

Ranik Raaen Wahlstrøm

Fredrik Forbord Helland

**Konkursprediksjon for norske selskaper – en analyse ved
maskinlæringsteknikker og tradisjonelle statistiske
metoder**

Bankruptcy prediction for Norwegian companies – an
analysis with machine learning techniques and
conventional statistical methods

MASTEROPPGAVE - Økonomi og administrasjon/siviløkonom

Trondheim, Mai 2016

Hovedprofil: Økonomistyring &
finansiering og investering

Veileder: Frode Kjærland



NTNU har intet ansvar for synspunkter eller innhold i oppgaven.
Framstillingen står utelukkende for studentens regning og ansvar.

Forord

Denne oppgaven er gjennomført som en del av masterstudiet i økonomi og administrasjon ved NTNU våren 2016, og utgjør ett semester med fulltidsstudier. Utredningen er en kombinasjon av de to hovedprofilene *finansiering og investering* og *økonomistyring*.

Konkursprediksjon ble ansett av oss begge som veldig interessant og relevant. I tillegg ble temaet ansett som adekvat for begge våre hovedprofiler, da kunnskaper om finans var viktig for forståelsen av kredittrisiko i utlånsporteføljer, mens økonomistyring var viktig for å tolke årsregnskaper og nøkkeltall. Temaet ga oss muligheten til å kombinere vår interesse for regnskapstall og finansmarkedet. Vi ønsket å benytte ulike metoder til å predikere konkurs, både de mer tradisjonelle og nyere maskinlæringsteknikker, for å se hvilken metode som predikerer best. Dette for komme med et nytt bidrag for bedre prediksjon av konkurs blant norske selskaper.

Vi ønsker å rette en stor takk til vår veileder førsteamanuensis Frode Kjærland for god veiledning og gode diskusjoner. En stor takk rettes også til førsteamanuensis Arild Brandrud Næss som har vært en god sparringspartner gjennom hele prosessen, samt har bistått med sin kompetanse innen maskinlæring, statistikk og bruk av Matlab.

Innholdet i denne oppgaven står for forfatterenes regning.

Trondheim, mai 2016

Ranik Raaen Wahlstrøm

Fredrik Forbord Helland

Sammendrag

Temaet for denne oppgaven er konkursprediksjon. Formålet er å undersøke hvorvidt statistiske modellens evne til å predikere selskapskonkurs kan forbedres. Det blir undersøkt om modellens uavhengige variabler har innvirkning på prediksjonsevne ved å estimere modeller med tre ulike variabelsett: et basert på Norges Banks SEBRA-modell, et basert på Altmans Z-score-modell og et egenutviklet. Det blir også undersøkt om maskinlæringsteknikker kan estimere modeller for konkursprediksjon bedre enn tradisjonelle statistiske metoder. De tradisjonelle metodene er representert ved generell additiv modell, generell lineær modell og diskriminantanalyse; mens nyere maskinlæringsteknikker er representert ved støttevektormaskiner, nevrale nettverk, k -nærmeste naboer og klassifiseringstrær. Modellene i denne oppgaven estimeres og testes ved å benytte samtlige årsregnskaper fra norske aksje- og allmennaksjeselskaper i perioden fra 2005 til 2012, noe som tilsvarer totalt 1 127 850 årsregnskaper fra over 470 000 selskaper. I lys av dette er det satt opp to problemstillinger:

- 1. I hvilken grad kan det implementeres et variabelsett som predikerer konkurs bedre enn variabelsettet benyttet i Norges Banks SEBRA-modell?*
- 2. Ved estimering av Norges Banks SEBRA-modell er den tradisjonelle statistiske metoden GAM (generell additiv modell) benyttet. I hvilken grad er det mulig å forbedre modellen ved å heller benytte maskinlæringsteknikker?*

Oppgaven konkluderer med at det er mulig å implementere en modell som er bedre enn Norges Banks SEBRA-modell, ved å benytte et nytt og egenutviklet variabelsett. I tillegg konkluderer oppgaven med at modeller estimert ved GAM oppnår best prediksjonsevne, men at maskinlæringsteknikker ikke er signifikant dårligere.

Abstract

This master thesis deals with the topic of bankruptcy prediction. The objective is to establish whether the ability of current statistical models to predict bankruptcy for companies can be improved. Three different sets of independent variables are created in order to examine whether a model's independent variables will influence the prediction. These three sets of variables are based on Norges Bank's SEBRA model, Altman's Z-score model and a self-developed model, respectively. The ability of machine learning techniques to estimate bankruptcy prediction models better than conventional statistical methods is also examined. Conventional statistical methods used in this thesis are general additive model, general linear model and discriminant analysis, while recent machine learning techniques are represented by support vector machines, neural networks, k -nearest neighbors and classification trees.

Models in this thesis are estimated and tested by using data from annual reports from every limited liability company in Norway from 2005 to 2012 as a basis. These reports consist of a total of 1 127 850 annual accounts from more than 470 000 limited liability companies. The following two problems are addressed:

- 1. To what extent is it possible to implement a set of variables, which predicts bankruptcy better than a set of variables based on Norges bank's SEBRA model?*
- 2. Norges bank's SEBRA model is estimated by the conventional statistical method GAM (general additive model). To what extent can machine learning techniques improve the existing model?*

The conclusions of the examinations are as follows:

- It is feasible to implement a model with a self-developed set of variables, which is superior to bankruptcy predictions by Norges bank's SEBRA model.
- Models estimated by GAM achieve the best bankruptcy predictions. Note, however, that the bankruptcy predictions of the best machine learning techniques almost equals the estimations by GAM.

Innholdsfortegnelse

1	Innledning.....	1
1.1	Motivasjon.....	1
1.2	Problemstillinger	3
1.3	Oppbygging av oppgaven.....	4
2	Tidligere arbeid	5
2.1	Konkursprediksjon	5
2.2	SEBRA-modellen ved generell additiv modell (GAM).....	6
2.2.1	Variabler i SEBRA-modellen.....	8
2.2.2	SEBRA-basis og SEBRA-utvidet	11
2.2.3	Bruk av SEBRA-modellen i dag	12
2.2.4	Svakheter ved SEBRA-modellen	12
2.3	Beavers univariate modell	13
2.4	Altmans multivariate modeller	16
2.4.1	Den reviderte Altman Z-Score-modellen	18
2.4.2	Zeta-modellen.....	18
2.4.3	Sammenligning av Z-score-modellen og Zeta-modellen	20
2.5	Ohlsons logistiske regresjonsmodell	20
2.6	Maskinlæringsteknikker	22
2.6.1	Klassifiseringstrær.....	23
2.6.1	k -nærmeste naboer.....	24
2.6.2	Støttevektormaskiner.....	24
2.6.3	Nevrale nettverk	26
2.7	Oppsummering av tidligere arbeid	27

3	Forskningsmetode	29
3.1	Valg av variabelsett	29
3.1.1	Avhengig variabel	29
3.1.2	Uavhengige variabler – Variabelsett 1	30
3.1.3	Uavhengige variabler – Variabelsett 2	31
3.1.4	Uavhengige variabler – Variabelsett 3	31
3.2	Valg av metode.....	32
3.3	Evaluering	39
3.3.1	Treffprosent.....	39
3.3.2	Receiver Operating Characteristics (ROC)	40
3.3.3	McNemars test.....	43
3.3.4	Bruk av McNemars test, ROC-verdi og treffprosent i tidligere arbeid	44
3.4	Matlab-kode og repliserbarhet	44
3.5	Datagrunnlag	45
3.6	Definering av konkursselskaper – avhengig variabel.....	46
3.7	Ekskludering av enkeltobservasjoner	46
3.8	Ekskludering av regnskapsår 2013 og 2014.....	47
3.9	Datautvalg og kryssvalidering.....	47
4	Resultater, analyse og drøfting	50
4.1	Resultater fra variabelsett 1 – basert på SEBRA-modellen	50
4.2	Resultater fra variabelsett 2 – basert på Altmans Z-score-modell	52
4.3	Resultater fra variabelsett 3 – egenutviklet	54
4.4	Drøfting	55

5	Konklusjon og implikasjoner	57
5.1	Forslag til videre forskning	58
	Referanser.....	59
	Appendiks 1: Variabelsett 1 – basert på Norges Banks SEBRA-modell.....	63
	Appendiks 2: Variabelsett 2 – basert på Altmans Z-score-modell.....	67
	Appendiks 3: Variabelsett 3	68
	Appendiks 4: Resultater for ulike parametere for metoder	71
	Appendiks 5: Resultater fra variabelsett 1 uten transformering av uavhengige variabler til standardisert verdi	78
	Appendiks 6: McNemars test og <i>p</i> -verdier.....	79
	Appendiks 7: Resultater ved flere ikke-konkursselskaper	81
	Appendiks 8: Antall årsregnskaper i datautvalget.....	85

Figuroversikt

Figur 1: Åpnede konkurser blant aksje- og allmennaksjeselskaper	3
Figur 2: Gjennomsnittlig Zeta-score	19
Figur 3: Eksempel på struktur for klassifiseringstrær (Breiman et al., 1984)	23
Figur 4: Todimensjonale vektorer innenfor to forskjellige klasser.	25
Figur 5: Grunnmodell for et nevralt nettverk (Graupe, 2013).....	26
Figur 6: Parametervalg for nevrale nettverk (NN), basert på variabelsett 1	34
Figur 7: KNN: ROC-verdi som funksjon av antall k	35
Figur 8: Parametervalg for støttevektormaskin (SVM), basert på variabelsett 1	36
Figur 9: Parametervalg for klassifiseringstrær (CT), basert på variabelsett 1	37
Figur 10: Parametervalg for dypt nevralt nettverk skapt av to autoencodere (DNNA)	38
Figur 11: Parametervalg for dypt nevralt nettverk med back-propagation (DNNB)	39
Figur 12: ROC-kurve og ROC-verdi ved tilfeldig klassifisering.	41
Figur 13: Eksempel på ROC-kurve og ROC-verdi	42
Figur 14: Feilklassifiseringsrate for modeller basert på variabelsett 1	51
Figur 15: Feilklassifiseringsrate for modeller basert på variabelsett 2	52
Figur 16: Feilklassifiseringsrate for modeller basert på variabelsett 3	54
Figur 17: Feilklassifiseringsrate for modell basert på variabelsett 3, uten transformering av uavhengige variabler til standardisert verdi.	78
Figur 18: Feilklassifiseringsrate for modell basert på 41 580 selskaper i datautvalget.	82
Figur 19: Feilklassifiseringsrate for modell basert på 114 345 selskaper i datautvalget.	83
Figur 20: Feilklassifiseringsrate for modell basert på 738 045 selskaper i datautvalget.	84

Tabelloversikt

Tabell 1: Risikoinndeling basert på inntjening, likviditetsutvikling og egenkapitalandel	10
Tabell 2: Variabler som inngår i SEBRA-basis og SEBRA-utvidet	12
Tabell 3: Prosentvis oversikt av feilklassifiserte selskaper	15
Tabell 4: Oppsummering av resultatene til Altman (1968).....	18
Tabell 5: Sammenligning av Altmans Zeta-modell og Altmans Z-score-modell	20
Tabell 6: Avhengig variabel	30
Tabell 7: Uavhengige variabler inspirert av Norges Banks utvidede SEBRA-modell	30
Tabell 8: Uavhengige variabler inspirert av Altmans Z-score-modell.....	31
Tabell 9: Uavhengige variabler for variabelsett 3	32
Tabell 10: Metoder valgt til estimering av modeller	33
Tabell 11: Parametere som ble testet for metode KNN.....	34
Tabell 12: Parametere som ble testet for metode SVM.	35
Tabell 13: Parametere som ble testet for metode CT.	36
Tabell 14: Mulige klassifiseringsutfall; hypotetisk og faktisk (Fawcett, 2006).....	40
Tabell 15: Kategoriseringen ved McNemars test	43
Tabell 16: År mellom siste avlagte årsregnskap og år erklært konkurs	47
Tabell 17: ROC-verdier og treffprosjenter for modeller basert på variabelsett 1	51
Tabell 18: ROC-verdier og treffprosjenter for modeller basert på variabelsett 2	53
Tabell 19: ROC-verdier og treffprosjenter for modeller basert på variabelsett 3	55
Tabell 20: ROC-verdier ved ulike antall skjulte lag og noder for metode NN	71
Tabell 21: ROC-verdi ved ulike antall k for metode KNN ved estimering av modell.....	71
Tabell 22: Treffprosjent ved ulike antall k og avstandsberegninger	72
Tabell 23: ROC-verdier ved ulike verdier for kernel scale og box constraint metode SVM...	72

Tabell 24: ROC-verdier ved modeller estimert ved CT	74
Tabell 25: ROC-verdier ved ulike verdier for minste antall løv og minste antall noder for metode CT	75
Tabell 26: ROC-verdier ved ulike antall skjulte lag for autoencoder 1 og 2 for DNNA	76
Tabell 27: ROC-verdier ved ulike antall skjulte lag og noder for metode DNNB.....	77
Tabell 28: ROC-verdi og treffprosent uten transformering av uavhengige variabler til standardisert verdi	78
Tabell 29: <i>p</i> -verdier ved McNemars test for modeller basert på variabelsett 1	79
Tabell 30: <i>p</i> -verdier ved McNemars test for modeller basert på variabelsett 2	79
Tabell 31: <i>p</i> -verdier ved McNemars test for modeller basert på variabelsett 3	79
Tabell 32: <i>p</i> -verdier ved McNemars test for modeller estimert ved GAM	80
Tabell 33: <i>p</i> -verdier ved McNemars test for modeller estimert ved DNNB	80
Tabell 34: ROC-verdi og treffprosent for modell basert på variabelsett 3 og med 41 580 selskaper i datautvalget.	82
Tabell 35: ROC-verdi og treffprosent for modell basert på 114 345 selskaper i datautvalget	83
Tabell 36: ROC-verdi og treffprosent for modell basert på 738 045 selskaper i datautvalget.	84
Tabell 37: Antall årsregnskaper fra aksje- og allmennaksjeselskaper i datautvalget.....	85

1 Innledning

1.1 Motivasjon

«Det å se optimistisk på fremtiden når man utarbeider regnskaper, er ikke et særnorsk fenomen, det har jeg sett ofte også i utenlandske banker. Selv om aksje- eller obligasjonsmarkedet har sett de mørke skyene i lang tid, liker banksjefene å fokusere på at solen er der, den er bare midlertidig skjult bak en sky» (Håland, 2016)

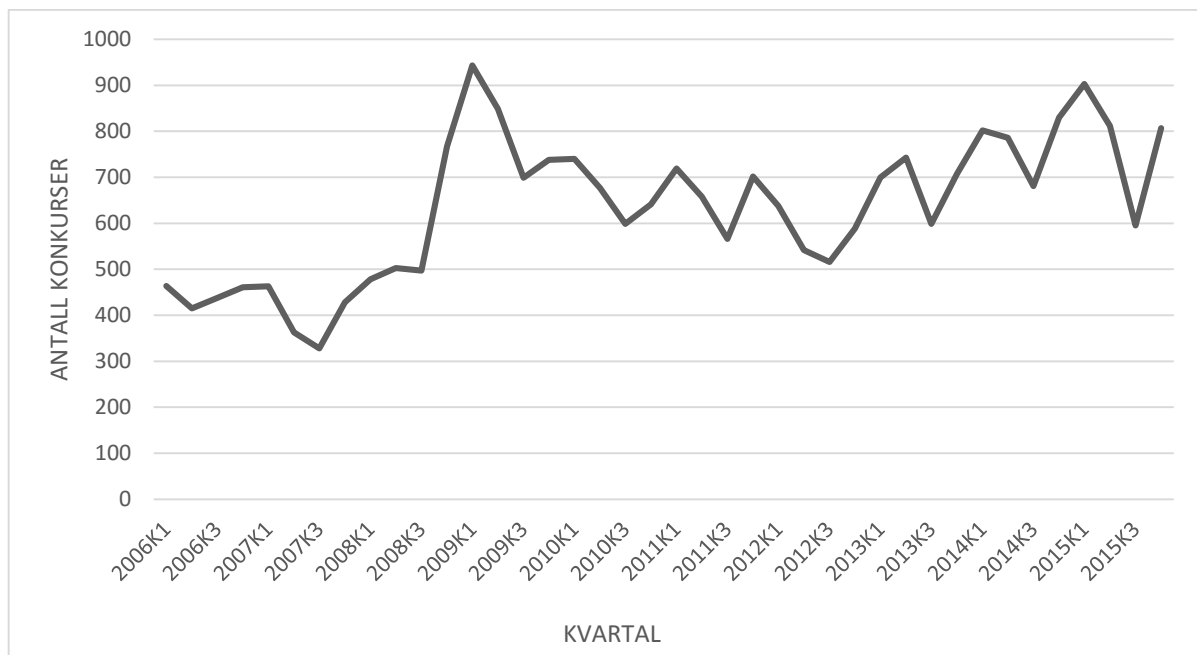
Sitatet er hentet fra en artikkel i Dagens Næringsliv som tar for seg det de kaller for «Banker i en drømmeverden». Å se optimistisk på fremtiden gjelder nødvendigvis ikke bare banker, men også andre bransjer.

Optimismen var også stor i tiden før finanskrisen i 2008 inntraff. Akay og Winters (2011) påpeker at finanskrisen i sterk grad ble påvirket av et for optimistisk syn på eiendomsmarkedet, boliglånsmarkedet og de såkalte *mortgage-backed securities-ene* (MBS) i USA. Dette medførte store lån til investering i eiendom i perioden før finanskrisen. Disse lånene var ofte av typen *subprime*-lån (Kleppe, 2016), altså lån til privatpersoner med lav kredittverdighet. Boligprisene i USA hadde en jevn stigning i tiden før 2006 (Timiraos, 2013), så da det ble en nedgang på 23% i juni samme år kom det som et sjokk for de fleste. Ratingbyråer som Standard & Poor's (S&P), Moody's og Fitch ga den gang flere investeringsobjekter og lån toppvurdering som følge av samme optimistiske markedssyn. Likevel vil nok flere si at mange av disse investeringene fikk en noe ufortjent toppscore, noe som delvis ble bekreftet da flere av topprangeringene gikk konkurs. I ettertid har ratingbyrået S&P foreslått at en toppscore kun skal være forbundet med investeringer med tilnærmet ingen konkursrisiko (Bielecki et al., 2011).

Etter finanskrisen i 2008 har et det meldt seg et økende behov for modeller for å predikere konkurs. Dette er i samsvar med Agarwal og Taffler (2008), som i etterkant av boligprisnedgangen i 2006, rapporterte om økt interesse for vurdering av kredittrisiko grunnet en kraftig vekst i kredittderivatmarkedet i samsvar med økende regulatorisk krav i den internasjonale standarden Basel II, som krever at banker har en kapitaldekning på totalt 8% av sine totale utlån (Martens et al., 2009). For norske banker er kravet om kapitaldekning enda strengere, og fra 1. juli 2016 kreves det en kapitaldekning opp mot 13,5% (Finanstilsynet, 2015a).

Modeller for å predikere konkurs regnes som viktige hjelpemiddel for blant annet banker, investorer, forvaltere, ratingbyråer og kriserammede selskaper (Altman et al., 2014). Chen et al. (2011) hevder at å predikere konkurs blir stadig viktigere i perioder med økende konkursrate. Kostnaden ved å feilklassifisere konkurs rettferdiggjør økt oppmerksomhet for konkursprediksjonsmodeller, da dette kan bidra til å redusere disse kostnadene. I den senere tid har det også blitt tatt i bruk nyere maskinlæringsteknikker til konkursprediksjon (Ravi Kumar og Ravi, 2007, Kirkos, 2015). Varian (2014) mener at slike maskinlæringsteknikker egner seg bedre enn tradisjonelle statistiske metoder på store datasett fordi: (1) størrelsen på datasettet trolig krever et kraftigere datamanipulasjonsverktøy, (2) det kan være at det er for mange variabler med i estimeringen og man må derfor gjøre en form for variabelseleksjon og (3) store datasett kan gi rom for mer fleksible relasjoner enn enkle lineære modeller. Maskinlæringsteknikker har i alle fall innenfor Norges landegrenser fått vesentlige lavere oppmerksomhet enn de eldre statistiske teknikkene for å predikere konkurs, så det er ikke utenkelig at maskinlæringsteknikker fortjener et større fokus.

Det er et kjent fenomen at konkurserisiko er lavere blant større selskaper (Baker og Martin, 2011). Likevel har flere store selskaper, deriblandt større allmennaksjeselskaper, gått konkurs i tiden etter finanskrisen. I den siste tiden kan det blant annet nevnes at Dolphin Group ASA, Cecon ASA, Northland Resources SE og Norse Energy Corp.ASA har gått konkurs (Oslo Børs, 2016). En samlet oversikt over antall konkurser for aksje- og allmennaksjeselskaper for perioden 1. kvartal 2006 til 3. kvartal 2015 følger i figur 1. Figuren viser en kraftig prosentvis økning i antall åpnete konkurser i Norge per kvartal, fra 464 i 1. kvartal 2006 til 807 i 3. kvartal 2015. Trenden er tydelig stigende, noe som indikerer at behovet for økonomistyringsverktøy for å predikere konkurs i sterkeste grad er til stede. Dette for at kreditorers tap kan begrenses og en eventuell ny finansiell krise kan unngås.



Figur 1: Åpnede konkurser blant aksje- og allmennaksjeselskaper kvartalsvis fra 1. kvartal 2006 til 3. kvartal 2015 (SSB, 2015)

1.2 Problemstillinger

I denne oppgaven skiller det mellom modell, variabelsett og metode. En *modell* defineres som en forenkling av virkeligheten, hvor en avhengig variabel blir forklart av en eller flere uavhengige variabler. I alle modeller som omtales i denne oppgaven forteller den avhengige variabelen om et selskap går konkurs eller ikke, mens de uavhengige variablene er nøkkeltall fra selskapets regnskaper. Alle modellene har altså samme avhengig variabel, mens de uavhengige variablene varierer i antall og form. Et *variabelsett* defineres som en samling av nøkkeltall fra årsregnskaper som benyttes som de uavhengige variablene i en gitt modell. *Metoder* er ulike teknikker for å estimere modeller ved hjelp av et datautvalg og et gitt variabelsett.

Denne oppgaven er gjennomført i følgende tre faser:

1. Implementere et variabelsett basert på offentlig tilgjengelige beskrivelser av Norges Banks SEBRA-modell, som er den modellen for konkursprediksjon som benyttes av Norges Bank og Finanstilsynet.
2. Implementere to andre variabelsett, hvor det ene er utarbeidet basert på nøkkeltallene benyttet i Altmans Z-score-modell, mens det andre er tilsvarende som i fase 1, men i tillegg blir det lagt til ytterligere flere nøkkeltall.

3. Estimere modeller basert på de tre implementerte variabelsettene ved hjelp av ti forskjellige statistiske metoder, hvorav fire anses som tradisjonelle og seks anses som maskinlæringsteknikker. Datautvalget som modellene blir estimert på består av årsregnskaper fra alle norske aksje- og allmennaksjeselskaper i perioden fra 2005 til 2012.

Dette leder til følgende to problemstillinger:

1. I hvilken grad kan det implementeres et variabelsett som predikerer konkurs bedre enn variabelsettet benyttet i Norges Banks SEBRA-modell?
2. Ved estimering av Norges Banks SEBRA-modell er den tradisjonelle statistiske metoden GAM (generell additiv modell) benyttet. I hvilken grad er det mulig å forbedre modellen ved å heller benytte maskinlæringsteknikker?

Oppsummert vil denne oppgaven forsøke å finne innfallsvinkler som gir bedre konkursprediksjon enn etablerte modeller, basert på samtlige årsregnskaper fra norske aksje- og allmennaksjeselskaper i perioden fra 2005 til 2012, noe som er totalt 1 127 850 årsregnskaper fra over 470 000 selskaper. I tillegg vil bruken av maskinlæringsteknikker gi et unikt bidrag på metodiske tilnærminger som i liten grad tidligere er behandlet innen konkursprediksjon på norske regnskapsdata.

1.3 Oppbygging av oppgaven

I kapittel 2 av oppgaven presenteres tidligere arbeid som er gjort innenfor både konkursprediksjon og statistiske metoder, deriblant maskinlæringsteknikker. Deretter introduseres forskningsmetoden og empirien anvendt i oppgaven i kapittel 3. Til slutt presenteres resultater og det gis en konklusjon på problemstillingene.

2 Tidligere arbeid

Tre viktige studier om konkursprediksjon er «Bankruptcy Prediction in Banks and Firms via Statistical and Intelligent Techniques – A Review» av Ravi Kumar og Ravi (2007), «A Review of Bankruptcy Prediction Studies: 1930-Present» av Bellovary et al. (2007) og «Assessing Methodologies for Intelligent Bankruptcy Prediction» av Kirkos (2015).

Sistnevnte oversiktsartikkel refererer til den førstnevnte. Referanselistene i disse tre artiklene ga et godt utgangspunkt for å finne andre artikler om temaet.

I artiklene til Ravi Kumar og Ravi (2007) og Bellovary et al. (2007) ble det fremlagt en grundig oversikt over utviklingen av konkursprediksjonsmodeller siden 1930-tallet. Dette ga grunnlaget for omtalene av modellene i denne oppgavens teoridel. I tillegg ble det søkt på Norges Bank sine hjemmesider og Finanstilsynets årlige publikasjon «Finansielt Utsyn» (2013a, 2014, 2015b) for å finne modeller for konkursprediksjon fra Norge. Den mest omtalte modellen i Norge er Norges Banks SEBRA-modell, så det er naturlig at denne blir behandlet videre i denne oppgaven.

I resten av dette kapitlet presenteres tidligere arbeid gjort innenfor konkursprediksjon. En definisjon av konkurs introduseres først, før Norges Banks SEBRA-modell presenteres. Deretter blir flere etablerte konkursprediksjonsmodeller presentert, før kapitlet avsluttes med en gjennomgang av maskinlæringsteknikker for å predikere konkurs.

2.1 Konkursprediksjon

I Lov om gjeldsforhandling og konkurs (1986) § 60 står det at dersom en skyldner er *«insolvent, skal vedkommendes bo tas under konkursbehandling når det begjæres av skyldneren eller en fordringshaver.»* Videre står det i § 61 at *«Skyldneren er insolvent når denne ikke kan oppfylle sine forpliktelser etter hvert som de forfaller, medmindre betalingsudyktigheten må antas å være forbigående.»*

Det står på Konkursrådet (2012) sine hjemmesider: *«For at en skyldner skal anses å være insolvent, må skyldneren være både illikvid og insuffisient.»* Videre står det forklart at illikviditet vil si at skyldner ikke har nok penger til å betale sine forpliktelser ved forfall, og at insuffisient betyr at skyldner heller ikke har nok penger til å betale sine forpliktelser ved salg av alle sine eiendeler.

Siden børskrakket i 1929 har man sett behovet for modeller som kan predikere mislighold og konkurs. Det har siden da blitt utarbeidet ulike typer modeller og det skiller i denne oppgaven mellom univariate modeller, multivariate modeller, logistiske modeller og modeller basert på maskinlæringsteknikker. Analyser av 165 studier gjort av Bellovary et al. (2007) innen konkursprediksjon gir en historisk utvikling som starter med diskriminantanalyser på 60- og 70-tallet, til logistiske modeller og nevralt nettverk på 80- og 90-tallet.

