)	0.603*** (0.085)
total mva-restanse	4.361*** (0.312)	6.829*** (0.225)	7.872*** (0.209)
total SOFIE-restanse	0.885** (0.391)	2.144*** (0.267)	2.257*** (0.250)
betalbar avgift	0.151 (0.414)	0.633*** (0.160)	0.666*** (0.132)
Konstantledd	-1.870*** (0.393)	-1.103*** (0.200)	-0.857*** (0.155)
Observasjoner	356,054	356,054	356,054
Log Likelihood	-9,063.522	-28,485.040	-42,795.770
AUC 2012	0.7796	0.7973	0.7934
AUC 2013	0.7548	0.7962	0.7913

Note:

*p<0.1; **p<0.05; ***p<0.01

A8. Modell 3 - Parametrisk

Modell 3 - alle konkurshorisonter

	<i>Avhengig variabel:</i>		
	konk1 (1)	konk2 (2)	konk3 (3)
inntjening	-0.336*** (0.034)	-0.557*** (0.026)	-0.642*** (0.024)
soliditet	-0.016 (0.011)	-0.012* (0.007)	-0.021 (0.013)
likviditet	0.328*** (0.072)	0.547*** (0.044)	0.606*** (0.036)
størrelse	-0.146*** (0.023)	-0.146*** (0.011)	-0.139*** (0.009)
arbeidskapital	0.313*** (0.072)	0.308*** (0.044)	0.300*** (0.036)
bransjesnitt inntjening	0.005 (0.004)	-0.001 (0.002)	-0.0002 (0.001)
bransje stdavvik inntjening	-0.0004*** (0.0001)	-0.0004*** (0.0001)	-0.0003*** (0.00005)
bransjesnitt soliditet	-0.003*** (0.001)	-0.002*** (0.0004)	-0.002*** (0.0003)
alder_1	1.091*** (0.099)	1.302*** (0.046)	1.294*** (0.036)
alder_2	0.848*** (0.091)	1.014*** (0.043)	1.011*** (0.033)
alder_3	0.941*** (0.086)	0.911*** (0.043)	0.854*** (0.034)
alder_4	0.740*** (0.095)	0.773*** (0.047)	0.705*** (0.037)
alder_5	0.763*** (0.099)	0.609*** (0.052)	0.592*** (0.040)
alder_6	0.537*** (0.113)	0.514*** (0.056)	0.498*** (0.044)
alder_7	0.644*** (0.119)	0.585*** (0.060)	0.502*** (0.048)
alder_8	0.464*** (0.133)	0.410*** (0.067)	0.366*** (0.052)
andel mangler	1.506*** (0.150)	1.233*** (0.087)	1.195*** (0.074)
total mva-restanse	6.099*** (0.257)	9.064*** (0.211)	9.741*** (0.203)
total SOFIE-restanse	3.576*** (0.431)	5.671*** (0.352)	6.683*** (0.363)
betalbar avgift	-2.232*** (0.297)	-2.243*** (0.201)	-1.400*** (0.164)
Konstantledd	-3.915*** (0.362)	-2.585*** (0.186)	-2.172*** (0.145)
Observasjoner	354,666	354,666	354,666
Log Likelihood	-9,271.164	-29,946.850	-44,742.040
AUC 2012	0.8501	0.8234	0.8057
AUC 2013	0.8048	0.806	0.7967

Note:

*p<0.1; **p<0.05; ***p<0.01