William H. Beaver, Edward I. Altman og James Ohlson er noen av de som står bak de mest anerkjente internasjonale modellene. Her til lands har Norges Bank utviklet en modell kalt SEBRA-modellen, som er basert på norske regnskapsdata. Denne har blitt benyttet av Norges Bank siden 2001 (Bernhardsen og Larsen, 2007) og benyttes fortsatt (Finanstilsynet, 2015b), og vil bli omtalt i det følgende delkapitlet.

2.2 SEBRA-modellen ved generell additiv modell (GAM)

Ved estimering av SEBRA-modellen benytter Bernhardsen og Larsen (2007) en metode som kalles generell additiv modell (GAM), som er en utvidelse av generell lineær modell (GLM). Bakgrunnen for at GAM velges som metode er en studie av Berg (2007) som kommer frem til at GAM er signifikant bedre enn metodene lineær diskriminantanalyse (LDA), forovermatet nevralt nettverk og GLM.

GLM er en metode som forsøker å forklare en avhengig variabel ved hjelp av en funksjon av den lineære kombinasjonen av uavhengige variabler (Gaussier og Yvon, 2013). I denne oppgaven kan den avhengige variabelen kun ha én av to mulige verdier, «konkurs» eller «ikke konkurs», og det er derfor naturlig å benytte en binomisk fordeling. Ved en slik fordeling kan GLM uttrykkes på følgende form (Hastie og Tibshirani, 1995):

$$\log\left(\frac{\mu}{1-\mu}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n \quad (1)$$

Her er $x_1, x_2 \dots x_n$ de uavhengige variablene, $\beta_1, \beta_2 \dots \beta_n$ koeffisientene til de uavhengige variablene, α er funksjonens konstantledd og μ er den estimerte verdien av den binære avhengige variabelen. Sistnevnte blir i denne oppgaven transformert ved hjelp av den logaritmiske funksjonen $\log\left(\frac{\mu}{1-\mu}\right)$.

GAM erstatter GLM sin lineære form $\sum_n \beta_n$ med en additiv form $\sum_j f_j(x_j)$ (Berg, 2007), og kan skrives på følgende form:

$$\log\left(\frac{\mu}{1-\mu}\right) = \alpha + f_1(x_1) + f_2(x_2) + \dots + f_j(x_j) \quad (2)$$

der f_j er en uspesifisert ikke-parametrisk funksjon, det vil si en funksjon som ikke inneholder statistiske parametere. At funksjonene f_j ikke er parametriske bidrar til å gjøre modellen mer fleksibel (Hastie og Tibshirani, 1995). Funksjonen (2) kan omskrives til følgende form, hvor μ blir en funksjon av $x_1, x_2 \dots x_j$, og får en verdi mellom 0 og 1:

$$\mu = \frac{1}{1 + \exp(-[f_1(x_1) + f_2(x_2) + \dots + f_j(x_j)])} \quad (3)$$

Ved testing av modellens prediksjonsevne blir den estimerte verdien, μ , omregnet til verdiene 0 ved «ikke konkurs» eller 1 ved «konkurs». Dette blir gjort ved å bestemme en terskelverdi, hvor μ får verdien 1 om den er over terskelverdien, og 0 om μ er under terskelverdien. Man kan tenke seg at denne terskelverdien blir satt til 0,5. I denne oppgaven blir det imidlertid estimert den terskelverdien som gir flest riktige antall estimeringer blant dataene som modellen trenes på.

SEBRA er en forkortelse for *System for EDB-Basert regnskapsanalyse* (Finanstilsynet, 2013a). SEBRA-modellen er en empirisk regnskapsmodell utviklet av Norges Bank basert på SEBRA-databasen, hvor nøkkeltall for likviditet, inntjening, soliditet, bransje, alder og størrelse er inkludert (Bernhardsen, 2001). Nøkkeltallene vil bli nærmere behandlet i et senere avsnitt.

På det tidspunktet modellen ble estimert besto SEBRA-databasen av 407 125 selskaper fra perioden mellom slutten av 80-tallet og fram til 1999, hvorav 398 689 ikke hadde gått konkurs og 8 436 hadde gått konkurs (Eklund et al., 2001). Det er siden den gang foretatt omfattende testing og validering av modellen som bekrefter at den er relativt god til å predikere mislighold og konkurs ett år frem i tid.

Et av SEBRA-modellens viktigste bruksområder er ved beregning av bankenes forventede tap på utlånsporteføljer, hvor forventet utlånstap finnes ved å multiplisere konkurssannsynligheten med samlet utlån (Eklund et al., 2001). SEBRA-modellen brukes også til framskrivninger og stresstesting av utlånstap for opptil tre år frem i tid, ved at den anvender prognoser basert på SSB sine data (Finanstilsynet, 2013b). Disse stresstestene brukes blant annet av Finanstilsynet i deres analyser av finansmarkedet og i deres tilsyn med bankene (Finanstilsynet, 2015b).

2.2.1 Variabler i SEBRA-modellen

I forrige delkapittel ble SEBRA-modellens bruksområder og datagrunnlag presentert. I dette delkapitlet behandles SEBRA-modellens uavhengige variabler ved at likviditet, inntjening, soliditet, alder, størrelse og bransjekjennetegn og deres underliggende finansielle nøkkeltall presenteres. Variablene er hentet fra Eklund et al. (2001).

Likviditet

Hvor mye penger et selskap trenger for å betjene sine kostnader avhenger av hva selskapet driver med. Man bør derfor være forsiktig med å bruke nøkkeltall knyttet opp mot likviditet. Likevel er det ikke uvanlig at selskapets likvide midler gradvis forsvinner før konkurs.

I Norges Banks SEBRA-modell benyttes variabelen betalingsmidler fratrukket kortsiktige gjeld i forhold til inntekter, da den dekker både kortsiktig gjeld i forhold til inntekter fra drift og andel kontanter i forhold til inntekter fra driften (Bernhardsen, 2001).

Under følger oversikt over alle tre variabler tilhørende likviditet som er valgt inkludert i den opprinnelig SEBRA-modellen.

lik = betalingsmidler – kortsiktig gjeld, i prosent av driftsinntekter

ube = skyldige offentlig avgifter i prosent av total kapital

lev = leverandørgjeld i prosent av total kapital

Inntjening

Inntjening påvirker både likviditeten og soliditeten til et selskap. På lang sikt må et selskap generere en høy nok inntekt, og således bidra til positive resultater slik at forpliktelse og gjeld kan betjenes. Negative resultater over lengre perioder vil påvirke et selskap negativt, spesielt i forhold til soliditet og likviditet. I SEBRA-modellen defineres en årlig kontantinntjening på 20% eller mer av langsiktig gjeld som tilfredsstillende. Dette er basert på en antagelse om at et selskap bruker i gjennomsnitt 5 år på å nedbetale langsiktig gjeld (Larsen og Sæther, 1999). Under inntjening benyttes kun ett nøkkeltall.

tkr = inntjening i prosent av total kapital

Soliditet

Den finansielle risikoen kan knyttes opp mot egenkapitalprosenten et selskap har. Altså hvor mye egenkapital i forhold til total kapital. Jo høyere egenkapitalandel, jo lavere risiko og vice versa. En høyere egenkapitalgrad gjør at selskap kan overleve lengre perioder med svak inntjening. I SEBRA-modellen er egenkapitalprosent på 20% av total kapitalen satt som en tilfredsstillende andel. Dette er i henhold til gjennomsnittsbetraktninger og statistisk testing av den norske selskapssektoren (Larsen og Sæther, 1999). Det er i SEBRA-modellen valgt ut tre nøkkeltall under soliditet. To av disse er dummyvariabler, som innebærer at de kun kan ha én av to mulige verdier.

e_{ka} = bokført verdi av egenkapital som andel av total kapital

t_{aptek} = bokført verdi er mindre enn innskutt egenkapital (dummyvariabel)

div = utbetalt dividende i siste regnskapsår (dummyvariabel)

Alder

Konkursrisikoen er høyest i etableringsfasen. En høyere alder reduserer statistisk sett selskapets risiko for å bli insolvent. Dette samsvarer med Eklund et al. (2001) sine testresultater og studier i andre land. Årsaken til at det er slik er at det ofte tar tid å bygge opp relevant kompetanse på sentrale områder som for eksempel økonomi- og likviditetsstyring, organisering, innkjøp, salg og produksjon. I SEBRA-modellen benytter man en parameter som er representert av dummyvariabelen a_x .

a_x = år siden opprettelse av selskap, $x = 1, \dots, 8$ (dummyvariabler)

Størrelse

Størrelse kan tolkes likt som alder. Jo større selskapet er, desto lavere er risikoen for insolvens. I SEBRA-modellen benytter man logaritmen av total kapitalen til selskapet. Mindre selskaper operer ofte innenfor mindre geografiske områder og har ofte et mindre produktsortiment. De blir dermed mer sårbare for enkelthendelser sammenlignet med selskaper som operer på flere geografiske områder og har et større produktsortiment (Eklund et al., 2001). I SEBRA-modellen benyttes altså variabel som beskrevet her.

$size = \ln(\text{total kapital})$

Bransjekjennetegn

Konkursrisikoen varierer fra bransje til bransje. Bransjer med lav egenkapitalprosent og høy gjennomsnittlig leverandørgjeld er spesielt utsatt. Selskap innenfor disse bransjene blir ofte hardest rammet i dårlige tider. Det kan for eksempel være dagligvarebutikker og restauranter. Bransjespesifikke parametervariabler som er valgt ut i modellen følger under.

$mean_{ek}$ = gjennomsnittlig verdi av variabelen ek_a (Variabel under soliditet)

$mean_{lev}$ = gjennomsnittlig verdi av variabelen lev (variabel under likviditet)

sd_{tkr} = variansen til variabelen tkr (variabel under inntjening)

Ut i fra ulike kombinasjoner av forholdstallene har Larsen og Sæther (1999) i sin artikkel for Norges Bank dannet ulike risikogrupper. Det er her laget et skille etter hvilken inntjeningsgrad, egenkapitalandel og likviditetsutvikling selskapene har. Dette benyttes igjen til å danne et skille mellom de ulike selskapenes risikoprofiler. En grovere inndeling på 8 grupper er laget og følger i tabell 1.

Gruppe 1-2 anses som lav risiko. Gruppe 3-5 anses som moderat risiko, mens gruppe 6 til 8 betraktes som høy risiko. Med andre ord gruppe 1 og 2 anses som de lånene hvor det er lavest sannsynlighet for mislighold og således der risikopåslaget på lån bør være minst.

I senere tid har det blitt utviklet en SEBRA-basis- og en SEBRA-utvidet-modell, som begge er en utvidelse av den generelle SEBRA-modellen (Bernhardsen og Larsen, 2007).

Tabell 1: Risikoinndeling basert på inntjening, likviditetsutvikling og egenkapitalandel (Larsen og Sæther, 1999)

Inntjening	Likviditets- utvikling	Egenkapitalandel		
		Over 20%	0-20%	Negativ
Mer enn 20%	God	Gruppe 1	Gruppe 3	
	Svak	Gruppe 2	Gruppe 4	
0-20%	God	Gruppe 5	Gruppe 6	
	Svak			
Negativ	God	Gruppe 7	Gruppe 8	
	Dårlig			

2.2.2 SEBRA-basis og SEBRA-utvidet

I forrige delkapittel ble den opprinnelige SEBRA-modellen introdusert. I dette delkapitlet presenteres SEBRA-basis og SEBRA-utvidet. SEBRA-basis inkluderer som den generelle modellen de opprinnelige nøkkeltallene for inntjening, soliditet, likviditet og alder. I tillegg inkluderes en dummyvariabel for svekket egenkapital. Et gjennomgående trekk i modellen er at konkurssanssynligheten for større selskaper viser seg å være høyere enn i opprinnelige SEBRA-modellen og i SEBRA-utvidet (Bernhardsen og Larsen, 2007). Årsaken til dette er man i basismodellen i mindre grad vektlegger størrelsesrelaterte variabler. Variabler som tar hensyn til størrelse, bidrar til å trekke ned konkurssanssynlighet for større selskaper. Likevel gir de ulike modellene samme gjennomsnittlige konkurssanssynlighet uansett hvilken variant av SEBRA-modellen som benyttes.

SEBRA-basis er mer treffsikker når det skal beregnes forventet framtidig utlånstap og framtidig utvikling av utlånstap. Forskning tilsier at SEBRA-basis gir gode anslag for bankenes bokførte utlånstap (Bernhardsen og Larsen, 2007). At størrelse på selskapet vektet i mindre grad er en viktig grunn til at modellen er bedre til å predikere utlånstap enn SEBRA-modellen og den utvidede SEBRA-modellen. SEBRA-utvidet er lik basisversjonen, men har i tillegg med variabler som hensyn tar omfang av leverandørgjeld, ubetalte offentlige avgifter og størrelse. Modellen er marginalt bedre til å predikere konkurs. Med andre ord bør det vektlegges hvorvidt om det er konkurs eller mislighold av gjeld som skal predikeres i analysen når valget står mellom basis-modellen og utvidet-modellen. Hvis forventet utlånstap skal beregnes benyttes SEBRA-basis, men hvis det skal predikeres konkurs slik som i denne oppgaven benyttes SEBRA-utvidet.

Datautvalget i SEBRA-modellene består av ikke-finansielle aksjeselskaper med total kapital over en halv million. Variabelen som forklares er definert ved sammenfallet av begivenhetene: «Selskapet slutter å levere regnskap neste år» og «Konkurs åpnes». I om lag 20 prosent av tilfellene åpnes konkurs 3 år etter det siste leverte regnskapet. Det betyr at modellen bare kan testes og re-estimeres på regnskap som ligger 2 til 3 år i forkant av siste tilgjengelige regnskap (Bernhardsen og Larsen, 2007). Tabell 2 viser en oversikt over hvilke faktorer som er med i beregningen av SEBRA-basis og SEBRA-utvidet.

Tabell 2: Variabler som inngår i SEBRA-basis (mørkere tonet felt) og SEBRA-utvidet (hele tabellen) er hentet fra Bernhardsen og Larsen (2007).

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

2.2.3 Bruk av SEBRA-modellen i dag

SEBRA-modellen benyttes av Norges Bank (Bernhardsen og Larsen, 2007) og Finanstilsynet (Syversten et al., 2015). I dag anvendes modellen blant annet til å foreta stikkprøver og som et verktøy for foreta betraktninger i forhold til risiko i bedriftsmarkedsporføljen. SEBRA-modellen anvendes også til å beregne misligholds- og konkurssannsynlighet hos de ulike institusjonene. De ulike porteføljene sorteres etter sannsynlighet for mislighold. Selskaper som har misligholdsannsynlighet større enn 3,0 prosent blir definert som et selskap med høy risiko for å gå konkurs. Med bakgrunn i SEBRA-modellen sine beregninger får man en risikoklassifisert portefølje som igjen danner et sammenligningsgrunnlag mot institusjonens egen. Man kan således si at modellen er en viktig verktøy for utvelgelse av enkeltengasjement risikoklassifisering (Finanstilsynet, 2013b). SEBRA-modellen er fortsatt høyaktuell i dag da den er blitt benyttet i de siste tre års Finansielt Utsyn, publisert av Finanstilsynet (2015b, 2014, 2013a). Modellen er med andre ord fremdeles en viktig modell for Finanstilsynet og Norges Bank.

2.2.4 Svakheter ved SEBRA-modellen

SEBRA-modellen benytter kun norske aksjeselskaper hvor regnskap kan hentes i Brønnøysund sine register. Det betyr at modellen ikke nødvendigvis er representativ på aksjeselskaper i andre land enn Norge og på ikke-aksjeselskaper i Norge. I denne oppgaven vil dette imidlertid ikke by på problemer, da det kun anvendes norske regnskapsdata. Eiendomsselskaper med en høy andel prosjekter under utvikling er vanskelig å analysere for

SEBRA-modellen på grunn av at det ofte er lang tid mellom når investering gjennomføres og når penger kommer inn (Finanstilsynet, 2013b). Det vil med andre ord oppstå store tids-lag. Dette problemet oppstår på grunn av at SEBRA-modellen baserer seg på årsregnskapsdata.

Til nå i dette kapitlet har Norges Banks SEBRA-modell blitt gjennomgått. I fortsettelsen gis en gjennomgang av flere anerkjente internasjonale modeller for konkursprediksjon, samt en gjennomgang av hvordan modellene er estimert.

2.3 Beavers univariate modell

Univariate modeller antar at en enkelt uavhengig variabel kan benyttes for å predikere en avhengig variabel, i dette tilfellet konkurs eller ikke konkurs. I analysene ble det benyttet forhåndsdefinerte faktorer til å skille mellom konkurs og ikke-konkursselskaper. Bidrag innenfor denne estimeringsmetoden kommer fra blant annet *Bureau of Business Research* (1930), FitzPatrick (1932) og Beaver (1966), for å nevne noen. I denne oppgaven vil kun Beavers univariate modell behandles nærmere.

Da Beaver lanserte sin univariate modell, hadde det blitt forsket på konkursprediksjon i 35 år uten at man hadde funnet modeller som klarte å vise til signifikante resultater eller gode prediksjonsevner (Lawrence et al., 2015). Beaver brukte i sin modell et parret utvalg, hvor et insolvent selskap ble parret med et solvent selskap. Selskapene ble klassifisert som insolvent ved:

- (1) at finansielle forpliktelser ikke gjøres opp til forfall,
 - (2) konkurs,
 - (3) mislighold av obligasjonsgjeld,
 - (4) overtrukket bankkonto,
 - (5) eller manglende betaling av foretrukket aksjeutbytte
- (Laitinen og Laitinen, 2000).

Beaver mente at å pare et stort og et mindre selskap kunne gi gale indikasjoner. Han valgte derfor å pare selskaper med tilnærmet samme total kapital innen samme bransje. Valget ble begrunnet med at store selskap går nesten aldri konkurs. Å være mindre solvent enn et stort selskap trenger for eksempel ikke å bety at selskapet er insolvent.

Beaver hentet data fra Moody's sine databaser fra perioden 1954 til 1964, hvor 79 konkurssekskaper fra 38 forskjellige bransjer ble parett med 79 ikke-konkurssekskaper med tilnærmet samme størrelse innen samme bransje. Industriefeffekten og størrelseeffekten ble på den måten minimert ved at man sammenlignet et insolvent selskap og et solvent selskap av samme størrelse (Beaver, 1966). Totalkapitalen til selskapene i utvalget strakte seg fra \$0,6 millioner til \$45,0 millioner, med et gjennomsnitt på \$6,0 millioner. Beaver benyttet 30 finansielle nøkkeltall som han trodde kunne ha innvirkning på et selskaps solvens eller insolvens, hvor nøkkeltallene ble delt opp i seks grupper:

1. *Nøkkeltall knyttet opp mot kontantstrøm*
2. *Nøkkeltall knyttet opp mot nettoinntekt*
3. *Nøkkeltall knyttet opp mot gjeld i forhold til totalkapital*
4. *Nøkkeltall knyttet opp mot likvide eiendeler i forhold til totalkapital*
5. *Nøkkeltall knyttet opp mot likvide eiendeler i forhold til nåværende gjeld*
6. *Nøkkeltall knyttet opp mot omsetning*

Nøkkeltallene ble valgt ut fra popularitet og tidligere bruk i tilsvarende modeller. At nøkkeltallene også enkelt kunne forklares ved hjelp av en kontantstrømmodell ble vektlagt.

Beaver så på selskapet som et reservoar av eiendeler, og endringer i dette reservoaret påvirket selskapets solvens og videre drift. Med bakgrunn i dette trakk Beaver (1966) fire konklusjoner:

1. *Et større reservoar av likvide eiendeler gir lavere sannsynlighet for å gå konkurs*
2. *Jo større netto kontantstrøm fra driften av selskapet, jo lavere sannsynlighet for konkurs*
3. *Sannsynligheten for konkurs øker med økte driftsutgifter,*
4. *Sannsynligheten for konkurs øker med økende gjelden*

Analysen av data ble gjennomført i en trestegs prosess. Fase én besto av en profilanalyse for å sammenligne gjennomsnittsverdier til konkurs og ikke-konkurssekskaper. I fase to ble det gjennomført en dikotom klassifiseringstest. De finansielle nøkkeltallene analyseres i fase tre for å finne sannsynligheten for feilklassifiseringer. Et normtall ble benyttet for å klassifisere hvorvidt et selskap var konkurstruet eller ikke (Beaver, 1966).

Nøkkeltallenes prediksjonsevne ble så testet i utvalget og ble plukket ut med hensyn på hva som faktisk predikerer best. Beaver endte da opp med fem variabler slik som illustrert i tabell 3.

Tabell 3: Prosentvis oversikt av feilklassifiserte selskaper: dikotom klassifiseringstest. Nederste rad (i parantes) representerer resultater fra en første test utført av Beaver (1966), mens øverste rad representerer resultater fra en andre test.

Ratio	Year before Failure				
	1	2	3	4	5
$\frac{\text{kontantstrøm}}{\text{samlet gjeld}}$	0,13 (0,10)	0,21 (0,18)	0,23 (0,21)	0,24 (0,24)	0,22 (0,22)
$\frac{\text{netto resultat}}{\text{totalkapital}}$	0,13 (0,12)	0,20 (0,15)	0,23 (0,22)	0,29 (0,28)	0,28 (0,25)
$\frac{\text{samlet gjeld}}{\text{totalkapital}}$	0,19 (0,19)	0,25 (0,24)	0,34 (0,28)	0,27 (0,24)	0,28 (0,27)
$\frac{\text{arbeidskapital}}{\text{totalkapital}}$	0,24 (0,20)	0,34 (0,30)	0,33 (0,33)	0,45 (0,35)	0,41 (0,35)
likviditetsgrad 1	0,20 (0,20)	0,32 (0,27)	0,36 (0,31)	0,38 (0,32)	0,45 (0,31)
likvide eiendeler – kortsiktig gjeld	0,23 (0,23)	0,38 (0,31)	0,43 (0,30)	0,38 (0,35)	0,37 (0,30)
totalkapital	0,38 (0,38)	0,42 (0,42)	0,45 (0,42)	0,49 (0,41)	0,47 (0,38)

Fra tabell 3 kommer det frem at forholdet mellom kontantstrøm og total gjeld har færrest feilklassifiseringer, mens forholdet mellom nettoinntekt og totalkapital har nest færrest feilklassifiseringer. I tabellen ser man at de har en feilklassifisering på 13% i år 1 og henholdsvis 22% og 28% i år 5, som markert i fet tekst i tabell 3.

Fordelen med en univariat modell, slik som Beaver benytter, er at den er enkel å ta i bruk i praksis. Krevende beregninger utelates fra modellen og man trenger dermed ikke ta i bruk avanserte statistiske verktøy. Likevel er det noen svakheter med modellen – man kan for eksempel ikke trekke en konklusjon basert på en enkelt observasjon. Dette fordi de ulike nøkkeltallene faktisk kan indikere to forskjellige utfall; enten er det et konkursselskap eller et ikke-konkursselskap. Man kan også få upresise prediksjoner hvis det er sterk korrelasjon mellom variablene (Beaver, 1966). Han presiserte at hans egen forskning kun var et bidrag til den videre forskningen innenfor temaet.

2.4 Altmans multivariate modeller

I dette delkapitlet behandles multivariate modeller hvorav Z-score-modellen, Zeta-modellen og den reviderte Z-score-modellen er representert. Z-score-modellen er utviklet av Edward I. Altman i 1968 og er av flere regnet som en av de mest kjente og sentrale kredittrisikomodelle (Bellovary et al., 2007). Altman så klare svakheter i univariate modeller.

I motsetning til Beaver som benyttet en univariat modell, som kun ser på et forholdstall om gangen, valgte Altman å benytte seg av multivariate modeller, som ser på flere forholdstall samtidig. Fordelen med dette er at da blir det også tatt hensyn til samvariasjon mellom forholdstallene. I denne typen analyse klassifiseres en observasjon inn i en av flere a priori grupperinger ut i fra observasjonens individuelle kjennetegn (Altman, 1968). Når observasjonene er inndelt i grupperinger vil en multivariat modell lage en lineærkombinasjon som best skiller mellom de ulike gruppene.

Modellen kan uttrykkes på følgende form:

$$Z = v_1x_1 + v_2x_2 + \dots + v_nx_n \quad (4)$$

der Z er den avhengige variabelen og $x_1, x_2 \dots x_n$ er uavhengige variabler. $v_1, v_2, \dots v_n$ angir vektene til henholdsvis $x_1, x_2 \dots x_n$, og ved estimering av modellen er det nettopp disse vektene som estimeres.

For å estimere den multivariate modellen benyttet Altman seg av diskriminantanalyse, hvor observasjoner skiller ved å kategorisere deres tilhørende avhengige variabler (Wetcher-Hendricks, 2014). Diskriminantanalyse benyttes ofte for å løse klassifiseringsproblemer (Guo et al., 2007). Metoden ble først introdusert av Ronald A. Fisher (1936), da kjent som den lineære diskriminantanalysen (LDA). Ordet lineær kommer av antakelsen om lik kovariansmatrise blant kategoriene (Guo et al., 2007).

Dersom kovariansmatrisen til kategoriene derimot er forskjellige er det vanlig å benytte kvadratisk diskriminantanalyse (QDA) (Guo et al., 2007). Joy og Tollefson (1975) hevder også at den beste tilnærmingen er å benytte kvadratisk diskriminantanalyse når kovariansmatrisen til de gjensidig utelukkende gruppene er forskjellig.

For å estimere sin modell benyttet Altman et datautvalg fra 66 selskaper hvorav 33 var konkursselskaper og 33 var ikke-konkursselskaper. Gjennomsnittlig total kapital i utvalget var på \$6,4 millioner, med et intervall i total kapital fra \$0,7 millioner til \$25,9 millioner. Med

bakgrunn i at det forekommer industri- og størrelsesforskjeller ble utvalget av konkursselskaper grundig valgt ut (Altman, 1968).

Tidligere forskning hadde påvist mange signifikante variabler, og Altman brukte denne forskningen i sitt innledende arbeid. Han endte til slutt opp med 22 variabler som han mente kunne ha betydning for studiet, basert på hvor ofte de var behandlet i annen litteratur, samt hva han selv mente ville ha betydning for prediksjonen. Altman delte de 22 variablene i fem kategorier:

- Likviditet
- Lønnsomhet
- Gjeld
- Soliditet
- Aktivitetsnøkkeltall

For å finne beste kombinasjonen av variabler ble de ulike nøkkeltallene kombinert med hverandre og følgende ble vektlagt:

1. *Hvor signifikante er variablene*
2. *Korrelasjon mellom variablene*
3. *Modellens treffsikkerhet*
4. *Evaluerings av analysen*

Dette resulterte i en multipl diskriminant-funksjon:

$$Z = 0,012x_1 + 0,014x_2 + 0,033x_3 + 0,006x_4 + 0,999x_5 \quad (5)$$

der $x_1 = \frac{\text{arbeidskapital}}{\text{totalkapital}}$, $x_2 = \frac{\text{tilbakeholdt overskudd}}{\text{totalkapital}}$, $x_3 = \frac{\text{resultat før renter og skatt}}{\text{totalkapital}}$,

$x_4 = \frac{\text{markedsverdi egenkapital}}{\text{samlet gjeld}}$ og $x_5 = \frac{\text{salgsinntekter}}{\text{totalkapital}}$

Målingene ble objektivt vektet og summert sammen til en total Z-score som skilte konkursselskaper fra ikke-konkursselskaper. En Z-verdi over 2,99 og under 1,81 ga en korrekt klassifisering av selskapene. Verdier under 1,81 tilsa konkurs, mens verdier over 2,99 indikerte ikke konkurs. For selskaper som fikk en Z-verdi mellom 1,81 og 2,99 kunne predikeringen være noe usikker.

Tabell 4: Oppsummering av resultatene til Altman (1968)

År før konkurs	Antall treff	Antall ikke treff	Treffprosent
1 år, n = 33	31	2	95
2 år, n = 33	23	9	72
3 år, n = 33	14	15	48
4 år, n = 33	8	20	29
5 år, n = 33	9	16	36

I tabell 4 gis en oversikt over Altmans modells treffsikkerhet opp til fem år før konkurs. Tabellen viser at modellen predikerer godt til og med to år før konkurs, med treffprosent på 95% og 72% for henholdsvis ett og to år før konkurs.

2.4.1 Den reviderte Altman Z-Score-modellen

Altman (2000) reviderte i 2000 sin opprinnelige Z-score-modell fra 1968. Denne ble kalt den reviderte fire-variable Z-score-modellen. Ved å utelate variabelen x_5 , $\frac{\text{salgsinntekter}}{\text{total kapital}}$, reduserte Altman sannsynligheten for industrieffekt. I tillegg valgte han også å inkludere selskaper som ikke driver med vareproduksjon. Til slutt valgte Altman å erstatte egenkapitalens markedsverdi med egenkapitalens bokførte verdi (Oluwo, 2007). Dette ga følgende koeffisienter for den reviderte modellen:

$$Z = 6,567x_1 + 3,26x_2 + 6,72x_3 + 1,05x_4 \quad (6)$$

der x_1 , x_2 , x_3 og x_4 er som i den alminnelige Z-score-modellen angitt i ligning (5). I tillegg til at x_5 er utelatt, skiller (6) seg fra (5) ved at x_1 har fått en høyere vektning, mens x_4 har fått en lavere vektning. En samlet Z-score på 1,10 eller mindre indikerte en høy sannsynlighet for konkurs. En Z-score høyere enn 2,60 indikerte at sannsynligheten for konkurs var minimal. Verdier mellom 1,10 og 2,60 blir ofte kalt en overlappingszone eller en gråzone. Den reviderte modellen har vist seg nyttig spesielt i leasing-industrien hvor eiendelfinansieringen varierer fra selskap til selskap (Oluwo, 2007).

2.4.2 Zeta-modellen

Zeta-modellen ble utviklet av Altman et al. (1977) på midten av 70-tallet. I motsetning til Z-score-modellen, valgte de å inkludere større selskaper i Zeta-modellen. Av den grunn predikerer Zeta-modellen bedre for større selskaper sammenlignet med Z-score-modellen. For å få flest mulig signifikante variabler startet de med 27 variabler som de antok kunne ha

betydning for et selskaps videre solvens. Ikke-signifikante variabler ble eliminert helt til de sto igjen med kun de syv variablene som ga den mest presise prediksjonen.

Zeta-modellen består av følgende variabler:

x_1 : Totalkapitalrentabilitet

x_2 : Stabilitet i inntektene

x_3 : Rentedekningsgrad

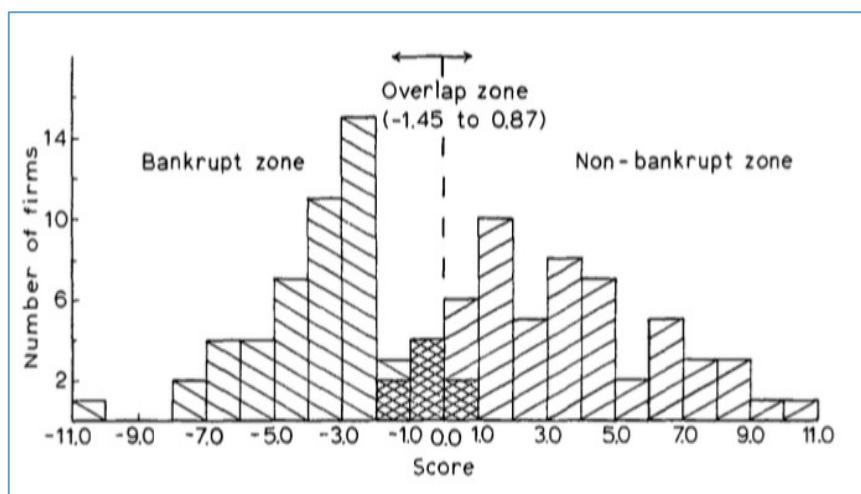
x_4 : Kumulativ profitt ($\frac{\text{tilbakeholdt overskudd}}{\text{totalkapital}}$)

x_5 : Likviditet

x_6 : Egenkapitalprosent

x_7 : Størrelse (målt i totalkapital)

hvor x_7 er logaritmisk transformert for å oppnå homoskedastisitet og normalfordeling i datautvalget. Homoskedastisitet innebærer at observasjoner av feilleddet har konstant varians, mens når variansen ikke er konstant er det heteroskedastisitet. En test av variablene viser at totalkapitalrentabilitet var variabelen som forklarte mest av variasjonen, etterfulgt av stabilitet i inntektene. Totalkapitalrentabilitet forklarte hele 25% av variasjonen (Altman et al., 1977). I utvalget brukte de data fra 53 insolvente selskaper og 58 solvente selskaper i estimeringen, og som tidligere nevnt ble også større selskap inkludert her. I figur 2 er de ulike selskapene kategorisert langs x-aksen. Y-aksen viser antall selskap som er på de ulike Z-verdiene i figuren.



Figur 2: Gjennomsnittlig Zeta-score på de ulike selskapene i benyttet datautvalg fordelt etter soner (Altman et al., 1977)

Figur 2 er delt inn i tre kategorier. Man kan se at Z-verdi på mellom $-1,45$ og $0,87$ er en overlappingszone. En Z-verdi under $-2,675$ innebærer høy sannsynlighet for insolvens og en verdi over $2,675$ innebærer høy sannsynlighet for at selskapet er solvent (Altman et al., 1977).

2.4.3 Sammenligning av Z-score-modellen og Zeta-modellen

I de forrige delkapitlene ble Zeta-modellen og Z-score-modellen behandlet. I tabell 5 ser vi en sammenligning av prediksjonsevnen til disse metodene.

Tabell 5: Sammenligning av Altmans Zeta-modell og Altmans Z-score-modell (Altman, 1968, Altman et al., 1977)

År før konkurs	Prediksjonsevne Zeta-modellen		Prediksjonsevne Z-score-modellen	
	Konkurs	Ikke konkurs	Konkurs	Ikke konkurs
1	96,2	89,7	93,9	97,0
2	84,9	93,1	71,9	93,9
3	74,5	91,4	48,2	
4	68,1	89,5	28,6	
5	69,8	82,1	36,0	

Tabellen viser at på ett års sikt predikerer modellene omtrentlig like godt. Når prediksjonen er for en lengre tidsperiode kommer det frem av tabellen at prediksjonsevnen til de ulike modellene spriker. For Z-score-modellen ble det ikke testet for prediksjonsevne for ikke konkurs for mer enn de to siste årene før konkurs, derfor er disse cellene blanke i tabell 5. Zeta-modellen er langt mer presis når det skal predikeres konkurs for mer enn to år. Når det skal predikeres for ikke konkurs viser tabellen derimot at Z-score-modellen predikerer mer presist enn Zeta-modellen, i alle fall for siste to år før konkurs.

2.5 Ohlsons logistiske regresjonsmodell

Multivariate modeller ble behandlet delkapittel 2.4, mens modellen i dette delkapitlet er av logistisk form. I følge Aldrich og Nelson (1984) må følgende betingelser være oppfylt før logistisk regresjon kan tas i bruk:

1. Avhengig variabel kan ha en av to verdier; 1 eller 0
2. Sannsynligheten for at den avhengige variabelen y er lik 1 er avhengig av de uavhengige variablene slik som illustrert i følgende formel:

$$P(Y = 1) = \frac{e^{(b_0 + b_1x_1 + \dots + b_nx_n + e)}}{(1 + e^{(b_0 + b_1x_1 + \dots + b_nx_n + e)})} \quad (7)$$

3. Observasjonene av den avhengige variabelen antas uavhengige av hverandre
4. Det er ikke signifikant lineær sammenheng mellom to eller flere uavhengige variabler, dvs. ikke multikollinearitet

Modellen presentert av James A. Ohlson (1980) regnes som en av de mer sentrale modellene innen konkursprediksjon i den vestlige delen av verden (Lawrence et al., 2015). Den er bedre kjent som O-score-modellen og ble utviklet sent på 70-tallet. Utvalget benyttet i modellen besto av 105 konkursselskaper og 2058 ikke-konkursselskaper. Modellen var den første til å benytte et representativt utvalg, og ble laget som en reaksjon på tidligere modeller innen kredittrisiko. Ohlson var spesielt kritisk til det parede utvalget som ble brukt i noen tidligere modeller, fordi han mente at dette blant annet kunne føre til at viktig informasjon gikk tapt. Ohlson viste også til flere svakheter ved den vanligste metoden, multippel diskriminantanalyse. Multippel diskriminantanalyse antar for eksempel at varians-kovariansmatrisen er lik for alle grupper som skal predikeres. Ohlson mente at dette sjelden er tilfelle og at metoden derfor kan gi lite presise prediksjoner. Resultatet man får fra multippel diskriminantanalysen gir også ifølge Ohlson et lite intuitivt resultat, da det i utgangspunktet er et verktøy for å rangere på ordinalnivå. Ohlson (1980) hevdet også at paringen i modellen i henhold til bransje og størrelse hadde en tendens til å være noe vilkårlig. Han testet tidligere modeller innenfor området, blant annet Altmans Z-score-modell, og fant større feilprediksjon enn det forskningen tilsa ut i fra data hentet fra nyere tid. Patti (2001) hevder at eldre modeller var mer opptatt av forklaring enn av prediksjon, mens de senere modellene derimot er mer opptatt av selve prediksjonen.

Ohlson brukte ni variabler i sin prediksjonsmodell for å predikere insolvens, hvor to av variablene kun kan ha verdiene 0 eller 1, og dermed er dummyvariabler. Ohlson laget tre modeller: en modell for et år, en for to år gitt at selskapet ikke går konkurs i år 1 og en som predikerer konkurs i løpet av to år. Det ble satt en grenseskille på 0,5 for alle modellene, som skilte mellom konkurs og ikke konkurs. Ohlson fant at et grenseskille på 0,038 reduserte prosentvis antall feilestimeringer, men likevel ga det samlet feilprosent på 14,9% (Ohlson, 1980). Dette er noe dårligere enn tidligere modeller.

Følgende variabler ble benyttet i Ohlsons modell;

$$x_1: \text{Size} = \log\left(\frac{\text{totalkapital}}{\text{BNP prisindeks}}\right)$$

$$x_2: \text{TLTA} = \frac{\text{samlet gjeld}}{\text{totalkapital}}$$

$$x_3: \text{WCTA: } \frac{\text{arbeidskapital}}{\text{totalkapital}}$$

$$x_4: \text{CLCA: } \frac{\text{kortsiktig gjeld}}{\text{omløpsmidler}}$$

$x_5: \text{OENEG: } 1$ dersom samlet gjeld er større enn totalkapital, ellers 0

$$x_6: \text{NITA: } \frac{\text{netto resultat}}{\text{totalkapital}}$$

$$x_7: \text{FUTL: } \frac{\text{driftsresultat}}{\text{totalkapital}}$$

$x_8: \text{INTWO: } 1$ dersom negativt årsresultat de to siste årene; 0 ellers

$x_9: \text{CHIN:}$ Endring i nettoresultat, målt de to siste årene

y : Dummyvariabel: 1 dersom konkurs, ellers 0

Alle modellene som er nevnt til nå, fra SEBRA-modellen til Ohlsons O-score-modell, er estimert ved tradisjonelle statistiske metoder. I resten av dette kapitlet er det nyere statistiske metoder, såkalte maskinlæringsteknikker, som blir gjennomgått.

2.6 Maskinlæringsteknikker

For å utvikle og teste modeller for konkursprediksjon finnes det mange mulige metoder. Ravi Kumar og Ravi (2007) gir en oversikt over metoder benyttet ved konkursprediksjon i artikler fra perioden 1968-2005, og de har valgt å gruppere metodene i to hovedkategorier: (i) statistiske metoder og (ii) intelligente metoder. Sistnevnte omtales i denne oppgaven som maskinlæringsteknikker. Som statistiske metoder omtaler Ravi Kumar og Ravi (2007) blant annet diskriminantanalyse og logistisk regresjon, mens de omtaler blant annet nevrale nettverk, beslutningstrær og støttevektormaskin som intelligente metoder.

25 av de 62 artiklene som er omtalt i oversiktsartikkelen til Ravi Kumar og Ravi (2007) benytter nevrale nettverk (NN). Teknikken er dermed den som er benyttet i flest artikler i deres studie, noe som kan indikere at nevrale nettverk er den mest benyttede teknikken for konkursprediksjon i nyere tid.

Kirkos (2015) tar for seg tidligere forskning på konkursprediksjon ved bruk av maskinlæringsteknikker, og analyserer totalt 42 artikler utgitt av anerkjente journaler i

perioden 2009-2011. Foruten NN er det støttevektormaskin (engelsk: support vector machine) (SVM) som benyttes mest av disse artiklene.

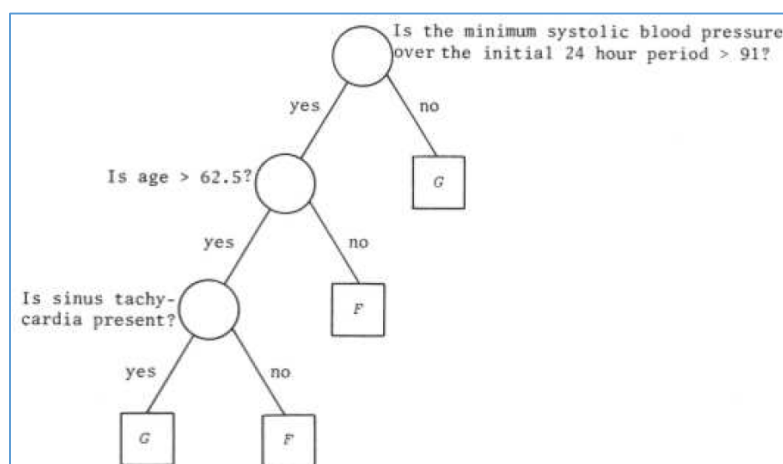
Med bakgrunn i dette er det i denne oppgaven valgt å benytte SVM og flere varianter av NN. I tillegg blir maskinlæringsteknikkene k-nærmeste naboer (KNN) og klassifiseringstrær (CT) benyttet. Alle disse fire blir omtalt i detalj nedenfor.

2.6.1 Klassifiseringstrær

Klassifiseringstrær er en variant av beslutningstrær, hvor formålet er å predikere en kategori for den avhengige variabelen, i denne oppgaven «konkurs» eller «ikke konkurs», ved å ta en lang rekke todelte beslutninger (Ledolter, 2013). Hver beslutning som må tas blir kalt *noder*, mens det endelige utfallet etter beslutninger i hver node kalles *løv*. Figur 3 viser et eksempel på strukturen i et beslutningstre, hvor nodene er illustrert ved rundinger og løvene illustrert ved firkanter.

Ved benyttelse av klassifiseringstrær er det forhåndsbestemt hvilke mulige verdier løvene kan ha, men det er uendelig mange kombinasjoner av antall noder og løv. For å finne den optimale kombinasjonen blir det i denne oppgaven testet med forskjellige parametere på metoden, som blir omtalt i kapittel 3.

Cielen et al. (2004) har benyttet denne teknikken i sitt arbeid innen konkursprediksjon og klarte den gang å oppnå en treffprosent på 80%.



Figur 3: Eksempel på struktur for klassifiseringstrær (Breiman et al., 1984)

2.6.1 *k*-nærmeste naboer

k-nærmeste naboer er blant de enkleste maskinlæringsalgoritmene (Ledolter, 2013). Ved bruk av denne metoden på et klassifiseringsproblem, som er tilfelle i denne oppgaven, plottes alle tidligere observasjoner i et rom med like mange dimensjoner som uavhengige variabler (Cover og Hart, 1967). I tillegg blir hver observasjon kategorisert i en klasse, som for denne oppgaven er klassene «konkurs» eller «ikke konkurs». Når man skal predikere klassen til en ny observasjon plottes denne i det flerdimensjonale rommet. Man finner så den eksisterende observasjonen som har minst avstand til den nye observasjonen, for så å sette klassen til den nye observasjonen lik klassen til denne nærmeste eksisterende observasjonen.

Når datagrunnlaget er stort nok er det ofte ønskelig å se på flere nærliggende observasjoner (*k*) da dette vil gi større sjanse for riktig verdi på den nye observasjonen (Cover og Hart, 1967). Samtidig er det ønskelig å ikke se på for mange eksisterende observasjoner (*k*), da det kan føre til at man ikke lengre kun ser på de nærmeste. I kapittel 3 blir det funnet en optimal verdi for hvor mange observasjoner (*k*) som skal tas i betraktning ved estimeringer av modeller som skal benyttes i denne oppgaven.

Ved kontinuerlige uavhengige variabler anbefales det å benytte en euklidsk avstand for å finne avstanden mellom nye og eksisterende observasjoner (Ledolter, 2013). Denne beregningsmåten gir summen av absoluttverdien til avstanden langs alle dimensjoner mellom to observasjoner. Jo mindre denne summen er, jo mindre er avstanden mellom observasjonene. Denne oppgavens datautvalg består også av noen ikke-kontinuerlige uavhengige variabler, så det blir testet for flere avstandsmål enn det euklidske, men siden det er euklidsk avstand som til slutt blir benyttet så blir ikke de andre avstandsmålene presentert her.

Park og Han (2002) benyttet seg av denne teknikken innen konkursprediksjon, og oppnådde da en treffprosent på 74%.

2.6.2 Støttevektormaskiner

Støttevektormaskiner (support vector machines, SVM) ble for første gang introdusert av Vapnik (1998). SVM var originalt utviklet for forskningsproblemer med binær klassifisering, altså to kategorier (Boyacioglu et al., 2009). Siden konkursprediksjon har to klassifiseringer, konkurs eller ikke konkurs, er metoden egnet til nettopp estimering av modeller for

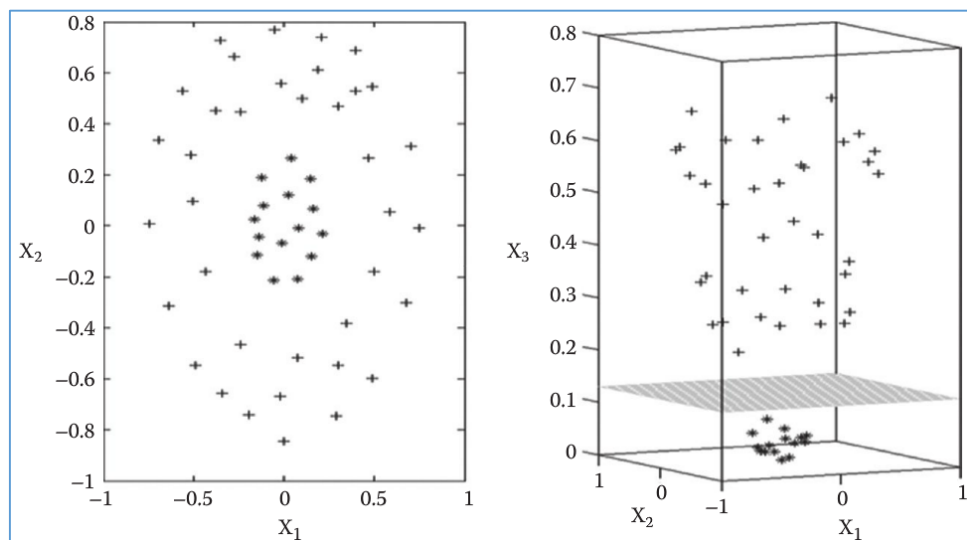
konkursprediksjon. Dette bekreftes også ved at flere av artiklene omtalt i Ravi Kumar og Ravi (2007) og Kirkos (2015) benytter metoden.

Ved å benytte SVM må man først tenke at alle uavhengige variabler i en observasjon utgjør en vektor. Vektorene med alle observerte variabelverdier transformeres så ved hjelp av kernel-funksjoner til vektorer med flere dimensjoner (Liang et al., 2011). Deretter plottes de i et vektorrom, hvor SVM benyttes for å finne det hyperplanet som skiller vektorene innenfor de to klassene (konkurs og ikke konkurs) fra hverandre på den måten som maksimerer avstanden mellom selve hyperplanet og de to nærmeste vektorene (Boyacioglu et al., 2009).

Transformasjonen av vektorene ved hjelp av kernel-funksjoner gjør det mulig å finne et hyperplan som skiller vektorene innenfor de to klassene, selv om det ikke var mulig før transformasjonen. Dette illustreres i figur 4 nedenfor fra Liang et al. (2011) som viser at det er mulig å skille stjernene fra plussene med et hyperplan kun etter de er transformert fra todimensjonale til tredimensjonale vektorer.

De vanligste kernel-funksjonene er lineær kernel-funksjon, polynomial kernel-funksjon, radial basis kernel-funksjon og sigmoidal kernel-funksjon (Liang et al., 2011). I denne oppgaven blir lineær kernel-funksjon benyttet, da denne er best forstått og enklest å anvende (Cristianini og Shawe-Taylor, 2000).

Det kanskje mest kjente studiet som benytter teknikken innenfor konkursprediksjon er utført av Boyacioglu et al. (2009), som oppnådde en treffprosent på 90,9%. Hung og Chen (2009) benyttet også SVM til konkursprediksjon, og oppnådde en treffprosent på 71%.



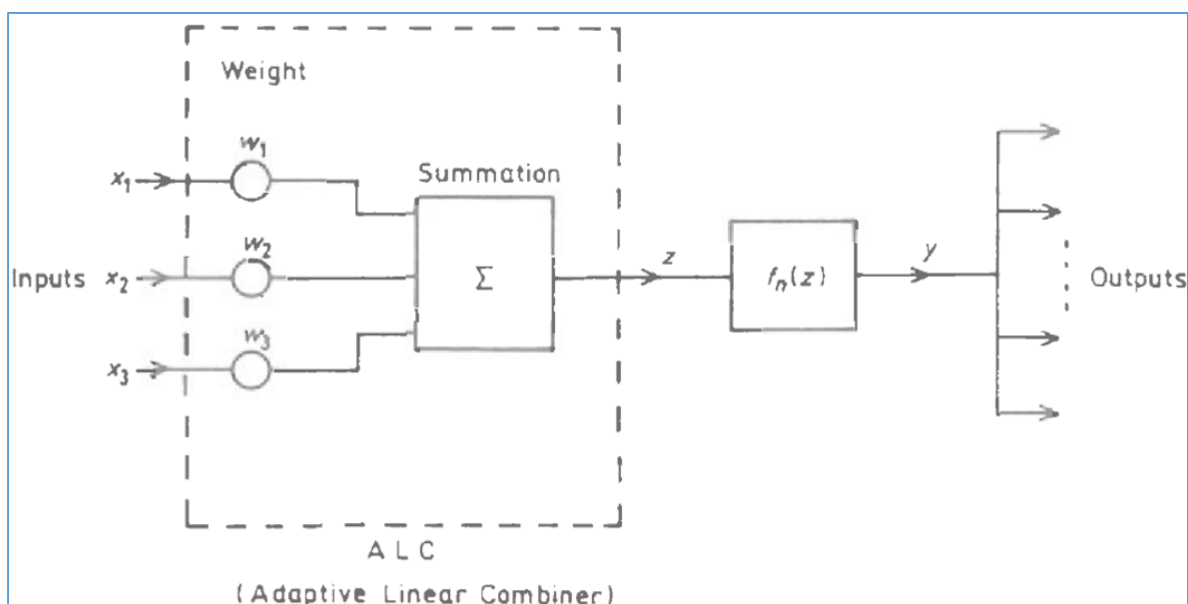
Figur 4: Todimensjonale vektorer innenfor to forskjellige klasser som kan skilles av et hyperplan først etter de er transformert til tredimensjonale vektorer (Liang et al., 2011).

2.6.3 Nevrale nettverk

Nevrale nettverk (NN) er metoder for behandling av empiri basert på hvordan biologiske nevrale systemer fungerer (Güres et al., 2012). Ifølge Daniel Graupe (2013) var perceptron-nettverket sannsynligvis det første NN, og det ble presentert av Frank Rosenblatt (1958). I ettertid har det kommet flere varianter av NN.

For å forklare hvordan NN fungerer blir elementene i et enkelt perceptron-nettverk, som presentert i figur 5. Dette NN mottar en rekke uavhengige variabler, her markert med x , som alle går inn i noder. I nodene blir verdien av de uavhengige variablene multiplisert med en vekt (w) som har forskjellig verdi for hver node. Produktet av alle uavhengige variabler (x) og tilhørende vekter (w) summeres så til en verdi som kalles perceptronet (z). Dette perceptronet er så den uavhengige variabelen til en funksjon (f). Resultatet fra denne funksjonen blir så nettverkets endelige avhengige variabel (y) som kan bestå av en eller flere verdier.

NN skal kunne trenes. Det vil si å bruke eksisterende datasett, med uavhengige variabler (x) og tilhørende avhengige variabler (y), for å tilpasse nettverket (Graupe, 2013). Én etter én kjøres settene med de uavhengige variablene (x) inn i det nevrale nettverket og den avhengige variabelen (y) måles. Deretter måles avviket mellom denne og den avhengige variabelen datasettet skulle hatt, og dette avviket benyttes for å tilpasse vektene (w) slik at det nevrale nettverket gir minst mulig totalavvik mellom de uavhengige variablene (x) og de avhengige variablene (y) i samtlige datasett som blir benyttet i treningen.



Figur 5: Grunnmodell for et nevralt nettverk (Graupe, 2013)

Funksjonen (f) er forhåndsdefinert, og dermed endres derfor ikke av treningen. I denne oppgaven er de uavhengige variablene (x) nøkkeltall hentet fra offisielle regnskaper. Videre kan resultatet (y) kun ha en av to verdier; «konkurs» eller «ikke konkurs». Funksjonen (f) blir derfor tilpasset slik at den gir resultatet (y) 1 eller 0.

Eksempelet ovenfor viser som nevnt et NN som er et etlags perceptron. Det er også mulig å lage NN med flere lag av denne typen (Graupe, 2013), såkalt flerlags perceptron. I NN med flere lag er input til det ene laget output i det neste. I tillegg er det mulig å tilpasse antall noder i hvert lag.

All informasjon i det beskrevne NN beveger seg fremover, fra input til output, og kalles på grunn av dette forovermatet (Graupe, 2013). Nevrale nettverk kan også være tilbakevendende, hvor output mates tilbake inn i nettverket.

Innenfor konkursprediksjon er NN blitt brukt flere ganger. Det kan blant annet nevnes at Boyacioglu et al. (2009) og Wilson og Sharda (1994) oppnådde treffprosjenter på henholdsvis 95,5% og 97% ved bruk av NN.

Dype nevrale nettverk er NN som inneholder svært mange lag og som gjør flere kalkuleringer enn ikke-dype NN (Bengio, 2009, Schmidhuber, 2015). En autoencoder er en type NN som rekonstruerer sine input til output, men med færre dimensjoner (Vishnubhotla et al., 2010, Hinton og Salakhutdinov, 2006). Backpropagation vil si at etter det nevrale nettverket er konstruert så benyttes differansen mellom estimerte avhengige variabler (\hat{y}) og faktiske avhengige variabler (y) til å rekalkulere vektene (w) (Graupe, 2013).

I denne oppgaven er det tre typer nevrale nettverk som blir benyttet ved analyse. Det ene er et flerlags forovermatet NN, det andre er et dypt nevralt nettverk skapt av to autoencodere (DNNA), mens det tredje er et dypt nevralt nettverk med backpropagation (DNNB).

2.7 Oppsummering av tidligere arbeid

I dette kapitlet er det introdusert både maskinlæringsteknikker og tradisjonelle statistiske metoder til bruk ved estimering av modeller for konkursprediksjon. Blant eksisterende modeller ble først Norges Banks SEBRA-modell presentert. Deretter ble Beavers univariate modell presentert, før det ble gitt en gjennomgang av Altmans multivariate modeller som av flere anses å være de mest anerkjente modellene innenfor konkursprediksjon. Dernest ble Ohlsons logistiske modell gjennomgått.

Til slutt ble maskinlæringsteknikker presentert, hvor nevrale nettverk, støttevektormaskin, k -nærmeste naboer og klassifiseringstrær ble nærmere gjennomgått. Varian (2014) hevder at maskinlæringsteknikker er mer passende for prediksjon enn tradisjonelle statistiske metoder, i alle fall ved større datasett. Dette blir analysert nærmere i denne oppgaven, og i det påfølgende kapitlet blir fremgangsmåten for denne analysen forklart.

3 Forskningsmetode

I denne oppgaven skilles det som nevnt mellom modell, variabelsett og metode. En *modell* defineres som en forenkling av virkeligheten, hvor en avhengig variabel blir forklart av en eller flere uavhengige variabler. Et *variabelsett* defineres som en samling av nøkkeltall fra årsregnskaper som benyttes som de uavhengige variablene i en gitt modell. *Metoder* er ulike teknikker for å estimere modeller ved hjelp av et datautvalg og et gitt variabelsett.

Konkursprediksjon er et typisk klassifiseringsproblem, og innebærer derfor en totrinnsprosess (Ravi Kumar og Ravi, 2007). I første trinn estimeres modellene ved å benytte et utvalg av datautvalget, kalt *treningssettet*. I det andre trinnet valideres variabelsettene ved å teste dem på et annet utvalg av datautvalget, kalt *valideringssettet*.

I denne oppgaven blir det estimert modeller basert på tre forskjellige variabelsett. For å estimere modellene benyttes ti forskjellige statistiske metoder, hvorav fire anses som tradisjonelle og seks anses som maskinlæringsteknikker. Med tre forskjellige variabelsett og ti forskjellige metoder for estimering blir det til sammen 30 forskjellige modeller som estimeres og testet.

3.1 Valg av variabelsett

Alle modellene som estimeres i denne oppgaven har den samme avhengige variabelen. De uavhengige variablene til modellene er imidlertid forskjellige, avhengig av hvilket variabelsett de baseres på. Variabelsett 1 består av uavhengige variabler som i stor grad er inspirert av variablene i Norges Banks utvidede SEBRA-modell (Bernhardsen og Larsen, 2007), variabelsett 2 består av uavhengige variabler som i stor grad er inspirert av variablene i Altmans Z-score-modell (Altman, 1968), mens variabelsett 3 består av de samme uavhengige variablene som i variabelsett 1 i tillegg til ytterligere ni uavhengige variabler.

3.1.1 Avhengig variabel

I denne oppgaven defineres det kun én avhengig variabel, som gjelder for alle modeller uavhengig av hvilket av de tre variabelsettene som benyttes. Denne avhengige variabelen kan kun ha verdiene «konkurs» eller «ikke konkurs», og er således en dummyvariabel. Forklaring av hvilke observasjoner som klassifiseres som «konkurs» eller «ikke konkurs» er beskrevet i kapittel 3.6.

Tabell 6: Avhengig variabel

Tag	Navn	Type	Verdi
konkurs	Konkurs	Dummy	1 = konkurs 0 = ikke konkurs

3.1.2 Uavhengige variabler – Variabelsett 1

De uavhengige variablene benyttet i variabelsett 1 er basert på variablene i den utvidede SEBRA-modellen (Bernhardsen og Larsen, 2007) som omtalt i kapittel 2.2.2 . Se tabell 7 for en oversikt, og Appendiks 1 for en utdypende forklaring.

Dummyvariabler med to mulige verdier har vanligvis verdiene «0» eller «1». Enkelte deler av analyseverktøyet benyttet i denne oppgave, Matlab, klarte ikke å håndtere «0» som verdi på den uavhengige dummyvariablen. På grunn av dette ble verdiene «1» eller «2» benyttet på dummyvariablene a_1 - a_8 . Dette hadde ingen effekt på hvordan modellene presterer.

Tabell 7: Uavhengige variabler inspirert av Norges Banks utvidede SEBRA-modell (Bernhardsen og Larsen, 2007)

Tag	Navn	Type	Formel
res	Resultat	Kontinuerlig	Ordinært resultat før av- og nedskrivninger i prosent av total gjeld
ek	Egenkapitalprosent	Kontinuerlig	Egenkapital i prosent av total kapital
likv	Likviditetsprosent	Kontinuerlig	Likvider minus kortsiktig gjeld i prosent av omsetning
eie	Eiendeler	Kontinuerlig	Sum eiendeler i faste kroner
lev	Leverandørgjeldprosent	Kontinuerlig	Leverandørgjeld i prosent av total kapital
avg	Avgiftsprosent	Kontinuerlig	Skyldige offentlige avgifter i prosent av total kapital
a_1	Alder over 1 år	Dummy	1 = selskapets alder er over ett år 2 = selskapets alder er ikke over ett år
a_2	Alder over 2 år	Dummy	1 = selskapets alder er over to år 2 = selskapets alder er ikke over to år
a_3	Alder over 3 år	Dummy	1 = selskapets alder er over tre år 2 = selskapets alder er ikke over tre år
a_4	Alder over 4 år	Dummy	1 = selskapets alder er over fire år 2 = selskapets alder er ikke over fire år
a_5	Alder over 5 år	Dummy	1 = selskapets alder er over fem år 2 = selskapets alder er ikke over fem år
a_6	Alder over 6 år	Dummy	1 = selskapets alder er over seks år 2 = selskapets alder er ikke over seks år
a_7	Alder over 7 år	Dummy	1 = selskapets alder er over syv år 2 = selskapets alder er ikke over syv år
a_8	Alder over 8 år	Dummy	1 = selskapets alder er over åtte år 2 = selskapets alder er ikke over åtte år

3.1.3 Uavhengige variabler – Variabelsett 2

De uavhengige variablene benyttet i variabelsett 2 er inspirert av de uavhengige variablene benyttet i Altmans Z-score-modell (Altman, 1968). I tabell 8 er en oversikt, og i Appendiks 2 er en utdypende forklaring.

Tabell 8: Uavhengige variabler inspirert av Altmans Z-score-modell (Altman, 1968)

Tag	Navn	Type	Formel
x_1	Likviditet	Kontinuerlig	$\frac{\text{arbeidskapital}}{\text{totalkapital}}$
x_2	Lønnsomhet	Kontinuerlig	$\frac{\text{tilbakeholdt overskudd}}{\text{totalkapital}}$
x_3	Gjeld	Kontinuerlig	$\frac{\text{resultat før rentekostnad og skatt}}{\text{totalkapital}}$
x_4	Soliditet	Kontinuerlig	$\frac{\text{markedsverdi til egenkapitalen}}{\text{samlet gjeld}}$
x_5	Aktivitetsnøkkeltall	Kontinuerlig	$\frac{\text{salgsinntekter}}{\text{totalkapital}}$

3.1.4 Uavhengige variabler – Variabelsett 3

Variabelsett 3 består av de samme uavhengige variablene som i variabelsett 1, men i tillegg også ytterligere ni uavhengige variabler, noe som gir totalt 23 uavhengige variabler. Oversikt over disse gis i tabell 9, og det gis en detaljert forklaring i Appendiks 3. Bakgrunnen for at det blir estimert modeller med enda flere uavhengige variabler er at modellen da dekker flere økonomiske aspekter, slik at flere potensielle trender i forkant av en konkurs kan fanges opp. Med så mange variabler kan det imidlertid oppstå problemer med multikollinearitet, men siden modellene i denne oppgaven er konstruert kun for prediksjon er dette allikevel ikke et problem (Berg, 2007).

Tabell 9: Uavhengige variabler for variabelsett 3. I tillegg til disse inngår også samtlige av de uavhengige variablene i variabelsett 1.

Tag	Navn	Type	Formel
opptjek	Opptjent egenkapital	Dummy	1 = opptjent egenkapital er negativ 2 = opptjent egenkapital er positiv
rentgje	Rentebærende gjeld	Kontinuerlig	$\frac{\text{gjeld til kredittinstitusjoner}}{\text{sum gjeld}}$
omseie	Omsetning per eiendel	Kontinuerlig	$\frac{\text{salgsinntekter}}{\text{totalkapital}}$
avkek	Avkastning på egenkapital	Kontinuerlig	$\frac{\text{årsresultat}}{\text{sum egenkapital}}$
omslo	Omsetning lønninger	Kontinuerlig	$\frac{\text{salgsinntekter}}{\text{lønnskostnader}}$
vare	Varelager som andel av totalkapital	Kontinuerlig	$\frac{\text{sum varelager}}{\text{totalkapital}}$
omlopkunde	Omløpshastighet kundefordringer	Kontinuerlig	$\frac{\text{salgsinntekter} \times 1,25}{\text{kundefordringer}}$
omlpst	Omløpsmidler som andel av totalkapital	Kontinuerlig	$\frac{\text{sum omløpsmidler}}{\text{totalkapital}}$
rev	Revisoranmerkninger	Dummy	1 = revisoranmerkning i årsregnskapet 2 = ikke revisoranmerkning

3.2 Valg av metode

For å se om metodevalg gir signifikant forskjell på prediksjonsevne ble det benyttet flere forskjellige metoder ved estimering av modeller basert på variabelsett 1, 2 og 3. GAM ble valgt fordi denne er benyttet ved estimering av Norges Banks SEBRA-modell (Bernhardsen og Larsen, 2007). I tillegg ble det estimert modeller ved å benytte de tradisjonelle statistiske metodene GLM, LDA og QDA, som alle er omtalt i kapittel 2. Det finnes svært mange forskjellige maskinlæringsteknikker, og i tillegg hybrider hvor forskjellige maskinlæringsteknikker kombineres. I denne oppgaven er det valgt å estimere forskjellige modeller ved bruk av SVM og forskjellige versjoner av nevrale nettverk, da disse er hyppigst benyttet i artiklene omtalt i Ravi Kumar og Ravi (2007) og Kirkos (2015). Det ble også estimert modeller ved bruk av maskinlæringsteknikkene KNN og CT.

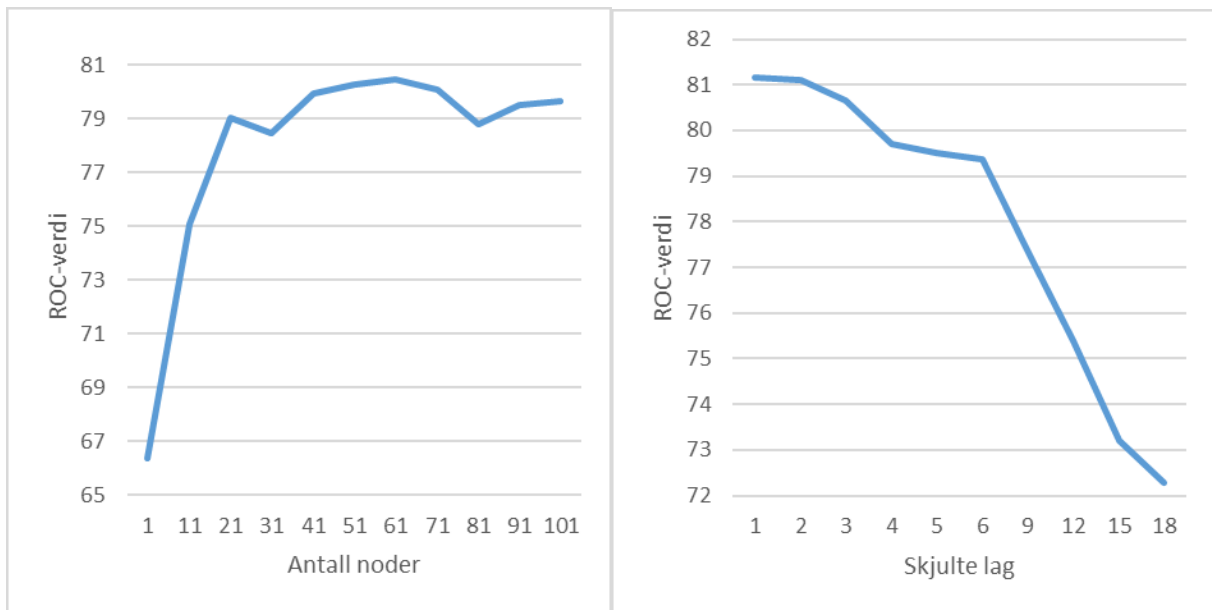
Tabell 10: Metoder valgt til estimering av modeller

Tradisjonelle statistiske metoder
Generell Lineær Modell (GLM)
Generell Additiv Modell (GAM)
Lineær diskriminantanalyse (LDA)
Kvadratisk diskriminantanalyse (QDA)
Maskinlæringsteknikker
Tolags forovermatet nevralt nettverk (NN)
k -nærmeste naboer (KNN)
Støttevektormaskin (SVM)
Klassifiseringstrær (CT)
Dypt nevralt nettverk skapt av to autoencodere (DNNA)
Dypt nevralt nettverk med back-propagation (DNNB)

All estimering og testing med disse ti forskjellige metodene ble utført ved bruk av Matlab, og nedenfor blir det diskutert valg av parametere for maskinlæringsteknikkene. Kriteriet for hvilke parametere som ble valgt var ROC-verdi, som er nærmere forklart i kapittel 3.3.2. Parameterne ble valgt basert på hvilke kombinasjoner av dem som ga høyest ROC-verdi ved estimering av modell basert på variabelsett 1 og oppgavens datautvalg. Etter dette ble det også gjort tilsvarende tester for modeller basert på variabelsett 2 og 3, og det viste seg at parameterne som ga høyest ROC-verdi var tilnærmet de samme også for modeller basert på disse variabelsettene. Nedenfor er det gjengitt hvilke parametere som ble testet, samt hvilke som estimerte best for modeller basert på variabelsett 1, om ikke annet er oppgitt.

Parametervalg for NN

For denne maskinlæringsteknikken ble det testet for antall skjulte lag og antall noder. Det viste seg at ROC-verdi er høyest ved ett skjult lag. Videre viste det seg at ROC-verdi øker med antall noder. Effekten av antall noder er stor i starten, før den så er avtagende. I denne oppgaven blir det valgt å benytte 1 skjult lag og 60 noder ved estimering av modeller estimert ved NN og basert på alle tre variabelsett. Se tabell 20 i Appendiks 4 for ROC-verdier for ulike antall skjulte lag og noder.



Figur 6: Parametervalg for nevrale nettverk (NN), basert på variabelsett 1.
 Til venstre: gjennomsnittlig ROC-verdi for 1, 2, 3, 4, 5, 6, 9, 12, 15 og 18 skjulte lag, som funksjon av antall noder.
 Til høyre: gjennomsnittlig ROC-verdi for 1, 11, 21, 31, 41, 51, 61, 71, 81, 91 og 101 noder, som funksjon av skjulte lag.

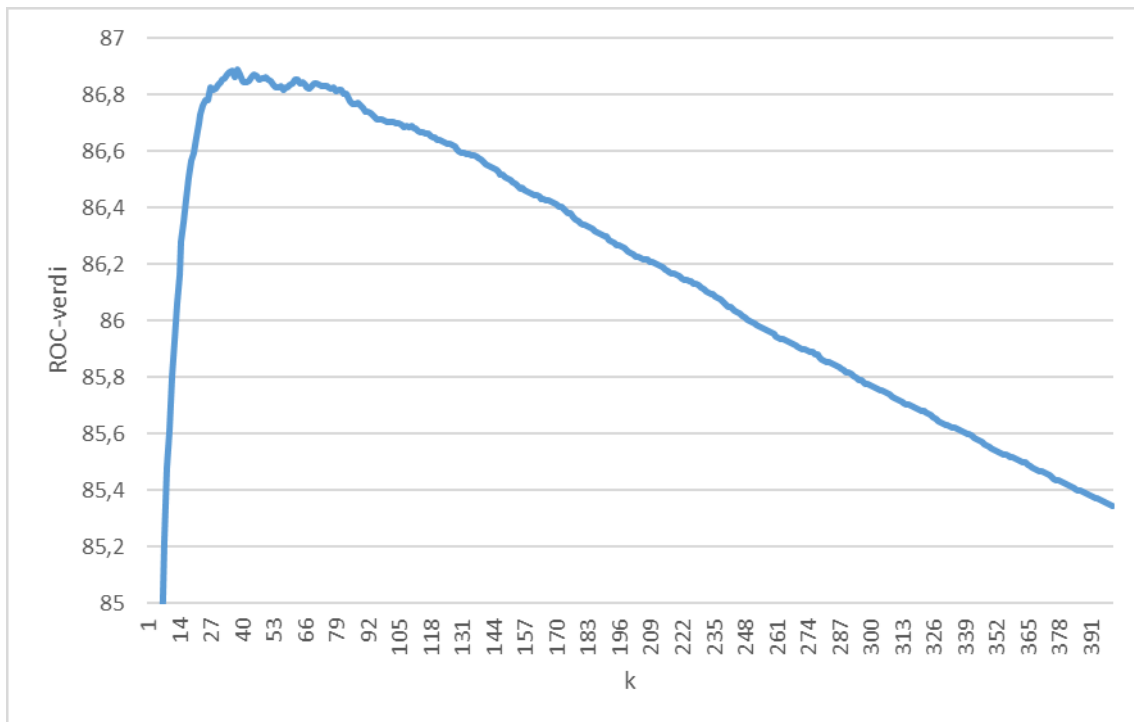
Parametervalg for KNN

For denne metoden ble det testet for alle kombinasjoner av verdiene på parametere som vist i tabell 11.

Tie-breaking-algoritme vil si hvordan metoden vekter observasjoner dersom de scorer likt ved avstandsberegningen. Det kom frem at parameterverdien «nærmeste» for tie-breaking-algoritme ga beste estimering av variabelsettet. Som avstandsberegning viste seg at euklidisk og Minkowski-avstand ga best estimering. Ved estimering ved KNN er det i denne oppgaven valgt å benytte euklidisk avstand.

Tabell 11: Parametere som ble testet for metode KNN.

Parameter	Verdier som ble testet
Antall nærmeste naboer (k)	Alle verdier mellom 1 og 100
Tie-breaking-algoritme	minst, nærmest, tilfeldig
Avstandsberegning	chebychev, correlation, euklidisk, hamming, jaccard, minkowski, spearman, cosine



Figur 7: KNN: ROC-verdi som funksjon av antall k ved estimering av modell basert på variabelsett 1.

Figur 7 viser ROC-verdi langs Y-aksen, og antall nærmeste naboer (k) langs X-aksen. Her kommer det frem at ROC-verdi stiger med antall k til og med 38, for så å synke med antall k etter dette. I denne oppgaven blir det derfor benyttet 38 k når metoden KNN benyttes. Se også tabell 21 og tabell 22 i Appendiks 4 for ROC-verdi og treffprosent ved ulike antall k og avstandsberegninger.

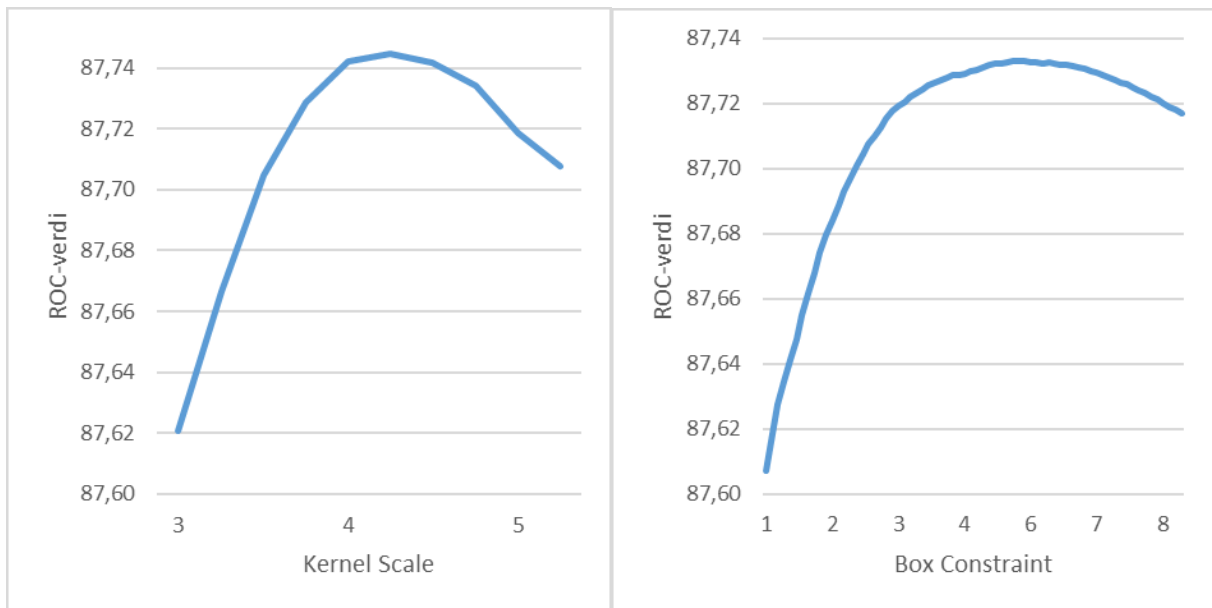
Parametervalg for SVM

For denne metoden ble det testet for alle kombinasjoner av verdiene på parametrene som vist i tabell 12.

Det viste seg at en gaussisk kernel-funksjon på følgende formel ga estimerte modeller med best ROC-verdi: $e^{-\|x_1 - x_2\|^2}$. Det er derfor denne som er valgt ved estimering ved SVM i denne oppgaven.

Tabell 12: Parametre som ble testet for metode SVM.

Parameter	Mulige verdier
Kernel-funksjon	gaussisk, lineær, polynomisk
Box constraint	Verdier mellom 1 og 100
Kernel scale	Verdier mellom 1 og 10



*Figur 8: Parametervalg for støttevektormaskin (SVM), basert på variabelsett 1.
Til venstre: gjennomsnittlig ROC-verdi for box constraint mellom 1,1 og 8,0,
som funksjon av kernel scale.
Til høyre: gjennomsnittlig ROC-verdi for kernel scale mellom 3,0 og 5,0,
som funksjon av box constraint.*

Fra de to grafene i figur 8 kommer det frem at ROC-verdi øker med både *box constraint* og *kernel scale* til en gitt verdi, for så å falle. Toppunktene i grafene er ved *box constraint* lik 5,3 og *kernel scale* lik 4,0. Det er derfor disse verdiene som ble valgt som parametere ved estimeringer av modeller ved SVM i denne oppgaven. Se tabell 23 i Appendiks 4 for ROC-verdi ved ulike verdier for antall *box constraint* og *kernel scale*.

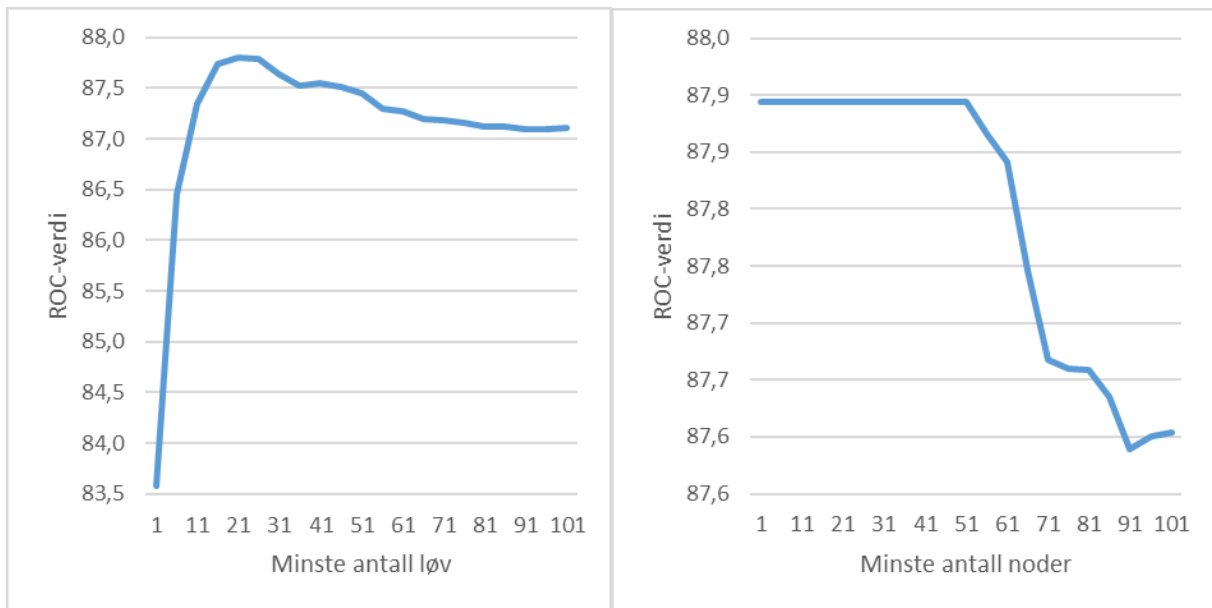
Parametervalg for CT

For denne metoden ble det testet for alle kombinasjoner av verdiene på parametere som vist i tabell 13.

Det viste seg at en symmetrisk score transform-funksjon på formen $2x - 1$ ga beste ROC-verdi, så det er denne som er valgt ved estimeringer i denne oppgaven. Tabell 24 i Appendiks 4 viser ROC-verdi for modeller estimert ved CT for ulike score transform-funksjoner.

Tabell 13: Parametere som ble testet for metode CT.

Parameter	Mulige verdier
Score transform-funksjon	ingen, dobbellogaritmisk, logaritmisk, ismax, symmetrisk, symmetrisk-logisk, symmetrisk-ismax
Minste antall løv	Verdier mellom 1 og 100
Minste antall noder	Verdier mellom 1 og 10

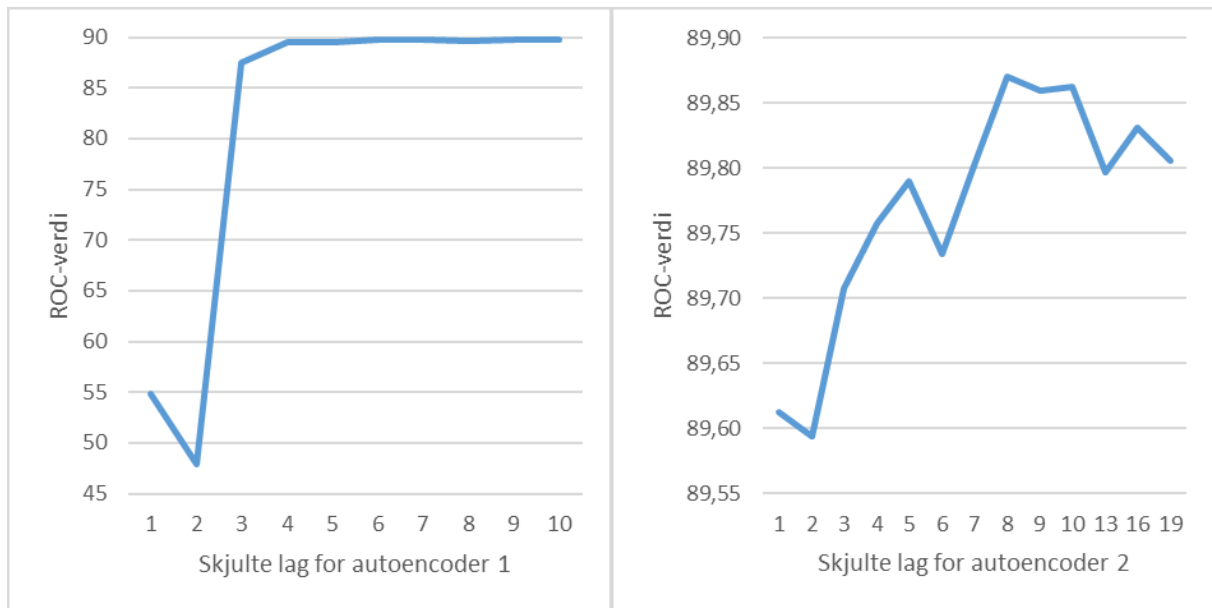


Figur 9: Parametervalg for klassifiseringstrær (CT), basert på variabelsett 1. Til venstre: gjennomsnittlig ROC-verdi for minste antall noder lik 1, 6, 11, 16, 21, 26, 31, 36, 41, 46, 51, 56, 61, 66, 71, 76, 81, 86, 91, 96 og 101, som funksjon av minste antall løv. Til høyre: ROC-verdi ved minste antall løv lik 26, som funksjon av minste antall noder.

Som figur 9 viser, ble høyest ROC-verdi oppnådd ved 26 som verdi på parameteren minste antall løv. Videre viste det seg at det var likegyldig hvilken verdi som ble angitt som minste antall noder, så lenge den var under en gitt verdi. I denne oppgaven ble det med bakgrunn i dette valgt å benytte 26 som minste antall løv og 1 som minste antall noder ved estimeringer av modeller ved CT. Se tabell 25 i Appendiks 4 for ROC-verdi ved ulike verdier for minste antall løv og minste antall noder.

Parametervalg for DNNA

For denne metoden ble det testet for antall skjulte lag for de to autoencoderne, henholdsvis autoencoder 1 og autoencoder 2. Som det fremgår av figur 10 viste det seg at økt antall skjulte lag for autoencoder 1 hadde positiv effekt på ROC-verdi til en gitt verdi, for så å ha minimal effekt. For antall skjulte lag for autoencoder 2 viste det seg at 8 var det optimale antallet, slik det fremgår av figur 10. Se tabell 26 i Appendiks 4 for ROC-verdier for ulike antall skjulte lag for autoencoder 1 og autoencoder 2.



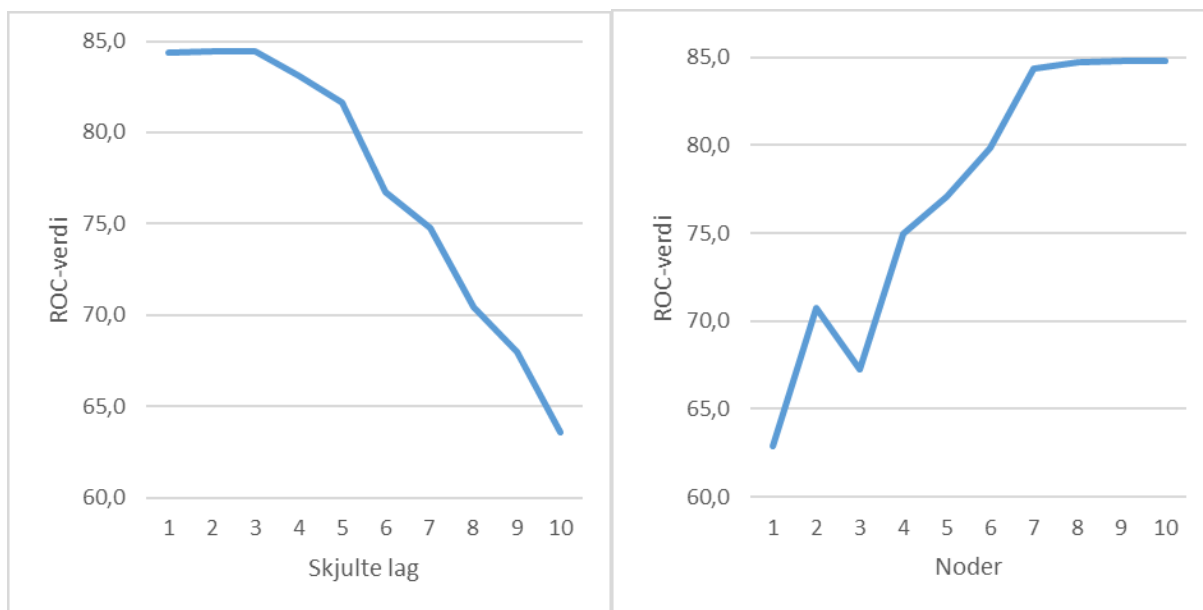
Figur 10: Parametervalg for dypt nevraitt nettverk skapt av to autoencodere (DNNA), basert på variabelsett 3.

Til venstre: gjennomsnittlig ROC-verdi ved 1-10 skjulte lag for autoencoder 2, som funksjon av skjulte lag for autoencoder 1.

Til høyre: ROC-verdi for 10 skjulte lag for autoencoder 1, som funksjon av skjulte lag for autoencoder 2.

Parametervalg for DNNB

For denne metoden ble det testet for antall skjulte lag og antall noder. Det viser seg at ROC-verdi ble høyest ved ett skjult lag. Videre viser det seg at ROC-verdi økte med antall noder til en gitt verdi, for så å ha minimal effekt. I denne oppgaven ble det valgt å benytte 1 skjult lag og 10 noder når det ble estimert modeller ved bruk av DNNB. I tabell 27 i Appendiks 4 vises ROC-verdier for ulike antall skjulte lag og noder.



Figur 11: Parametervalg for dypt nevralt nettverk med back-propagation (DNNB), basert på variabelsett 1.

Til venstre: gjennomsnittlig ROC-verdi for 1-10 noder, som funksjon av antall skjulte lag.
 Til høyre: gjennomsnittlig ROC-verdi for 1-10 skjulte lag, som funksjon av antall noder.

3.3 Evaluering

Ved evaluering av modellene i denne oppgaven ble det valgt å benytte treffprosent og ROC-verdi, som begge blir diskutert i dette delkapitlet. Det kan hevdes at to modeller predikerer forskjellig dersom de gir forskjellig treffprosent og ROC-verdi. Ved å utføre en McNemars test, som også blir diskutert i dette delkapitlet, kan det i tillegg anslås om forskjellen er signifikant.

3.3.1 Treffprosent

Treffprosent er andel av totalt hendelser som modellen predikerer korrekt. Jo høyere treffprosent, jo bedre er modellen. Denne måten å evaluere modeller på er enkel å forstå, og den krever heller ingen spesielle forkunnskaper om statistikk. Samtidig kan treffprosent i noen tilfeller være et mindre godt evalueringsmål, da det hverken tar høyde for klassefordelingen i datautvalget eller feilkostnader. På grunn av dette er det i denne oppgaven også valgt å benytte ROC-verdi som evalueringsmål, noe som blir utdypet nedenfor.

3.3.2 Receiver Operating Characteristics (ROC)

ROC er en forkortelse for *receiver operating characteristics*, og er en teknikk for å visualisere, organisere og kategorisere hendelser med kun to mulige utfall (Fawcett, 2006). Alle modeller i denne oppgaven klassifiserer hendelser til én av to verdier; «konkurs» eller «ikke konkurs», hvor førstnevnte anses som *positiv* og sistnevnte som *negativ*. Hver hendelse modellen predikerer kan plasseres i en av de fire utfallene som vist i tabell 14.

Tabell 14: Mulige klassifiseringsutfall; hypotetisk og faktisk (Fawcett, 2006)

	p	n
Y	Ekte positiv	Falsk positiv
N	Falsk negativ	Ekte negativ

Dersom modellen predikerer en hendelse som positiv, og den faktisk er positiv, så plasseres observasjonen i utfallet *ekte positiv*. Tilsvarende plasseres hendelsen i *ekte negativ* dersom modellen predikerer den som negativ og den faktisk også er det. *Falsk positiv* er en hendelse som blir predikert positiv, men som faktisk er negativ. I slike tilfeller gjør modellen en såkalt type-1-feil. Tilsvarende er *falsk negativ* hendelser som blir predikert negativ, men som faktisk er positive. I slike tilfeller gjør modellen en såkalt type-2-feil. I denne oppgavens resultater presenteres det stolpediagrammer for alle modeller som viser andel av positive hendelser som er ekte eller falske, og andel negative som er ekte eller falske. Dette blir illustrert i grafer med samme fargekoder som de i tabell 14, som dermed gir et bilde på type-1-feil og type-2-feil som samtlige modeller gjør.

En ROC-graf er en graf som er plottet i et diagram hvor y-aksen representerer ekte positivrate og x-aksen representerer falsk positivrate. *Ekte positivrate* er angitt ved

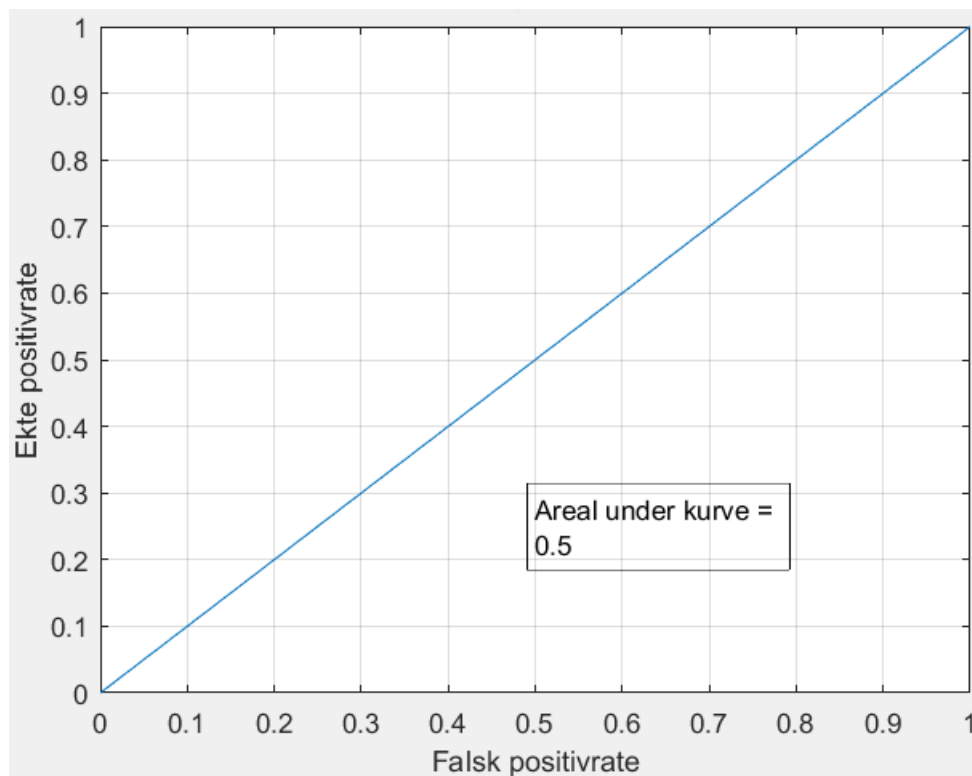
$\frac{\text{ekte positive hendelser}}{\text{sum ekte positive og falske positive hendelser}}$, mens *falsk positivrate* er angitt ved

$\frac{\text{falske negative hendelser}}{\text{sum falske negative og ekte negative hendelser}}$. En ROC-verdi er arealet under ROC-grafen, og er lik

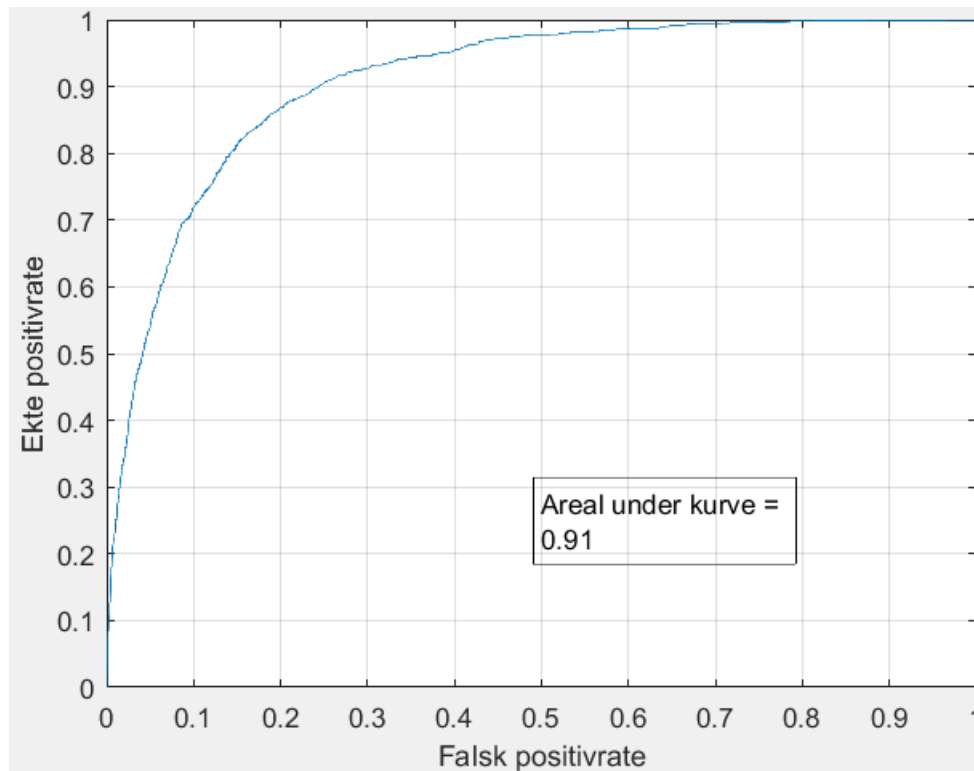
sannsynligheten for at en tilfeldig valgt positiv klassifisering klassifiseres høyere enn en tilfeldig valgt negativ klassifisering (Fawcett, 2006). Ved å tilfeldig klassifisere hendelser vil det oppnås en ROC-graf av formen $Y = X$ slik som illustrert i Figur 12. Dette er en tilfeldig klassifisering som generer en ROC-verdi på 0,5, og siden linjen også er rett forventes det at modellen predikerer riktig på halvparten av de negative og halvparten av de positive hendelsene. Derimot betyr en ROC-verdi på 1 at modellen predikerer perfekt. Jo lengre mot

det øverste venstre hjørnet ROC-grafen er, jo høyere blir ROC-verdien og jo bedre predikerer variabelsettet. Figur 13 viser et eksempel på en ROC-kurve gitt av en modell som blir presentert senere i oppgaven.

Treffprosent er kanskje enklest og mest intuitivt som evalueringsmål, mens ROC-verdi hevdes å være den beste endimensjonale skala for å måle en modells evne til å predikere riktig (Jin og Ling, 2005). En styrke med ROC-grafen og ROC-verdien er at de visualiserer og kategoriserer uten å ta hensyn til klassefordelingen og feilkostnader (Fawcett, 2006). Denne egenskapen er spesielt viktig når man arbeider med datasett som har en skjev fordeling og således ikke er normalfordelt. Det er også viktig når man jobber med kostnadssensitiv læring slik som i denne oppgaven. I denne oppgaven er det derfor valgt å benytte ROC-verdi som mål på hvor godt en modell predikerer.



Figur 12: ROC-kurve og ROC-verdi ved tilfeldig klassifisering.



Figur 13: Eksempel på ROC-kurve og ROC-verdi fra modell basert på variabelsett 3 og estimert med metode GAM.

At ROC-verdi kan være et bedre evalueringsmål enn treffprosent illustreres også i denne oppgavens resultater, gjengitt i kapittel 4 og i Appendiks 7. I denne oppgaven er det, som senere omtalt, valgt å estimere modeller basert på et datautvalg av 31 185 årsregnskaper, hvorav 33% er fra selskaper klassifisert som konkurs og 67% er klassifisert som ikke konkurs. For modeller som estimeres basert på dette datautvalget oppnås det en treffprosent på opp til 84% og en ROC-verdi på 91%. Baseres modellene imidlertid på et datautvalg av 738 045 årsregnskaper, hvorav kun 1,4% er fra selskaper klassifisert som konkurs og resterende 98,6% er klassifisert som ikke konkurs, oppnås en treffprosent på hele 99%. ROC-verdi er allikevel nesten uendret og har verdi på opp til 91%. Årsaken til denne unaturlige høye treffprosenten er skjevdeling i datautvalget mellom antallet årsregnskaper fra selskaper som er klassifisert som konkurs og antallet som er klassifisert som ikke konkurs. Dette fører til en høy treffprosent på bekostning av at modellen gjør mange type 1-feil. ROC-verdi tar imidlertid høyde for denne skjevdelingen, og gir derfor en tilnærmet lik verdi både med og uten skjevdeling.

3.3.3 McNemars test

I Dietterich (1998) presenteres fem forskjellige statistiske tester som kan avgjøre om en metode er bedre enn en annen. Av disse fem anbefales enten en såkalt «5 x 2cv t test» eller McNemars test (McNemar, 1947, Gillick og Cox, 1989). «5 x 2cv t test» presterer marginalt bedre, men krever mer datakraft og setter i tillegg begrensninger på antall observasjoner som kan benyttes ved estimering av modellene. På grunn av dette blir det i denne oppgaven valgt å benytte McNemars test. Dette er samme valg som blir gjort i Næss (2015), hvor McNemars test også benyttes for å avgjøre hvilken klassifiseringsmetode som presterer best.

Anta at raten for antall feilklassifiseringer ved en metode, benyttet på et datautvalg, er gitt ved $e_1 = \frac{M_1}{n}$, hvor n er antall observasjoner i datautvalget og M_1 er antall feilklassifiseringer blant disse. Raten for antall feilklassifiseringer ved en annen metode, benyttet på samme datautvalg, er gitt ved $e_2 = \frac{M_2}{n}$. I denne oppgaven benyttes McNemars test for å teste nullhypotesen om $e_1 = e_2$.

Tabell 15: Kategoriseringen ved McNemars test

		Metode 2	
		Riktig klassifisert	Feil klassifisert
Metode 1	Riktig klassifisert	A	B
	Feil klassifisert	C	D

I tabell 15 er A antall observasjoner som er riktig klassifisert av både metode 1 og 2, mens D er antall observasjoner som er feil klassifisert av både metode 1 og 2. C er antall observasjoner som er feil klassifisert av metode 1, men riktig klassifisert av metode 2. B er antall observasjoner som er riktig klassifisert av metode 1, og feil klassifisert av metode 2.

Tabell 15 minner mye på tabell 14 fra delkapittel 3.3.2. Det som skiller disse er at tabell 15 visualiserer hendelser som klassifiseres forskjellig mellom to modeller, mens tabell 14 visualiserer forskjeller mellom hvordan kun én modell klassifiserer hendelser i forhold til riktig klassifisering.

McNemars test kan på forenklet form skrives slik (Gillick og Cox, 1989):

$$W = \frac{|B - C| - 1}{\sqrt{B + C}} \quad (8)$$

B og C er som forklart ovenfor. Årsaken til at disse benyttes i formelen, og ikke A eller D , er at summen av B og C er de observasjonene hvor modellene predikerer forskjellig. Fra denne

formelen blir p -verdien beregnet av $p = 2P(Z \geq \omega)$, hvor $Z \sim N(0,1)$ og ω er den faktiske verdien av W . I denne oppgaven blir nullhypotesen om at $e_1 = e_2$ forkastet dersom p -verdien er under et signifikansnivå på 5%. Dersom p -verdien er lavere enn 5% kan man altså hevde signifikant forskjell mellom metodene.

3.3.4 Bruk av McNemars test, ROC-verdi og treffprosent i tidligere arbeid

Av artiklene omtalt i Ravi Kumar og Ravi (2007) er det kun 2 som benytter seg av McNemars test, noe som utgjør 1,7%. Blant artiklene omtalt i Kirkos (2015) er det kun 4 som anvender McNemars test, noe som utgjør 9,5%. Dette tyder på at tidligere forskning har konkludert med forskjeller mellom modeller for konkursprediksjon, selv om det ikke er testet for signifikant forskjell med McNemars test. At det i denne oppgaven benyttes McNemars test er derfor trolig en styrke som i liten grad er tilstede i tidligere forskning.

Omtrent alle artiklene omtalt i Ravi Kumar og Ravi (2007) og Kirkos (2015) benytter treffprosent til evaluering av modellene. Men kun 4 artikler i de to oversiktsartiklene benytter ROC-verdi, noe som utgjør 2,4% av alle artiklene. ROC-verdi blir allikevel benyttet i denne oppgaven, da dette evalueringsmålet anses som bedre enn treffprosent (Jin og Ling, 2005) og det benyttes av Norges Bank (Bernhardsen og Larsen, 2007). At ROC-verdi benyttes i denne oppgaven er trolig også en styrke som i liten grad er tilstede i tidligere forskning, i alle fall blant forskning presentert i Ravi Kumar og Ravi (2007) og Kirkos (2015). Dersom formålet med denne oppgaven kun hadde vært høyest mulig treffprosent, for enhver pris, kunne denne oppgaven ha konkludert med treffprosent på hele 99%, som omtalt i kapittel 3.3.2 og vist i Appendiks 7.

3.4 Matlab-kode og repliserbarhet

Matlab-kode som er benyttet i denne oppgaven er åpen, og tilgjengelig via <http://masterthesis.ranik.no/>.

I Chang og Li (2015) blir det påpekt at forskning innenfor fagfeltet økonomi ikke alltid er repliserbart. På grunn av dette er det i denne oppgaven gjort følgende grep som omtalt i artikkelen og som i tillegg anses som relevant her:

- Readme-filen i koden inneholder hvilket operativsystem og programvareversjon som ble benyttet ved analyse.

- Readme-filen klargjør hvilke filer som skal kjøres i hvilken rekkefølge for å oppnå resultatene som angitt i denne oppgaven.
- Det ble satt en standard verdi for «random seeds» for å klargjøre nummegeratoren i programvaren benyttet ved randomisering. Dette er omtalt i readme-filen.

3.5 Datagrunnlag

Det opprinnelige datagrunnlaget i denne oppgaven består av alle aksje- og allmennaksjeselskaper i Norge som leverte årsregnskaper for minst ett av årene fra og med 2005 til og med 2014. Dette ble levert av Proff Forvalt, som er en del av Eniro Norge AS.

Som tidligere nevnt blir det i denne oppgaven blant annet benyttet variabler som tar utgangspunkt i de to eksisterende variabelsettene Altmans Z-score-modell og Norges Banks SEBRA-modell. Ifølge Finanstilsynet (2013b) beregner Norges Banks SEBRA-modell sannsynlighet for mislighold av lån i løpet av kommende år basert på siste avlagte årsregnskap, og i tabell 5 kommer det frem at Altmans Z-score-modell predikerer konkurs best når regnskapstallene fra året før konkursen benyttes. Basert på dette er det i denne oppgaven valgt å analysere det siste leverte årsregnskapet, og ikke tidligere årsregnskaper, når analysen gjennomføres.

De fleste årsregnskapene i datagrunnlaget er oppgitt i norske kroner (NOK). I noen tilfeller er årsregnskapene oppgitt i annen valuta, som for eksempel USD, og i andre tilfeller er ikke valuta oppgitt i årsregnskapene. For variabelen e_{ie} i variabelsett 1 har det betydning hvilken valuta tallene er oppgitt i. Alle andre variabler med regnskapsdata, i alle variabelsettene, er brøker. Så disse er ikke avhengige av valuta, siden samme valuta alltid benyttes både i teller og nevner. På bakgrunn av dette, og at de aller fleste årsregnskaper er oppgitt i NOK, er det ikke gjort noe i denne oppgavens analyse for å omregne annen valuta til NOK.

I denne oppgaven blir ikke konsernregnskaper analysert. For alle morselskap i datagrunnlaget blir årsregnskapet til selskapet benyttet, og ikke årsregnskapet for konsernet. På den måten blir ikke regnskapstallene fra datterselskaper analysert to ganger, altså både i morselskapets konsernregnskap og i tillegg i datterselskapets egne selskapsregnskaper.

3.6 Definerings av konkursselskaper – avhengig variabel

Bernhardsen (2001) definerer den avhengige variabelen, konkurs, på følgende måte:

(...) the firm was registered bankrupt within 3 years and this year constitutes the last registered financial statement. (Bernhardsen, 2001)

På bakgrunn av dette, og funnene i delkapittelet ovenfor, blir variabelen «konkurs» definert på tilsvarende måte i denne oppgaven:

Variabelen «konkurs» får verdien 1 dersom begge følgende kriterier innfris:

1. det aktuelle årsregnskapet er det siste avlagte, og
2. selskapet blir erklært konkurs innen det 3. året etter året det aktuelle årsregnskapet gjelder for.

Om ikke får variabelen «konkurs» verdien 0.

3.7 Ekskludering av enkeltobservasjoner

Alle årsregnskaper med totale eiendeler under kr 500 000,- er ekskludert fra datagrunnlaget. Dette er i samsvar med hva som også er gjort av Bernhardsen og Larsen (2007).

Videre står det i Bernhardsen og Larsen (2007) at finansielle aksjeselskaper ikke er inkludert i deres datagrunnlag. På bakgrunn av dette er det i denne oppgaven valgt å ekskludere fra datagrunnlaget alle selskaper med NACE-koder som begynner på 64 (Finansieringsvirksomhet), 65 (Forsikringsvirksomhet og pensjonskasser, unntatt trygdeordninger underlagt offentlig forvaltning) eller 66 (Tjenester tilknyttet finansierings- og forsikringsvirksomhet). Dette er koder fra Standard for næringsgruppering (SN2007/NACE), som er gjeldende standard når denne oppgaven skrives.

I Finanstilsynet (2013b) står det at Norges Banks SEBRA-modell egner seg i mindre grad for å analysere eiendomsselskaper med stor andel av prosjekter under utvikling. På grunn av dette er det i denne oppgaven valgt å ekskludere fra datautvalget alle selskaper med NACE-kode som begynner på 41 (Oppføring av bygninger).

Det er verdt å merke seg at en del selskaper er registrert med NACE-kode 00.000 «Uoppgitt» i offentlige registre. I denne oppgaven er det ikke blitt kartlagt om disse reelt sett skulle hatt en NACE-kode med to første siffer som oppgitt ovenfor, og dermed burde vært ekskludert fra datautvalget.

3.8 Ekskludering av regnskapsår 2013 og 2014

I datagrunnlaget er det totalt 20 491 selskaper som har gått konkurs før 31.12.2015, og som samtidig har minst ett årsregnskap i perioden fra og med 2005 til og med 2014. Følgende tabell viser hvor mange av disse som ble erklært konkurs i et gitt år sammenlignet med siste avlagte årsregnskap.

Tabell 16: År mellom siste avlagte årsregnskap og år erklært konkurs for selskaper som har avlagt minst ett årsregnskap i perioden 2005-2014.

	Antall	Andel
Erklært konkurs samme år som siste avlagte årsregnskap	7	0,03%
Erklært konkurs året etter siste avlagte årsregnskap	5 664	27,64%
Erklært konkurs 2. året etter siste avlagte årsregnskap	13 415	65,47%
Erklært konkurs 3. året etter siste avlagte årsregnskap	1 362	6,65%
Erklært konkurs 4. året etter siste avlagte årsregnskap	33	0,16%
Erklært konkurs 5. året etter siste avlagte årsregnskap	6	0,03%
Erklært konkurs 6. året etter siste avlagte årsregnskap	4	0,02%
Totalt	20 491	100,00%

Alle selskaper som er konkurs per dato for uttrekk (2016-03-21) i oppgavens datautvalg, og som er erklært konkurs innen 3. året etter siste avlagte årsregnskap, utgjør 99,79% av totalen. Datagrunnlaget i denne oppgaven ble hentet fra Brønnøysundregistrene 21. mars 2016, så alle konkurserklæringer som er registrert i 2015 eller tidligere er med i datagrunnlaget. 2015 utgjør 3. året etter avlagte årsregnskaper 2012. På bakgrunn av dette funnet er det i denne oppgaven ikke valgt å analysere regnskapsårene 2013 og 2014.

Bernhardsen (2001) fant i sitt studie av årsregnskaper fra perioden 1990-1999 at tilsvarende 100% av alle selskapene var erklært konkurs innen 3. året etter siste avlagte årsregnskap. Også i videreutviklingen av SEBRA-modellen av Bernhardsen og Larsen (2007) kommer det frem at tester og retester kun kan utføres på regnskaper 2-3 år i forkant av siste avlagte årsregnskap. Begge disse studiene støtter altså opp om utelatelsen av analyse av årsregnskaper fra regnskapsårene 2013 og 2014 i denne oppgaven.

3.9 Datautvalg og kryssvalidering

Etter ekskluderinger som nevnt i kapittel 3.7 og 3.8 besto datautvalget av totalt 1 127 850 årsregnskaper i perioden 2005 til 2012. Av disse defineres 10 395 som konkurs, mens resterende defineres som ikke konkurs.

Samtlige som ble definert til konkurs ble inkludert i datautvalget som analysen ble utført på. I tillegg ble dobbelt så mange definert til ikke konkurs tatt med i datautvalget, slik som det også ble gjort i Boyacioglu et al. (2009). Dermed besto datautvalget av $\frac{1}{3}$ konkursselskaper og $\frac{2}{3}$ ikke-konkursselskaper. Matlab valgte tilfeldig hvilke ikke-konkursselskaper som ble inkludert i datautvalget.

Estimering og testing av modellene i denne oppgaven ble utført ved å benytte en fem-delt kryssvalidering, hvor datautvalget da deles inn i fem deler (Kirkos, 2015). Hvordan denne oppgavens datautvalg er inndelt er illustrert i tabell 37 i Appendiks 8. Modellene estimeres ved å benytte fire av delene, og så testes på den femte. Estimering og testing av modellene gjøres til sammen fem ganger, hvor hver av delene på tur er den delen modellen testes på, mens resterende fire deler benyttes til estimering. Til slutt blir gjennomsnittlig ytelse av de fem modellene beregnet, og det er denne gjennomsnittlige ytelsen som blir presentert i oppgaven. Denne formen for validering anbefales blant annet i artikkelen av Kohavi (1995), hvor det fremgår at en kryssvalidering er bedre enn flere andre valideringsmetoder.

Hver del besto av $\frac{1}{3}$ konkursselskaper og $\frac{2}{3}$ ikke-konkursselskaper, altså samme andeler som hele datautvalget. Matlab valgte tilfeldig hvilke selskaper som ble med i hver del.

Ingen av de fem delene inneholder de samme observasjonene, så derfor vil heller ikke de deler som benyttes til estimering inneholde samme data som benyttes til validering. Dette bidrar til å unngå overtilpassing (Kirkos, 2015).

I tillegg til estimering av modeller basert på årsregnskaper i perioden fra 2005 til 2012, ble det også estimert modeller basert på årsregnskapene for hvert av de åtte årene enkeltvis. Det ble funnet at prediksjonsevnen til modeller estimert på årsregnskaper for hvert år enkeltvis ikke skilte seg nevneverdig fra modeller estimert på årsregnskaper for alle årene fra 2005 til 2012 samlet. Det er derfor valgt å kun presentere og analysere modeller estimert med årsregnskaper fra alle årene fra 2005 til 2012.

Som nevnt er altså samtlige konkursselskaper inkludert, samt dobbelt så mange ikke-konkursselskaper. Dette utgjorde da et datautvalg på totalt 31 185 selskaper, hvorav 10 395 er klassifisert som konkurs og 20 790 klassifisert som ikke konkurs. Det ble også estimert modeller med 3, 10 og 70 ganger så mange ikke-konkursselskaper, noe som ga datautvalg på henholdsvis 41 580, 114 345 og 738 045 selskaper. Med et slikt større datautvalg ble treffprosenten høyere, men på bekostning av at modellen gjorde flere type 1-feil, som innebærer at modellen predikerte flere selskaper som ikke konkurs, mens de i realiteten gikk

konkurs. ROC-verdi var stort sett tilnærmet uendret, uavhengig av antall selskaper i datautvalget, noe som også støtter opp om påstanden om at ROC-verdi er et bedre evalueringsmål enn treffprosent, slik som også drøftet i delkapittel 3.3. Siden ROC-verdi er uendret, og tidligere forskning (Boyacioglu et al., 2009) benytter dobbelt så mange ikke-konkursselskaper som konkursselskaper, er tilsvarende også valgt i denne oppgaven. Resultater for modeller estimert med et datautvalg bestående av 41 580, 114 345 og 738 045 selskaper i datautvalget er allikevel presentert i Appendiks 7.

Transformering av uavhengige variabler

Ved estimering av modellene i oppgaven ble det forsøkt å transformere alle ikke-dummy uavhengige variabler til standardiserte variabler (Zhao og Zhang, 2014). Dette fordi verdiene på de uavhengige variablene da blir mer sammenlignbare, og dermed er sjansen mindre for at noen uavhengige variabler får høyere betydning enn andre ved estimering av modellene.

Følgende formel er benyttet ved transformering til standardiserte variabler:

$$\text{standardisert variabel}_i = \frac{x_i - E(X)}{SD(X)} \quad (9)$$

standardisert variabel_{*i*} er den standardiserte verdien til den uavhengige variabelen, x_i er den opprinnelige observerte variabelen, $E(X)$ er gjennomsnittet av alle observerte verdier for den uavhengige variabelen og $SD(X)$ er standardavviket til den uavhengige variabelen.

For modeller basert på variabelsett 3 viste det seg at fire av denne oppgavens seks benyttede maskinlæringsteknikker estimerte bedre modeller dersom de uavhengige variablene var transformert til standardisert verdi. For de resterende to maskinlæringsteknikkene og alle fire tradisjonelle statistiske metodene var det ingen endring dersom datautvalget ble transformert til standardisert verdi før estimering av modeller.

I denne oppgaven er det derfor valgt å transformere datautvalget ved estimering av modeller basert på variabelsett 3, men ikke for modeller basert på variabelsett 1 eller 2. For ordens skyld er det i Appendiks 5 presentert resultatene for modeller basert på variabelsett 3 når datautvalget på forhånd ikke er transformert til standardisert verdi.

4 Resultater, analyse og drøfting

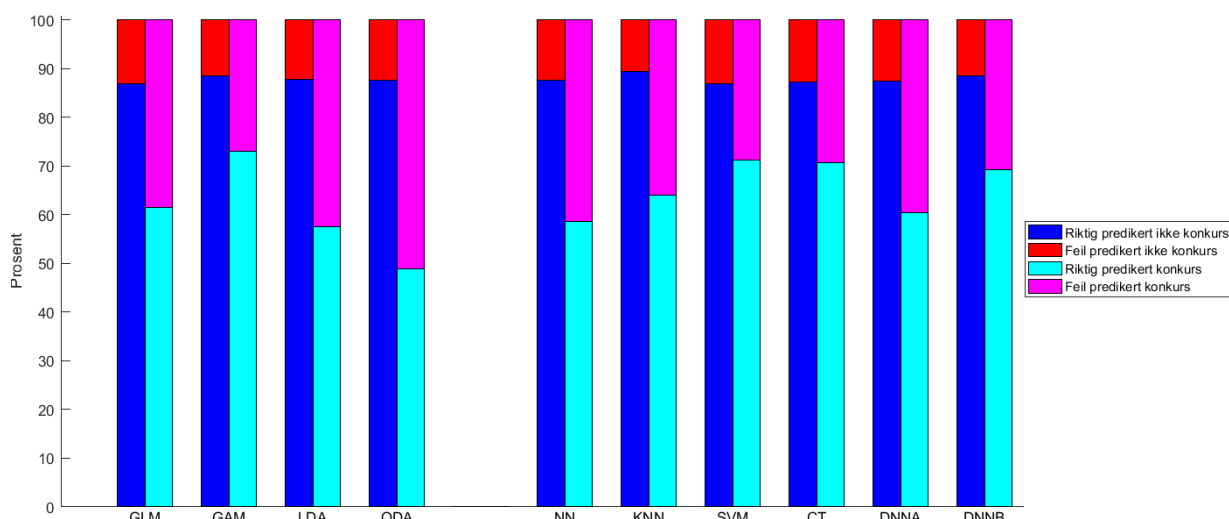
I dette kapitlet blir resultatene fra alle estimerte modeller presentert og sammenlignet. Det vil her komme frem hvilke metoder som estimerer bedre enn andre, og hvilke variabelsett som egner seg best. Resultatene for modellene basert på variabelsett 1, 2 og 3 vil bli presentert nedenfor i egne delkapitler, henholdsvis 4.1, 4.2 og 4.3. Prediksjonsevnen til modellene blir evaluert ved ROC-verdi og treffprosent, slik som beskrevet i kapittel 3. Imidlertid blir ROC-verdi mest vektlagt da denne anses som et bedre evalueringsmål enn treffprosent, som drøftet i delkapittel 3.3.2 og vist i Appendiks 7.

4.1 Resultater fra variabelsett 1 – basert på SEBRA-modellen

I dette delkapitlet blir resultater fra ti ulike modeller presentert. Alle disse ti er basert på variabelsett 1, og estimert ved denne oppgavens ti valgte metoder.

Figur 14 viser hvor bra de forskjellige modellene predikerer konkurs og ikke konkurs. Mørkeblå stolpe viser prosentvis riktig predikert ikke konkurs, mens lyseblå stolpe viser prosentvis riktig predikert konkurs. Fargene i figuren representerer det samme som i tabell 14. Feilklassifiserte ikke konkurs betegnes dermed som type 2-feil og er representert ved rød, mens feilklassifiserte konkurs betegnes som type 1-feil og har fargen rosa.

Det kommer også frem fra figur 14 at de ulike modellene predikerer ikke konkurs riktig i omtrent 85 til 90% av tilfellene, mens de predikerer konkurs riktig i 50% til 70% av tilfellene. Det kommer tydelig frem fra figuren at samtlige modeller gjør flere type 2-feil enn type 1-feil, altså at de i flere tilfeller klassifiserer konkursselskaper som ikke konkurs, enn ikke-konkursselskaper som konkurs.



Figur 14: Feilklassifiseringsrate for modeller basert på variabelsett 1

I tabell 17 presenteres oppnådd ROC-verdi og treffprosent for de ti modellene, som alle er basert på variabelsett 1 og estimert ved denne oppgavens ti valgte metoder. Blant de tradisjonelle statistiske metodene er det modellen ved GAM som oppnår best prediksjon med en ROC-verdi på 90% og en treffprosent på 83%. QDA estimerer dårligst modell blant de tradisjonelle statistiske metodene, med en ROC-verdi på 76% og en treffprosent på 75%. Av de nyere maskinlæringsteknikkene er det SVM, CT og DNNB som estimerer modeller som predikerer best, med en ROC-verdi på 88%-89% og treffprosent på 81%-82%. Bernhardsen og Larsen (2007) fra Finansmarkedsavdelingen i Norges Bank fikk på sine modeller, SEBRA-basis og SEBRA-utvidet, en ROC-verdi på henholdsvis 88% og 89%. I denne oppgaven oppnås altså en prediksjonsevne som er minst like god når GAM og variabelsett 1 benyttes.

Tabell 17: ROC-verdier og treffprosent for modeller basert på variabelsett 1

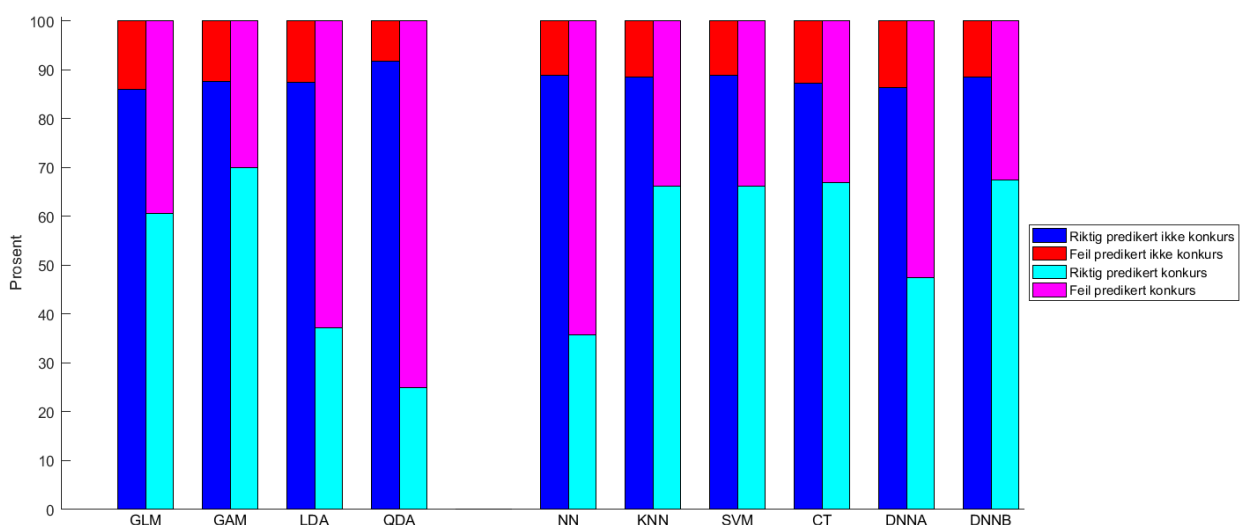
Tradisjonelle statistiske metoder	ROC-verdi	Treffprosent
Generell Lineær Modell (GLM)	82,7%	78,3%
Generell Additiv Modell (GAM)	89,6%	83,3%
Lineær diskriminantanalyse (LDA)	81,3%	77,7%
Kvadratisk diskriminantanalyse (QDA)	75,5%	74,7%
Maskinlæringsteknikker	ROC-verdi	Treffprosent
Tolags forovermatet nevralt nettverk (NN)	81,9%	78,0%
<i>k</i> -nærmeste naboer (KNN)	86,9%	81,0%
Støttevektormaskin (SVM)	87,8%	81,7%
Klassifiseringstrær (CT)	87,9%	81,7%
Dypt nevralt nettverk skapt av to autoencodere (DNNA)	83,2%	78,4%
Dypt nevralt nettverk med back-propagation (DNNB)	88,5%	82,1%

Tabell 17 viser at det er forskjeller mellom metodene ved estimering av modeller basert på variabelsett 1. For å finne ut om forskjellene er signifikante må det gjennomføres flere McNemars tester. Ved gjennomføringen av McNemars test ble det etablert hypoteser, der nullhypotesen hevdet ingen signifikant forskjell mellom prediksjonsevnen til modellene, mens alternativhypotesen hevdet signifikant forskjell. Signifikansnivået ble i denne oppgaven satt til 5%, noe som innebærer at ved en p-verdi under 5% forkastes nullhypotesen og det hevdes signifikant forskjell. Det presiseres at McNemars test ikke sier noe om hvor stor forskjellen er, men kun at det er signifikant forskjell.

Resultatene fra McNemars test følger i tabell 29 i Appendiks 6. Her kommer det frem at GAM er signifikant bedre enn alle de andre tradisjonelle statistiske metodene: GLM, LDA og QDA. Imidlertid er GAM kun signifikant bedre enn to av de seks maskinlæringsteknikkene: NN og DNNA. Mellom maskinlæringsteknikkene kunne ingen av de tre best egnede vise til signifikant forskjell i prediksjonsevne. DNNB, som er den metoden som gir best prediksjonsevne blant maskinlæringsteknikkene, er likevel signifikant bedre enn GLM, LDA, QDA, NN og DNNA. Selv om GAM ikke er signifikant bedre enn de maskinlæringsteknikkene som ga best prediksjonsevne, så er den allikevel den beste estimeringsmetoden totalt sett ved estimering av modell basert på variabelsett 1.

4.2 Resultater fra variabelsett 2 – basert på Altmans Z-score-modell

I figur 15 blir det gjengitt hvor bra de forskjellige modellene basert på variabelsett 2 predikerer konkurs, ettersom hvilken metode de estimeres med. Fargekodene representerer type 1- og type 2-feil, slik som i figur 14.



Figur 15: Feilklassifiseringsrate for modeller basert på variabelsett 2

Figuren viser at de ulike modellene predikerer ikke konkurs riktig i omtrent 85% til 90% av tilfellene, mens de predikerer konkurs riktig i 25% til 70% av tilfellene. Det kommer tydelig frem av figuren at også ved å basere modellene på variabelsett 2 gjør samtlige modeller flere type 2-feil enn type 1-feil .

I tabell 18 presenteres oppnådd ROC-verdi og treffprosent for de ti modellene, som alle er basert på variabelsett 2 og estimert ved denne oppgavens ti valgte metoder. Blant de tradisjonelle statistiske metodene er det GAM som estimerer den beste modellen, med en ROC-verdi på 88% og en treffprosent på 82%. QDA estimerer den modellen som predikerer dårligst blant de tradisjonelle statistiske metodene, med en ROC-verdi på 69% og en treffprosent på 69%. Av de nyere maskinlæringsteknikkene er det CT, KNN og DNNB som estimerer modeller som predikerer best, med en ROC-verdi på 86%-87% og treffprosent på 81%-82%.

Tabell 18 viser, slik som for tabell 17, at det også for modeller estimert med variabelsett 2 er forskjeller mellom deres prediksjonsevne. Tabell 30 i Appendiks 6 viser p -verdiene for gjennomførte McNemars tester. Her kommer det frem at modeller estimert ved GAM er signifikant bedre til å predikere konkurs enn de andre tradisjonelle statistiske metodene: GLM, LDA og QDA. Dette er i samsvar med modeller basert på variabelsett 1. GAM er også en signifikant bedre metode enn maskinlæringsteknikkene NN og DNNA. Den maskinlæringsteknikken som ga modell med best prediksjonsevne, DNNB, er signifikant bedre enn GLM, LDA, QDA, NN og DNNA. GAM er den metoden som totalt sett gir best prediksjonsevne for modeller basert på variabelsett 2, selv om den ikke kan sies å være signifikant bedre enn de beste maskinlæringsteknikkene.

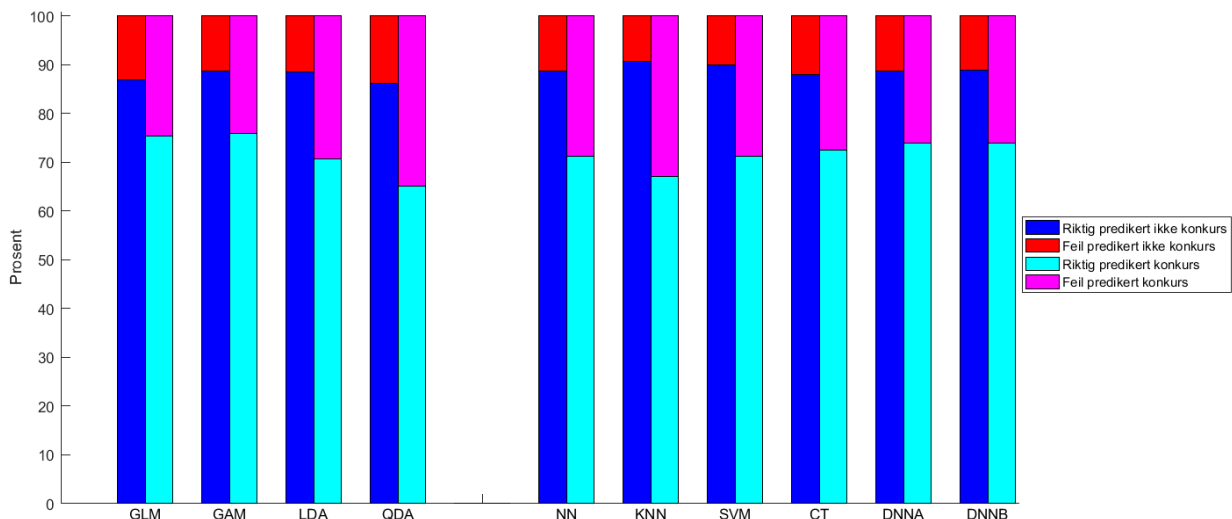
Tabell 18: ROC-verdier og treffprosent for modeller basert på variabelsett 2

Tradisjonelle statistiske metoder	ROC-verdi	Treffprosent
Generell Lineær Modell (GLM)	82,3%	77,5%
Generell Additiv Modell (GAM)	87,5%	81,7%
Lineær diskriminantanalyse (LDA)	73,8%	70,7%
Kvadratisk diskriminantanalyse (QDA)	68,8%	69,4%
Maskinlæringsteknikker	ROC-verdi	Treffprosent
Tolags forovermatet nevralt nettverk (NN)	74,4%	71,1%
k -nærmeste naboer (KNN)	86,7%	81,1%
Støttevektormaskin (SVM)	85,5%	81,2%
Klassifiseringstrær (CT)	86,4%	80,4%
Dypt nevralt nettverk skapt av to autoencodere (DNNA)	78,4%	73,3%
Dypt nevralt nettverk med back-propagation (DNNB)	86,8%	81,5%

4.3 Resultater fra variabelsett 3 – egenutviklet

I dette delkapitlet presenteres prediksjonsevnen til modeller basert på det egenutviklede variabelsettet. I figur 16 blir det gjengitt hvor bra de forskjellige modellene basert på variabelsett 3 predikerer konkurs, ettersom hvilken metode de estimeres med. Fargekodene representerer type 1- og type 2-feil, slik som i figur 14 og figur 15.

I tabell 19 presenteres oppnådd ROC-verdi og treffprosent for de ti modellene, som alle er basert på variabelsett 3 og estimert ved denne oppgavens ti valgte metoder. Blant de tradisjonelle statistiske metodene er det GAM som estimerer modellen som oppnår beste prediksjonsevne, med en ROC-verdi på 91% og en treffprosent på 84%. Av de nyere maskinlæringsteknikkene er det SVM, DNNA og DNNB som estimerer modeller med best prediksjonsevne, med en ROC-verdi på 90% og treffprosent på 84%. Her er det verdt å merke seg at modeller estimert ved GAM og minst tre av maskinlæringsteknikkene oppnår en ROC-verdi og treffprosent som er bedre enn det Norges Bank har offentliggjort for samtlige av sine tre versjoner av SEBRA-modellen: opprinnelige versjon, basisversjon og utvidet versjon (Bernhardsen og Larsen, 2007, Eklund et al., 2001). Dette er i tillegg en bedre ROC-verdi og treffprosent enn oppnådd med denne oppgavens variabelsett 1, som er basert på Norges Banks SEBRA-modell.



Figur 16: Feilklassifiseringsrate for modeller basert på variabelsett 3

Tabell 19: ROC-verdier og treffprosent for modeller basert på variabelsett 3

Tradisjonelle statistiske metoder	ROC-verdi	Treffprosent
Generell Lineær Modell (GLM)	89,6%	83,0%
Generell Additiv Modell (GAM)	91,1%	84,4%
Lineær diskriminantanalyse (LDA)	89,0%	82,5%
Kvadratisk diskriminantanalyse (QDA)	85,0%	79,1%
Maskinlæringsteknikker	ROC-verdi	Treffprosent
Tolags forovermatet nevralt nettverk (NN)	89,2%	82,8%
<i>k</i> -nærmeste naboer (KNN)	89,5%	82,8%
Støttevektormaskin (SVM)	90,0%	83,7%
Klassifiseringstrær (CT)	89,3%	82,8%
Dypt nevralt nettverk skapt av to autoencodere (DNNA)	90,1%	83,7%
Dypt nevralt nettverk med back-propagation (DNNB)	90,1%	83,8%

Tabell 19 viser at det er forskjeller i prediksjonsevne for modeller basert på variabelsett 3, avhengig av hvilken metode de estimeres med. *p*-verdiene ved McNemars test følger i tabell 31 i Appendiks 6. Det er her, i likhet med modellene basert på variabelsett 1 og 2, fremdeles GAM og DNNB som gir modeller med best prediksjonsevne blant henholdsvis tradisjonelle statistiske metoder og maskinlæringsteknikker. Derimot er nå både GAM og DNNB kun signifikant bedre enn QDA. Totalt sett er det likevel tilsvarende resultater hva gjelder signifikant forskjell som for modeller basert på de øvrige variabelsettene i denne oppgaven. For estimering av modeller basert på variabelsett 3 er også GAM den metoden som gir best prediksjonsevne, men den er altså ikke signifikant bedre enn noen av maskinlæringsteknikkene.

4.4 Drøfting

Fra de foregående delkapitlene er det presentert et resultat som tilsier at modeller estimert ved nyere maskinlæringsteknikker predikerer konkurs marginalt dårligere enn GAM, som viser seg å være den best egnede metoden for estimering av konkursprediksjonsmodeller.

Variabelsett 1, basert på variablene i Norges Banks SEBRA-modell, gir modeller som predikerer stort sett bedre enn modeller ved variabelsett 2, basert på variablene i Altmans Z-score-modell, både med tanke på oppnådd ROC-verdi og treffprosent, uavhengig av statistisk metode. Dette er ikke overraskende siden Norges Banks SEBRA-modell er utviklet basert på norske selskapsdata, mens Altmans Z-score-modell er basert på data fra amerikanske selskaper for om lag 50 år siden. Sistnevnte modell brukte også nøkkeltall som var ansett som

sentrale for det amerikanske markedet, noe som ikke nødvendigvis ikke er like relevante for Norge i dag. Det er likevel interessant å se at på tross av dette predikerer modeller basert på variabelsett 2 nesten like godt som modeller basert på variabelsett 1, noe som forsvare den sentrale posisjonen Altmans Z-score-modell har fått innenfor dette forskningsfeltet. Riktignok er ikke Norges Banks SEBRA-modell helt ny den heller, da den har vært i bruk siden 2001 av Norges Bank til å anslå konkurssannsynligheter (Bernhardsen og Larsen, 2007).

Det kan tenkes at i tiden etter etableringen av Norges Banks SEBRA-modell har nye nøkkeltall blitt viktigere og således kan bidra til mer presis prediksjon. Det ble derfor forsøkt predikert konkurs med modeller basert på et siste, og utvidet, variabelsett. Mange ulike variabler kan imidlertid by på problemer med multikollinearitet, men når modellen er konstruert kun for prediksjon er dette allikevel ikke et problem (Berg, 2007). Det viser seg at modeller basert på dette siste variabelsettet predikerer konkurs bedre enn modeller basert på variabelsett 1 og 2, både ved den beste maskinlæringsteknikken (DNNB) og den beste tradisjonelle statistiske metoden (GAM). At variabelsett 3 presterer best kan virke logisk, da dette benytter flest nøkkeltall og derfor dekker flest økonomiske aspekter. I tillegg kommer det frem at forskjellen i prediksjonsevne mellom modeller estimert ved tradisjonelle statistiske metoder og maskinlæringsteknikker er i gjennomsnitt mindre for variabelsett 3, enn variabelsett 1 og 2. Som det fremgår av p -verdiene i tabell 32 og tabell 33 i Appendiks 6 er det imidlertid ingen forskjell på prediksjonsevne mellom modeller avhengig av hvilket variabelsett de baserer seg på, men allikevel gir funnene i denne oppgaven en indikasjon om forskjell.

GAM er den metoden blant de tradisjonelle statistiske metodene som estimerer de beste modellene, mens DNNB er den mest egnede maskinlæringsteknikken. Dette gjelder uavhengig av hvilket variabelsett modellene baserer seg på. Samtidig gir modeller estimert ved GAM marginalt bedre prediksjonsevne enn modeller estimert ved DNNB. Allikevel er det verdt å merke seg at forskjellen ikke er signifikant, og at nyere maskinlæringsteknikker generelt predikerer godt, til tross for at de i mindre grad er benyttet ved konkursprediksjon enn de mer tradisjonelle statistiske metoder. Det kommer også frem at forskjellen mellom modeller basert på GAM og de best egnede maskinlæringsteknikkene blir gradvis mindre jo flere variabler som benyttes i modellene. Det kan således tenkes at de best egnede maskinlæringsteknikkene til slutt overgår GAM, dersom enda flere variabler benyttes.

5 Konklusjon og implikasjoner

I tråd med Bernhardsen (2001) og Eklund et al. (2001) finner også denne oppgaven at modeller estimert ved den statistiske metoden GAM oppnår best prediksjonsevne. Samtidig er det funnet at et egenutviklet variabelsett, bestående av flere variabler enn i Norges Bank SEBRA-modell, oppnår en bedre prediksjonsevne enn det Norges Bank viser til i sine rapporter. Det er også undersøkt hvorvidt det er signifikant forskjell mellom modellene ved å benytte McNemars test, noe det viser seg at kun et fåtall av tidligere studier har undersøkt.

Problemstilling 1

I hvilken grad kan det implementeres et variabelsett som predikerer konkurs bedre enn variabelsettet benyttet i Norges Banks SEBRA-modell?

Ved å estimere en modell ved GAM og basert på et egenutviklet variabelsett er det i denne oppgaven oppnådd en ROC-verdi på 91,1% og en treffprosent på 84,4%. Norges Bank opplyser at for deres SEBRA-modeller oppnås omtrent 1,5% lavere verdi for både ROC-verdi og treffprosent. Også modeller estimert ved flere maskinlæringsteknikker og basert på det samme egenutviklede variabelsettet gir en bedre prediksjonsevne enn det Norges Bank opplyser om for sin SEBRA-modell. Dette indikerer at det er mulig å implementere et variabelsett som predikerer konkurs bedre enn variabelsettet benyttet i Norges Banks SEBRA-modell. Tatt i betraktning av datasettets størrelse er det også ikke urimelig å hevde at dette er et representativt funn. Ved McNemars test kan det ikke hevdes signifikant forskjell mellom Norges Banks SEBRA-modell og en modell basert på et egenutviklet variabelsett, men funnene gir allikevel en indikasjon om forskjell.

Problemstilling 2

Ved estimering av Norges Banks SEBRA-modell er den tradisjonelle statistiske metoden GAM (generell additiv modell) benyttet. I hvilken grad er det mulig å forbedre modellen ved å heller benytte maskinlæringsteknikker?

I denne oppgaven kommer det frem at GAM estimerer modellene med best prediksjonsevne, uavhengig av hvilket variabelsett modellene baserer seg på. Maskinlæringsteknikker estimerer modeller som predikerer konkurs marginalt dårligere, men forskjellen er ikke signifikant. Det er verdt å merke seg at de aller fleste nyere maskinlæringsteknikker benyttet i denne oppgaven estimerer modeller som predikerer konkurs svært godt. Dette til tross for at de i mindre grad er omtalt i tidligere forskning, spesielt innenfor forskning på norske selskaper. Det er derfor på sin plass å poengtere at maskinlæringsteknikker med fordel kan anvendes ved

fremtidige estimeringer av modeller for konkursprediksjon. Dette er i tråd med Varian (2014) som hevder at ved store datasett kan nyere maskinlæringsteknikker egne seg bedre enn tradisjonelle statistiske metoder.

Implikasjoner

Det kan antas at kostnaden ved å estimere en bedre modell for konkursprediksjon enkelt kan forsvares dersom denne bistår med å begrense fremtidige utlånstap for banker. Forskjell i utlånstap på 1,5% utgjør et stort kronebeløp dersom man forutsetter at norske banker til enhver tid har utestående flere hundre milliarder kroner i lån. I tillegg til banker kan også andre institusjoner ha nytte av forbedrede modeller for konkursprediksjon, for eksempel selskaper som analyserer sannsynligheten for konkurs blant sine kunder, leverandører eller konkurrenter. Offentlige organer som stresstester markedet vil også ha stor nytte av modeller som predikerer konkurs mest mulig riktig, for å kunne forutsi og unngå fremtidige finansielle kriser. Dette indikerer at de samfunnskostnadene som potensielt unngås ved bedre modeller overgår kostnaden ved å finne måter å forbedre disse på.

Funnene i denne oppgaven indikerer et forbedringspotensial for etablerte konkursprediksjonsmodeller, og er således et godt bidrag til videre forskning innenfor konkursprediksjon basert på regnskapsdata.

5.1 Forslag til videre forskning

Det hadde vært interessant å gjøre tilsvarende analyser som i denne oppgaven også på internasjonale regnskapsdata. I tillegg hadde det vært interessant å estimere modeller med flere uavhengige variabler enn de som er gjennomgått i denne oppgaven, samt teste for ulike kombinasjoner av disse. Kanskje kan det tenkes at nyere maskinlæringsteknikker presterer bedre enn tradisjonelle statistiske metoder dersom enda flere variabler inkluderes, slik funnene i oppgaven indikerer, som nevnt i kapittel 4.4. Det hadde også vært interessant å estimere modeller som predikerer konkurs i opp mot fem år frem i tid, slik som i Altman (1968). Videre hadde det vært interessant å analysere flere metoder, som for eksempel hybride maskinlæringsteknikker og flere varianter av dyp maskinlæring, for å se hvordan disse presterer ved estimering av modeller for konkursprediksjon. Det kan også tenkes at flere av denne oppgavens 30 modeller predikerer de samme type selskaper feil, så en analyse av disse eventuelle selskapene kunne bidratt til ytterligere forbedring av prediksjonsevne.

Referanser

- (1986) Lov om gjeldsforhandling og konkurs (konkursloven). I: Beredskapsdepartementet, J.-O. (red.).
- Agarwal, V. og Taffler, R. (2008) Comparing the performance of market-based and accounting-based bankruptcy prediction models, *Journal of Banking and Finance*, 32 (8), s. 1541-1551.
- Akay, O. og Winters, D. (2011) Temporary open market operation on MBS repos: any foreshadowing of the financial crisis of 2008?, *Journal of Economics and Finance*, 35 (3), s. 260-273.
- Aldrich, J. H. og Nelson, F. D. (1984) *Linear probability, logit, and probit models*. Beverly Hills: Sage.
- Altman, E. I. (1968) Financial ratios, discriminant analysis and prediction of corporate bankruptcy *Journal of Finance*, 23 (4), s. 589-609.
- Altman, E. I. (2000) Predicting financial distress of companies: revisiting the Z-score and ZETA models, *Stern School of Business, New York University*, s. 9-12.
- Altman, E. I., Haldeman, R. G. og Narayanan, P. (1977) ZETATM analysis A new model to identify bankruptcy risk of corporations, *Journal of Banking & Finance*, 1 (1), s. 29-54.
- Altman, E. I., Iwanicz-Drozdowska, M., Laitinen, E. K. og Suvas, A. (2014) Distressed Firm and Bankruptcy Prediction in an International Context: A Review and Empirical Analysis of Altman's Z-Score Model, *Available at SSRN 2536340*.
- Baker, H. K. og Martin, G. S. (2011) *Capital Structure and Corporate Financing Decisions: Theory, Evidence, and Practice*. Hoboken, NJ, USA: Hoboken, NJ, USA: John Wiley & Sons, Inc.
- Beaver, W. H. (1966) Financial Ratios As Predictors of Failure, *Journal of Accounting Research*, 4, s. 71-111.
- Bellovary, J., Giacominio, D. og Akers, M. (2007) A Review of Bankruptcy Prediction Studies: 1930-Present, *Journal of Financial Education*, 33.
- Bengio, Y. (2009) Learning deep architectures for AI, *Foundations and trends® in Machine Learning*, 2 (1), s. 1-127.
- Berg, D. (2007) Bankruptcy prediction by generalized additive models, *Applied Stochastic Models in Business and Industry*, 23 (2), s. 129-143.
- Bernhardsen, E. (2001) A Model of Bankruptcy Prediction, *Working papers from Norges Bank*, 2001/10.
- Bernhardsen, E. og Larsen, K. (2007) Modelling av kredittrisiko i foretakssektoren ; videreutvikling av SEBRA-modellen, *Penger og kreditt*.
- Bernhardsen, E. og Syversten, B. D. (2009) Stress Testing the Enterprise Sector's Bank Debt: A Micro Approach, *International Journal of Central Banking*, 5 (3), s. 111-138.
- Bielecki, T., Brigo, D. og Patras, F. (2011) *Credit Risk Frontiers : Subprime Crisis, Pricing and Hedging, CVA, MBS, Ratings, and Liquidity*. Princeton: Wiley.
- Boyacioglu, M. A., Kara, Y. og Baykan, Ö. K. (2009) Predicting bank financial failures using neural networks, support vector machines and multivariate statistical methods: A comparative analysis in the sample of savings deposit insurance fund (SDIF) transferred banks in Turkey, *Expert Systems with Applications*, 36 (2), s. 3355-3366.
- Breiman, L., Friedman, J. H., Olshen, R. A. og Stone, C. J. (1984) *Classification and Regression Trees*. CRC Press LLC.
- Chang, A. C. og Li, P. (2015) Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say "Usually Not", *Finance and Economics Discussion Series 2015-083*.

- Chen, N., Ribeiro, B., Vieira, A. S., Duarte, J. og Neves, J. C. (2011) A genetic algorithm-based approach to cost-sensitive bankruptcy prediction, *Expert Systems with Applications*, 38 (10), s. 12939-12945.
- Cielen, A., Peeters, L. og Vanhoof, K. (2004) Bankruptcy prediction using a data envelopment analysis, *European Journal of Operational Research*, 154 (2), s. 526-532.
- Cover, T. og Hart, P. (1967) Nearest neighbor pattern classification, *IEEE Transactions on Information Theory*, 13 (1), s. 21-27.
- Cristianini, N. og Shawe-Taylor, J. (2000) *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge: Cambridge : Cambridge University Press.
- Dietterich, T. G. (1998) Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms, *Neural Computation*, 10 (7), s. 1895-1923.
- Eklund, T., Larsen, K. og Bernhardsen, E. (2001) Modell for analyse av kredittrisiko i foretakssektoren, *Penger og kreditt*, 2.
- Fawcett, T. (2006) An introduction to ROC analysis, *Pattern recognition letters*, 27 (8), s. 861-874.
- Finanstilsynet. (2013a) Finansielt utsyn 2013.
- Finanstilsynet. (2013b) Modul for Kredittrisiko.
- Finanstilsynet. (2014) Finansielt utsyn 2014.
- Finanstilsynet. (2015a) *Bankenes kapitalkrav* [Internett]. <http://www.finanstilsynet.no/>: Finanstilsynet. Tilgjengelig fra: http://www.finanstilsynet.no/no/Artikkelarkiv/Aktuelt/2015/4_kvartal/Bankenes-kapitalkrav/ (Hentet: 2016-03-31).
- Finanstilsynet. (2015b) Finansielt utsyn 2015.
- Fisher, R. A. (1936) The use of multiple measurements in taxonomic problems, *Annals of eugenics*, 7 (2), s. 179-188.
- FitzPatrick, P. J. (1932) *A comparison of the ratios of successful industrial enterprises with those of failed companies*. Ph.D.
- Gaussier, E. og Yvon, F. (2013) *Textual Information Access : Statistical Models*. Hoboken: Wiley.
- Gillick, L. og Cox, S. J. (1989) Some statistical issues in the comparison of speech recognition algorithms, *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on*, s. 532-535 vol.531.
- Graupe, D. (2013) *Principles of Artificial Neural Networks*. Singapore: World Scientific Publishing Company.
- Guo, Y., Hastie, T. og Tibshirani, R. (2007) Regularized linear discriminant analysis and its application in microarrays, *Biostatistics*, 8 (1), s. 86-100.
- Güres, S., Mendyk, A., Jachowicz, R., Dorożyński, P. og Kleinebudde, P. (2012) Application of artificial neural networks (ANNs) and genetic programming (GP) for prediction of drug release from solid lipid matrices, *International Journal of Pharmaceutics*, 436 (1-2), s. 877-879.
- Hastie, T. og Tibshirani, R. (1995) Generalized additive models for medical research, *Statistical methods in medical research*, 4 (3), s. 187-196.
- Hinton, G. E. og Salakhutdinov, R. R. (2006) Reducing the dimensionality of data with neural networks, *Science (New York, N.Y.)*, 313 (5786), s. 504.
- Hung, C. og Chen, J.-H. (2009) A selective ensemble based on expected probabilities for bankruptcy prediction, *Expert Systems with Applications*, 36 (3), s. 5297-5303.
- Håland, S. (2016) Banker i en drømmeverden, *Dagens næringsliv*, 2016-03-21.

- Jin, H. og Ling, C. X. (2005) Using AUC and accuracy in evaluating learning algorithms, *Knowledge and Data Engineering, IEEE Transactions on*, 17 (3), s. 299-310.
- Joy, O. M. og Tollefson, J. O. (1975) On the Financial Applications of Discriminant Analysis, *J. Financ. Quant. Anal.*, 10 (5), s. 723-739.
- Kirkos, E. (2015) Assessing methodologies for intelligent bankruptcy prediction, *Artificial Intelligence Review*, 43 (1), s. 83-123.
- Kleppe, H. (2016) Kan ha skjult 10.700 milliarder i råtne lån - mer enn det dobbelte av subprime-krisen, *Hegnar.no*, 2016-04-15. Tilgjengelig fra <http://hegnar.no/Nyheter/Boers-finans/2016/04/Kan-ha-skjult-10.700-milliarder-i-raatne-laan-mer-enn-det-dobbelte-av-subprime-krisen>
- Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Ijcai*, 1995. s. 1137-1145.
- Konkursrådet. (2012) *Innføring i konkurs* [Internett]. Tilgjengelig fra: <http://www.konkursradet.no/innfoering-i-konkurs.305266.no.html#2.1> (Hentet: 2016-01-26).
- Laitinen, E. K. og Laitinen, T. (2000) Bankruptcy prediction: Application of the Taylor's expansion in logistic regression, *International Review of Financial Analysis*, 9 (4), s. 327-349.
- Larsen, K. og Sæther, P. (1999) Kredittrisiko knyttet til foretakssektoren, *Penger og kreditt*, 4/1999, s. 546-554.
- Lawrence, J. R., Pongsatit, S. og Lawrence, H. (2015) The Use Of Ohlson's O-Score For Bankruptcy Prediction In Thailand, *Journal of Applied Business Research (JABR)*, 31 (6), s. 1067-1077.
- Ledolter, J. (2013) *Data Mining and Business Analytics with R*. Hoboken: Wiley.
- Liang, Y., Xu, Q.-S., Li, H.-D. og Cao, D.-S. (2011) *Support Vector Machines and Their Application in Chemistry and Biotechnology*. Hoboken: Taylor and Francis.
- Martens, D., Gestel, T. V., Backer, M. D., Haesen, R., Vanthienen, J. og Baesens, B. (2009) Credit rating prediction using Ant Colony Optimization, *Journal of the Operational Research Society*, 61 (4), s. 561.
- McNemar, Q. (1947) Note on the sampling error of the difference between correlated proportions or percentages, *Psychometrika*, 12 (2), s. 153-157.
- Næss, A. B. (2015) *Nearest Neighbor Frame Classification for Articulatory Speech Recognition*. ph.d., NTNU.
- Ohlson, J. A. (1980) Financial Ratios and the Probabilistic Prediction of Bankruptcy, *Journal of Accounting Research*, 18 (1), s. 109-131.
- Oluwo, M. (2007) *Strategic use of financial ratio to prevent bankruptcy: A study of opportunity for business enterprises*. CAPELLA UNIVERSITY.
- Oslo Børs. (2016) *Konkurs* [Internett]. Tilgjengelig fra: http://www.oslobors.no/content/search?SearchText=konkurs?param=1&SearchText=konkurs&search_radio=other (Hentet: 2016-02-25).
- Park, C.-S. og Han, I. (2002) A case-based reasoning with the feature weights derived by analytic hierarchy process for bankruptcy prediction, *Expert Systems with Applications*, 23 (3), s. 255-264.
- Patti, C. (2001) Description, explanation, prediction – the evolution of bankruptcy studies?, *Managerial Finance*, 27 (4), s. 29-44.
- Ravi Kumar, P. og Ravi, V. (2007) Bankruptcy prediction in banks and firms via statistical and intelligent techniques – A review, *European Journal of Operational Research*, 180 (1), s. 1-28.
- Rosenblatt, F. (1958) The perceptron: A probabilistic model for information storage and organization in the brain, *Psychological Review*, 65 (6), s. 386-408.

- Schmidhuber, J. (2015) Deep learning in neural networks: an overview, *Neural networks : the official journal of the International Neural Network Society*, 61, s. 85.
- SSB. (2015) *Opna konkurser* [Internett]. SSB.no. Tilgjengelig fra:
<https://www.ssb.no/statistikkbanken/selectvarval/Define.asp?subjectcode=&ProductId=&MainTable=Konkurs5&nvl=&PLanguage=0&nyTmpVar=true&CMSSubjectArea=&KortNavnWeb=konkurs&StatVariant=&checked=true>.
- Syversten, B. D., Johansen, R. M., Lind, Ø. A., Solheim, H. og Stefano, N. (2015) Bankmodellen og stresstesten i rapport om Finansiell stabilitet 2015. Staff Memo: Norges Bank.
- Timiraos, N. (2013) U.S. News: Housing Prices Cool As Loan Rates Rise. New York, N.Y. University of, I. og Bureau of Business, R. (1930) *A test analysis of unsuccessful industrial companies*. Urbana: University of Illinois.
- Vapnik, V. N. (1998) *Statistical learning theory*. New York: Wiley.
- Varian, H. R. (2014) Big data: new tricks for econometrics, *Journal of Economic Perspectives*, 28 (2), s. 3-28.
- Vishnubhotla, S., Fernandez, R. og Ramabhadran, B. (2010) An autoencoder neural-network based low-dimensionality approach to excitation modeling for HMM-based text-to-speech.
- Wetcher-Hendricks, D. (2014) *Analyzing Quantitative Data : An Introduction for Social Researchers*. Hoboken: Wiley.
- Wilson, R. L. og Sharda, R. (1994) Bankruptcy prediction using neural networks, *Decision Support Systems*, 11 (5), s. 545-557.
- Zhao, W. D. og Zhang, Y. D. (2014) Application development of student score analysis system for secondary school based on standard score, *Advanced Materials Research*, 989-994, s. 5062-5065.

Appendiks 1: Variabelsett 1 – basert på Norges Banks SEBRA-modell

Nedenfor er en beskrivelse av hvordan de 7 uavhengige variablene oppgavens i variabelsett 1 blir beregnet.

Resultat (res)

I Larsen og Sæther (1999) står det beskrevet at SEBRA-modellen benytter forholdstallet kontantinntjening, som ligner mye på det som omtales som «Ordinært resultat før av- og nedskrivninger i prosent av total gjeld» i Bernhardsen og Larsen (2007). Videre står det følgende i Larsen og Sæther (1999): *«I risikomodellen defineres kontantinntjeningen som resultat før ekstraordinære poster, eksklusive av-/nedskrivninger og inklusive skattekostnad. I SEBRA-modellen defineres tilfredsstillende kontantinntjening som 20 prosent eller mer av foretakets langsiktige gjeld.»* Med bakgrunn i dette blir den uavhengige variabelen Resultat (res) beregnet som beskrevet nedenfor.

Først ble det beregnet ordinært resultat før av- og nedskrivninger som følger:

$$\begin{aligned} & \text{Ordinært resultat før skattekostnad} \\ & - \text{Skattekostnad på ordinært resultat} \\ & + \text{Avskrivning varige driftsmidler/im. eiend.} \\ & + \text{Nedskrivning av driftsmidler/im. eiend.} \\ & + \text{Nedskrivn. av andre fin.omløpsmidler} \\ & + \text{Nedskrivning finansielle anleggsmidler} \\ & = \underline{\underline{\text{Ordinært resultat før av- og nedskrivninger}}} \end{aligned}$$

Deretter ble variabelen Resultat (res) beregnet slik:

$$\text{Resultat (res)} = \frac{\text{Ordinært resultat før av- og nedskrivninger}}{\text{Sum langsiktig gjeld}}$$

I rapporten Finansielt Utsyn 2015 velger Finanstilsynet (2015b) å ikke trekke ut av- og nedskrivninger, slik at variabelen skal være en indikator på selskapets langsiktige gjeldsbetalingssevne. Dette synes å være aktuelt ved beregning av forventet utlånstap. I denne oppgaven blir det predikert konkurs av enkeltsselskaper, så av- og nedskrivninger blir trukket

ut av variabelen. I rapporten Finansielt Utsyn 2013 (Finanstilsynet, 2013a) blir det også valgt å trekke ut av- og nedskrivninger da disse kun er regnskapsmessige disposisjoner.

Variabelen gir den inverse av hvor mange år det vil ta å tilbakebetale den langsiktige gjelden med det resultatet som ble oppnådd i det aktuelle året (Bernhardsen og Syversten, 2009).

Egenkapitalprosent (ek)

Som kjent er totalkapital lik summen av gjeld og egenkapital. Dermed ble variabelen Egenkapitalprosent (ek) beregnet slik:

$$\text{Egenkapitalprosent (ek)} = \frac{\text{Sum egenkapital} - \text{goodwill og andre ikke-materielle eiendeler}}{\text{Sum egenkapital} - \text{goodwill og andre ikke-materielle eiendeler} + \text{Sum gjeld}}$$

Goodwill og andre ikke-materielle eiendeler trekkes fra siden egenkapitalprosent skal si noe om kapitalbuffer til dårligere tider (Finanstilsynet, 2015b).

Likviditetsprosent (likv)

I Larsen og Sæther (1999) står det at i «(...) *SEBRA-modellen er likviditetsindikatoren definert som betalingsmidler (kontanter og andre likvide betalingsmidler) fratrukket kortsiktig gjeld i prosent av driftsinntekter.*» På bakgrunn av dette er variabelen Likviditetsprosent (likv) definert slik:

$$\text{Likviditetsprosent (likv)} = \frac{\text{Kasse/bank/post} - \text{sum kortsiktig gjeld}}{\text{Sum driftsinntekter}}$$

I rapporten Finansielt Utsyn 2015 velger Finanstilsynet (2015b) å benytte omløpsmidler i stedet for kasse/bank/post. I denne oppgaven blir det valgt å heller benytte kasse/bank/post i variabelen, da ikke alle omløpsmidler antas å kunne dekke kreditorers krav ved eventuell begjæring av konkurs.

Eiendeler (eie)

Denne variabelen er lik regnskapstallet «Sum eiendeler», altså totalkapital, som oppgitt fra Proff Forvalt.

Leverandørgjeldprosent (lev)

Denne variabelen blir beregnet slik:

$$\text{Leverandørgjeldprosent (lev)} = \frac{\text{Leverandørgjeld}}{\text{Sum egenkapital} + \text{Sum gjeld}}$$

Avgiftsprosent (avg)

Denne variabelen blir beregnet slik:

$$\text{Avgiftsprosent (avg)} = \frac{\text{Skyldige offentlige utgifter}}{\text{Sum egenkapital} + \text{Sum gjeld}}$$

Alder (alder)

For alder blir det i denne oppgaven benyttet en dummyvariabel for hvert år mellom ett og åtte som selskapet har eksistert, slik som nevnt i Berhardsen og Syversten (2009).

Selskapets alder blir definert som antall år selskapet har eksistert ved utgangen av regnskapsåret, altså differansen mellom stiftelsesdato og dato 31. desember det aktuelle året, som analyseres. Alderen blir altså rundet nedover til nærmeste hele år.

Variablene $a_1, a_2, a_3, a_4, a_5, a_6, a_7$ og a_8 er dummyvariabler, med verdiene 1 eller 2.

Verdiene settes slik:

- Om selskapet er 1 år eller eldre blir variabelen a_1 satt til 1, om ikke blir den satt til 2.
- Om selskapet er 2 år eller eldre blir variabelen a_2 satt til 1, om ikke blir den satt til 2.
- Om selskapet er 3 år eller eldre blir variabelen a_3 satt til 1, om ikke blir den satt til 2.
- Om selskapet er 4 år eller eldre blir variabelen a_4 satt til 1, om ikke blir den satt til 2.
- Om selskapet er 5 år eller eldre blir variabelen a_5 satt til 1, om ikke blir den satt til 2.
- Om selskapet er 6 år eller eldre blir variabelen a_6 satt til 1, om ikke blir den satt til 2.
- Om selskapet er 7 år eller eldre blir variabelen a_7 satt til 1, om ikke blir den satt til 2.
- Om selskapet er 8 år eller eldre blir variabelen a_8 satt til 1, om ikke blir den satt til 2.

Utbytte

I tidlige versjoner av SEBRA-modellen ble det inkludert en uavhengig dummyvariabel for hvorvidt det ble utbetalt utbytte siste regnskapsår (Eklund et al., 2001). Argumentasjonen for å ha en slik variabel var at den reflekterer eiernes forventning til fremtidig inntjening i selskapet. Men i siste versjoner er denne variabelen utelatt fordi utbytte kan skyldes like mye skattetilpasning som forventning til fremtidig inntjening (Bernhardsen og Larsen, 2007). I denne oppgaven er det derfor valgt å ikke inkludere en uavhengig variabel knyttet til utbytte i siste regnskapsår.

Appendiks 2: Variabelsett 2 – basert på Altmans Z-score-modell

Basert på hvordan Altman (1968) definerte sine 5 variabler i sin Z-score-modell er variablene i variabelsett 2 definert slik:

$$\text{Likviditet } (x_1) = \frac{\text{Sum omløpsmidler} - \text{sum kortsiktig gjeld}}{\text{Totalkapital}}$$

$$\text{Lønnsomhet } (x_2) = \frac{\text{Overføringer til/fra annen egenkapital}}{\text{Totalkapital}}$$

$$\text{Gjeld } (x_3) = \frac{\text{Ordinært resultat før skattekostnad} + \text{Sum annen rentekostnad}}{\text{Totalkapital}}$$

$$\text{Soliditet } (x_4) = \frac{\text{Sum egenkapital}}{\text{Totalkapital}}$$

$$\text{Aktivitetsnøkkeltall } (x_5) = \frac{\text{Sum salgsinntekter}}{\text{Totalkapital}}$$

Legg merke til at Altman (1968) benytter markedsverdien til egenkapitalen ved beregning av Soliditet (x_4). Siden denne ikke finnes i datagrunnlaget benyttes i stedet bokført egenkapital, slik også Altman (2000) gjorde i sin reviderte Z-Score-modell.

Appendiks 3: Variabelsett 3

Variabelsett 3 består av samtlige variabler i variabelsett 1, som omtalt i Appendiks 1, i tillegg til ni variabler som definert nedenfor.

Opptjent egenkapital (*opptjek*)

Dersom opptjent egenkapital er negativ i årsregnskapet blir denne variabelen satt til 1, om ikke blir den satt til 2.

Opptjent egenkapital er den delen av selskapets overskudd som ikke tas ut i utbytte. Dersom denne er negativ betyr det at akkumulert overskudd, fratrukket eventuelt utbytte, siden oppstart av selskapet er negativt. Selskapet har altså da totalt sett drevet ulønnsom virksomhet siden oppstart, og det kan da tyde på at konkurs er nærliggende. På grunn av dette er det naturlig å anta at variabelen *opptjek* gir en indikasjon på sannsynligheten for konkurs, og den er derfor inkludert i variabelsett 3.

Rentebærende gjeld (*rentgje*)

$$\text{Rentebærende gjeld (rentgje)} = \frac{\text{Gjeld til kredittinstitusjoner}}{\text{Sum gjeld}}$$

Det kan antas at gjeld til kredittinstitusjoner har lengre løpetid enn annen gjeld, og at denne er den eneste gjelden som gir rentekostnader. Variabelen *rentgje* er derfor valgt inkludert i variabelsett 3 da denne altså antas å fange opp både forskjellig tidshorisont på gjelden og andel gjeld som gir kostnader.

Omsetning per eiendeler (*omseie*)

$$\text{Omsetning per eiendeler (omseie)} = \frac{\text{Salgsinntekter}}{\text{totalkapital}}$$

Dette gir et anslag på hvor godt kapitalen i selskap utnyttes og kan gi et inntrykk av hvor effektivt et selskap drives, som igjen kan antas å ha innvirkning på sannsynligheten for om et selskap går konkurs.

Avkastning på egenkapital (avkek)

$$\text{Avkastning på egenkapital (avkek)} = \frac{\text{Årsresultat}}{\text{Sum egenkapital}}$$

Denne variabelen sier noe om hvor godt egenkapitalen i selskapet utnyttes. Dersom denne variabelen er høy kan det antas at selskapet er et godt investeringsobjekt, noe som bidrar til enklere tilgang til investorkapital. Videre kan det tenkes at dette minker sannsynligheten for konkurs ved eventuelle likviditetsproblemer, noe som dermed trolig gjør variabelen avkek relevant ved konkursprediksjon.

Omsetning på lønninger (omslo)

$$\text{Omsetning på lønninger (omslo)} = \frac{\text{Salgsinntekter}}{\text{Lønnskostnader}}$$

Lønnskostnader kan være en vesentlig kostnad for norske selskaper. Variabelen omslo antas derfor å være relevant da den gir indikasjoner på hvorvidt selskapet klarer å organisere og motivere ansatte totalt sett slik at disse klarer å dekke sine lønnskostnader.

Andel varelager av kapital (vare)

$$\text{Andel varelager av kapital (vare)} = \frac{\text{Sum varelager}}{\text{totalkapital}}$$

Denne variabelen sier hvor mye av totalkapitalen som er varelager. Andel varelager vil naturlig nok variere fra bransje til bransje, men variabelen antas allikevel å være relevant da den kan gi indikasjoner på om et selskap ikke får omsatt sine varer, og dermed bygger opp et større varelager. Dette er en uheldig utvikling for ethvert selskap, som i verste fall kan bidra til økt sannsynlighet for fremtidig konkurs.

Omløpshastighet kundefordringer (omlopkunde)

$$\text{Omløpshastighet kundefordringer (omlopkunde)} = \frac{\text{Salgsinntekter} * 1,25}{\text{Kundefordringer}}$$

Denne variabelen antas å være relevant for sannsynlighet for konkurs da den gir indikasjoner på hvor lang tid et selskap bruker på å få betalt for sine kundefordringer. Dersom det tar lang tid å få betalt av kunder kan det føre til likviditetsutfordringer, som igjen kan bidra til konkurs. Årsaken til at salgsinntekter multipliseres med 1,25 er antakelsen om at størsteparten av kundene faktureres med 25% merverdiavgift.

Omløpsmidler som andel av total kapital (omlpst)

$$\text{Omløpsmidler som andel av total kapital (omlpst)} = \frac{\text{Sum omløpsmidler}}{\text{total kapital}}$$

Omløpsmidler er eiendeler som forventes å gå i omløp innen ett år. Jo større andel av total kapitalen som er lett omsettelig, jo større kan det antas at sannsynligheten er for at selskapet skal klare seg ved eventuelle likviditetsutfordringer. Dette vil igjen trolig ha innvirkning på sannsynligheten for om et selskap går konkurs, noe som i så fall gjør variabelen omlpst relevant.

Revisoranmerkninger (rev)

Dersom selskapet har fått en eller flere revisoranmerkninger for regnskapsåret så blir denne variabelen satt til 1, om ikke blir den satt til 2.

Når revisor har kommet med noen kommentarer i form av presisering eller forbehold til regnskapet vil det påføres en revisoranmerkning. Dette kan tyde på at selskapet ikke har optimal økonomistyring og full kontroll på regnskaper, noe som igjen kan antas å påvirke sannsynligheten for om et selskap går konkurs.

Appendiks 4: Resultater for ulike parametere for metoder

Tabell 20: ROC-verdier ved ulike antall skjulte lag og noder for metode NN ved estimering av modell basert på variabelsett 1.

Skjulte lag → Noder ↓	1	2	3	4	5	6	9	12	15	18
1	79,8	80,0	75,9	70,9	70,3	74,5	63,0	62,2	43,4	43,6
11	81,0	80,5	82,1	77,8	76,3	72,9	75,3	68,2	67,4	69,6
21	81,9	81,9	78,4	80,3	79,4	79,2	78,1	80,1	75,8	75,2
31	81,2	80,6	81,6	81,5	79,8	77,5	80,9	74,5	75,9	71,1
41	81,0	82,2	81,6	81,5	81,2	80,9	77,5	78,0	78,0	77,5
51	81,6	82,1	81,4	81,2	81,5	82,5	79,4	77,6	74,4	81,2
61	81,9	81,3	82,1	80,5	81,4	82,0	80,6	78,8	77,9	78,2
71	80,8	81,4	81,9	82,0	81,8	81,0	77,3	81,1	78,5	75,3
81	81,9	79,3	81,8	79,6	81,4	80,9	79,0	73,6	76,5	74,0
91	80,9	81,0	81,7	80,9	80,3	79,9	79,0	80,2	76,8	74,5
101	80,8	81,7	78,8	80,7	81,1	81,9	80,6	75,3	80,8	74,7

Tabell 21: ROC-verdi ved ulike antall k for metode KNN ved estimering av modell basert på variabelsett 1. Her benyttes en nærmest tie-breaking-algoritme, og euklidisk avstand for avstandsberegning.

k	1	2	3	4	5	6	7	8	9	10
ROC-verdi	72,1	78,4	81,2	82,7	83,6	84,4	84,8	85,2	85,5	85,6
k	11	12	13	14	15	16	17	18	19	20
ROC-verdi	85,8	85,9	86,1	86,2	86,3	86,3	86,4	86,5	86,6	86,6
k	21	22	23	24	25	26	27	28	29	30
ROC-verdi	86,7	86,7	86,7	86,8	86,8	86,8	86,8	86,8	86,8	86,8
k	31	32	33	34	35	36	37	38	39	40
ROC-verdi	86,8	86,9	86,9	86,9	86,9	86,9	86,9	86,9	86,9	86,8
k	41	42	43	44	45	46	47	48	49	50
ROC-verdi	86,8	86,8	86,8	86,9	86,9	86,9	86,9	86,9	86,9	86,9
k	51	52	53	54	55	56	57	58	59	60
ROC-verdi	86,9	86,9	86,8	86,8	86,8	86,8	86,8	86,8	86,8	86,8
k	61	62	63	64	65	66	67	68	69	70
ROC-verdi	86,8	86,9	86,9	86,8	86,8	86,8	86,8	86,8	86,8	86,8
k	71	72	73	74	75	76	77	78	79	80
ROC-verdi	86,8	86,8	86,8	86,8	86,8	86,8	86,8	86,8	86,8	86,8
k	81	82	83	84	85	86	87	88	89	90
ROC-verdi	86,8	86,8	86,8	86,8	86,8	86,8	86,8	86,8	86,8	86,8
k	91	92	93	94	95	96	97	98	99	100
ROC-verdi	86,7	86,7	86,7	86,7	86,7	86,7	86,7	86,7	86,7	86,7
k	101	102	103	104	105	106	107	108	109	110
ROC-verdi	86,7	86,7	86,7	86,7	86,7	86,7	86,7	86,7	86,7	86,7
k	111	112	113	114	115	116	117	118	119	120
ROC-verdi	86,7	86,7	86,7	86,7	86,7	86,7	86,7	86,7	86,6	86,6
k	121	122	123	124	125	126	127	128	129	130
ROC-verdi	86,6	86,6	86,6	86,6	86,6	86,6	86,6	86,6	86,6	86,6

<i>k</i>	131	132	133	134	135	136	137	138	139	140
ROC-verdi	86,6	86,6	86,6	86,6	86,6	86,6	86,6	86,6	86,6	86,6

Tabell 22: Treffprosent ved ulike antall *k* og avstandsberegninger for metode KNN ved estimering av modell basert på variabelsett 1. Her benyttes en nærmest tie-breaking-algoritme.

<i>k</i>	chebychev	correlation	euclidean	hamming	jaccard	minkowski	spearman	cosine
1	71,4	71,0	71,7	39,7	39,7	71,7	42,4	71,0
11	77,4	76,6	77,5	47,1	47,1	77,5	59,1	76,5
21	78,1	77,2	78,1	50,2	50,2	78,1	67,3	77,1
31	78,2	77,3	78,2	52,3	52,3	78,2	69,3	77,3
41	78,3	77,3	78,3	53,2	53,2	78,3	72,5	77,4
51	78,4	77,7	78,4	55,4	55,4	78,4	74,7	77,7
61	78,4	77,6	78,5	55,8	55,8	78,5	74,6	77,6
71	78,4	77,7	78,4	56,1	56,1	78,4	74,6	77,6
81	78,4	77,8	78,5	56,1	56,1	78,5	76,5	77,7
91	78,4	77,8	78,5	56,5	56,5	78,5	76,5	77,7
101	78,5	77,8	78,6	56,6	56,6	78,6	76,5	77,7
Mean	77,6	76,9	77,7	52,6	52,6	77,7	69,5	76,9

Tabell 23: ROC-verdier ved ulike verdier for kernel scale og box constraint metode SVM ved estimering av modell basert på variabelsett 1. Her benyttes en gaussian kernel-funksjon.

Kernel scale → Box constraint ↓	3	3,2	3,4	3,6	3,8	4	4,2	4,4	4,8	5	Mean
1,1	87,62	87,64	87,64	87,64	87,63	87,62	87,60	87,59	87,56	87,54	87,61
1,2	87,63	87,64	87,65	87,65	87,64	87,63	87,62	87,60	87,57	87,56	87,62
1,3	87,63	87,65	87,66	87,66	87,65	87,64	87,63	87,62	87,58	87,57	87,63
1,4	87,63	87,65	87,67	87,66	87,65	87,65	87,64	87,63	87,59	87,58	87,63
1,5	87,64	87,66	87,67	87,67	87,66	87,65	87,65	87,63	87,60	87,59	87,64
1,6	87,64	87,66	87,68	87,68	87,67	87,66	87,65	87,64	87,61	87,60	87,65
1,7	87,64	87,67	87,69	87,68	87,67	87,66	87,66	87,65	87,62	87,61	87,65
1,8	87,64	87,67	87,69	87,69	87,68	87,67	87,66	87,66	87,63	87,62	87,66
1,9	87,65	87,68	87,69	87,70	87,69	87,68	87,67	87,67	87,64	87,62	87,67
2	87,65	87,68	87,70	87,70	87,69	87,68	87,68	87,67	87,65	87,63	87,67
2,1	87,65	87,68	87,70	87,71	87,70	87,69	87,68	87,68	87,65	87,64	87,68
2,2	87,65	87,69	87,70	87,72	87,71	87,70	87,69	87,68	87,66	87,64	87,68
2,3	87,66	87,69	87,71	87,72	87,71	87,70	87,69	87,69	87,67	87,65	87,69
2,4	87,66	87,69	87,71	87,72	87,72	87,71	87,70	87,69	87,67	87,66	87,69
2,5	87,66	87,69	87,71	87,73	87,73	87,72	87,70	87,69	87,68	87,66	87,70
2,6	87,66	87,69	87,72	87,73	87,73	87,72	87,71	87,70	87,68	87,67	87,70
2,7	87,66	87,69	87,72	87,74	87,74	87,72	87,71	87,70	87,69	87,68	87,70
2,8	87,66	87,69	87,72	87,74	87,74	87,73	87,72	87,71	87,69	87,68	87,71
2,9	87,66	87,69	87,72	87,74	87,74	87,73	87,72	87,71	87,69	87,68	87,71

7,8	87,57	87,62	87,68	87,72	87,75	87,77	87,77	87,77	87,76	87,77	87,72
7,9	87,57	87,62	87,68	87,72	87,75	87,77	87,77	87,77	87,76	87,77	87,72
8	87,57	87,62	87,67	87,72	87,75	87,77	87,77	87,77	87,76	87,77	87,72
Mean	87,62	87,67	87,70	87,73	87,74	87,74	87,74	87,73	87,72	87,71	

Tabell 24: ROC-verdier ved modeller estimert ved CT for ulike score transform-funksjoner, minste antall løv og minste antall noder.

ROC-verdi	Minste antall løv	Minste antall noder	Score transform-funksjon
74,43	1	1	ingen
74,43	1	1	dobbellogaritmisk
74,43	1	1	ismax
74,43	1	1	logaritmisk
74,43	1	1	symmetrisk
74,43	1	1	symmetrisk-logisk
74,43	1	1	symmetrisk-ismax
79,15	1	10	ingen
79,15	1	10	dobbellogaritmisk
74,54	1	10	ismax
79,15	1	10	logaritmisk
79,15	1	10	symmetrisk
79,15	1	10	symmetrisk-logisk
74,54	1	10	symmetrisk-ismax
81,87	5	1	ingen
81,87	5	1	dobbellogaritmisk
75,01	5	1	ismax
81,87	5	1	logaritmisk
81,87	5	1	symmetrisk
81,87	5	1	symmetrisk-logisk
75,01	5	1	symmetrisk-ismax
81,87	5	10	ingen
81,87	5	10	dobbellogaritmisk
75,01	5	10	ismax
81,87	5	10	logaritmisk
81,87	5	10	symmetrisk
81,87	5	10	symmetrisk-logisk
75,01	5	10	symmetrisk-ismax

Tabell 25: ROC-verdier ved ulike verdier for minste antall løv og minste antall noder for metode CT ved estimering av modell basert på variabelsett 1. Her benyttes en symmetrisk score transform-funksjon.

Minste antall løv →	1	6	11	16	21	26	31	36	41	46
Minste antall noder ↓										
1	73,4	83,6	86,9	87,8	87,9	87,9	87,7	87,5	87,6	87,5
6	75,6	83,6	86,9	87,8	87,9	87,9	87,7	87,5	87,6	87,5
11	78,5	83,6	86,9	87,8	87,9	87,9	87,7	87,5	87,6	87,5
16	80,7	84,4	86,9	87,8	87,9	87,9	87,7	87,5	87,6	87,5
21	81,8	85,3	86,9	87,8	87,9	87,9	87,7	87,5	87,6	87,5
26	82,6	85,8	87,1	87,8	87,9	87,9	87,7	87,5	87,6	87,5
31	83,3	86,2	87,3	87,8	87,9	87,9	87,7	87,5	87,6	87,5
36	83,8	86,5	87,3	87,8	87,9	87,9	87,7	87,5	87,6	87,5
41	84,4	86,7	87,4	87,8	87,9	87,9	87,7	87,5	87,6	87,5
46	84,7	86,9	87,5	87,8	87,7	87,9	87,7	87,5	87,6	87,5
51	85,1	87,1	87,6	87,8	87,8	87,9	87,7	87,5	87,6	87,5
56	85,2	87,3	87,6	87,8	87,8	87,9	87,7	87,5	87,6	87,5
61	85,5	87,3	87,5	87,8	87,9	87,8	87,7	87,5	87,6	87,5
66	85,7	87,4	87,5	87,8	87,8	87,7	87,6	87,5	87,6	87,5
71	86,0	87,5	87,5	87,7	87,7	87,7	87,6	87,5	87,6	87,5
76	86,1	87,5	87,6	87,7	87,8	87,7	87,6	87,5	87,6	87,5
81	86,3	87,6	87,6	87,7	87,8	87,7	87,6	87,5	87,6	87,5
86	86,5	87,7	87,7	87,7	87,8	87,6	87,6	87,5	87,6	87,5
91	86,6	87,7	87,7	87,8	87,8	87,6	87,5	87,5	87,5	87,5
96	86,7	87,8	87,4	87,7	87,7	87,6	87,5	87,5	87,5	87,5
101	86,8	87,7	87,4	87,5	87,7	87,6	87,5	87,5	87,5	87,5
Mean	83,6	86,5	87,3	87,7	87,8	87,8	87,6	87,5	87,6	87,5
Minste antall løv →	51	56	61	66	71	76	81	86	91	96
Minste antall noder ↓										
1	87,4	87,3	87,3	87,2	87,2	87,2	87,1	87,1	87,1	87,1
6	87,4	87,3	87,3	87,2	87,2	87,2	87,1	87,1	87,1	87,1
11	87,4	87,3	87,3	87,2	87,2	87,2	87,1	87,1	87,1	87,1
16	87,4	87,3	87,3	87,2	87,2	87,2	87,1	87,1	87,1	87,1
21	87,4	87,3	87,3	87,2	87,2	87,2	87,1	87,1	87,1	87,1
26	87,4	87,3	87,3	87,2	87,2	87,2	87,1	87,1	87,1	87,1
31	87,4	87,3	87,3	87,2	87,2	87,2	87,1	87,1	87,1	87,1
36	87,4	87,3	87,3	87,2	87,2	87,2	87,1	87,1	87,1	87,1
41	87,4	87,3	87,3	87,2	87,2	87,2	87,1	87,1	87,1	87,1
46	87,4	87,3	87,3	87,2	87,2	87,2	87,1	87,1	87,1	87,1
51	87,4	87,3	87,3	87,2	87,2	87,2	87,1	87,1	87,1	87,1

56	87,4	87,3	87,3	87,2	87,2	87,2	87,1	87,1	87,1	87,1
61	87,4	87,3	87,3	87,2	87,2	87,2	87,1	87,1	87,1	87,1
66	87,4	87,3	87,3	87,2	87,2	87,2	87,1	87,1	87,1	87,1
71	87,4	87,3	87,3	87,2	87,2	87,2	87,1	87,1	87,1	87,1
76	87,4	87,3	87,3	87,2	87,2	87,2	87,1	87,1	87,1	87,1
81	87,4	87,3	87,3	87,2	87,2	87,2	87,1	87,1	87,1	87,1
86	87,4	87,3	87,3	87,2	87,2	87,2	87,1	87,1	87,1	87,1
91	87,4	87,3	87,3	87,2	87,2	87,2	87,1	87,1	87,1	87,1
96	87,4	87,3	87,3	87,2	87,2	87,2	87,1	87,1	87,1	87,1
101	87,4	87,3	87,3	87,2	87,2	87,2	87,1	87,1	87,1	87,1
Mean	87,4	87,3	87,3	87,2	87,2	87,2	87,1	87,1	87,1	87,1

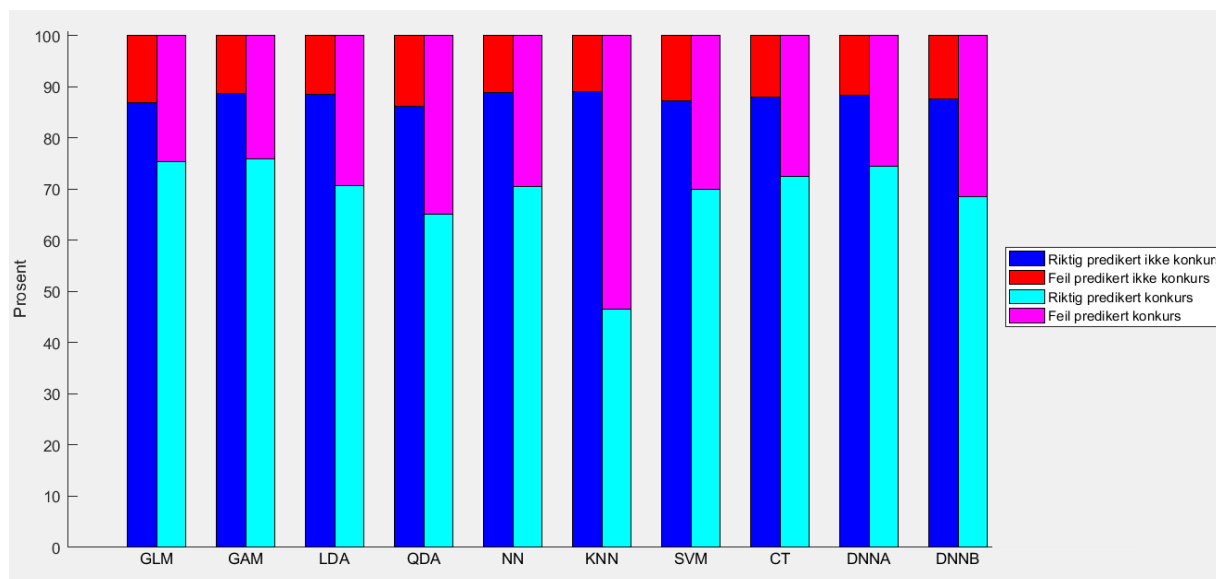
Tabell 26: ROC-verdier ved ulike antall skjulte lag for autoencoder 1 og 2 for metode DNNA ved estimering av modell basert på variabelsett 3.

skjulte lag for autoencoder 1 → skjulte lag for autoencoder 2 ↓	1	2	3	4	5	6	7	8	9	10
1	69,35	62,49	81,66	89,57	89,56	89,48	89,46	89,39	89,54	89,61
2	40,32	39,82	89,22	89,46	89,47	89,63	89,55	89,52	89,61	89,59
3	69,35	62,77	89,54	89,64	89,52	89,66	89,84	89,58	89,79	89,71
4	69,35	37,64	89,65	89,55	89,63	89,59	89,64	89,57	89,77	89,76
5	30,68	62,14	89,55	89,58	89,53	89,82	89,75	89,61	89,89	89,79
6	69,35	38,16	89,52	89,54	89,49	89,88	89,50	89,83	89,81	89,73
7	69,31	37,78	89,58	89,49	89,54	89,77	89,80	89,70	89,88	89,80
8	30,65	38,63	89,44	89,55	89,66	89,59	89,69	89,55	89,84	89,87
9	69,35	37,42	81,13	89,64	89,56	89,82	89,91	89,75	89,81	89,86
10	30,65	62,16	85,27	89,55	89,51	89,79	89,82	89,64	89,76	89,86
13	69,35	38,00	89,62	89,50	89,63	89,74	89,56	89,81	89,86	89,80
16	69,35	37,57	83,15	89,59	89,54	89,74	89,57	89,66	89,78	89,83
19	69,35	87,48	89,56	89,56	89,60	89,69	89,88	89,53	89,77	89,81

Tabell 27: ROC-verdier ved ulike antall skjulte lag og noder for metode DNNB ved estimering av modell basert på variabelsett 1.

Skjulte lag → Noder ↓	1	2	3	4	5	6	7	8	9	10	Mean
1	83,5	83,3	82,6	69,7	55,6	54,1	46,2	46,2	53,8	53,8	62,9
2	83,6	84,5	84,2	84,4	83,6	75,3	34,7	63,8	72,9	40,5	70,8
3	84,4	83,6	83,8	84,0	84,6	45,3	76,9	49,4	20,7	60,2	67,3
4	84,4	84,5	84,8	84,4	84,0	84,7	82,7	38,1	71,9	50,2	75,0
5	84,4	84,9	84,6	84,7	84,4	84,8	84,0	83,6	71,5	23,8	77,1
6	84,5	84,7	84,7	84,7	84,8	84,5	84,7	84,7	50,4	70,9	79,9
7	84,5	84,3	85,0	84,8	84,9	84,7	83,9	84,2	84,0	83,1	84,3
8	84,6	84,8	84,8	84,6	85,2	84,4	85,0	84,7	84,7	84,5	84,7
9	84,6	84,8	85,0	84,9	84,8	85,0	84,9	84,8	84,9	84,1	84,8
10	84,9	84,6	84,9	84,9	84,9	84,9	84,7	84,7	84,9	84,4	84,8
Mean	84,3	84,4	84,4	83,1	81,7	76,8	74,8	70,4	68,0	63,6	

Appendiks 5: Resultater fra variabelsett 1 uten transformering av uavhengige variabler til standardisert verdi



Figur 17: Feilklassifiseringsrate for modell basert på variabelsett 3, uten transformering av uavhengige variabler til standardisert verdi.

Tabell 28: ROC-verdi og treffprosent for modell basert på variabelsett 3, uten transformering av uavhengige variabler til standardisert verdi.

Tradisjonelle statistiske metoder	ROC-verdi	Treffprosent
Generell Lineær Modell (GLM)	89,6%	83,0%
Generell Additiv Modell (GAM)	91,1%	84,4%
Lineær diskriminantanalyse (LDA)	89,0%	82,5%
Kvadratisk diskriminantanalyse (QDA)	85,0%	79,1%
Maskinlæringsteknikker	ROC-verdi	Treffprosent
Tolags forovermatet nevralt nettverk (NN)	89,2%	82,8%
<i>k</i> -nærmeste naboer (KNN)	82,3%	74,8%
Støttevektormaskin (SVM)	88,2%	81,5%
Klassifiseringstrær (CT)	89,3%	82,8%
Dypt nevralt nettverk skapt av to autoencodere (DNNA)	90,0%	83,6%
Dypt nevralt nettverk med back-propagation (DNNB)	87,8%	81,2%

Appendiks 6: McNemars test og *p*-verdier

Tabell 29: *p*-verdier ved McNemars test ved sammenligning av modeller basert på variabelsett 1 og estimert ved ulike metoder.

	GLM	GAM	LDA	QDA	NN	KNN	SVM	CT	DNNA	DNNB
GLM		0,002	0,702	0,020	0,813	0,102	0,039	0,040	0,966	0,020
GAM	0,002		0,001	0,000	0,001	0,161	0,331	0,327	0,003	0,472
LDA	0,702	0,001		0,054	0,892	0,043	0,014	0,014	0,663	0,007
QDA	0,020	0,000	0,054		0,038	0,000	0,000	0,000	0,018	0,000
NN	0,813	0,001	0,892	0,038		0,060	0,021	0,021	0,772	0,010
KNN	0,102	0,161	0,043	0,000	0,060		0,674	0,680	0,114	0,500
SVM	0,039	0,331	0,014	0,000	0,021	0,674		0,998	0,044	0,808
CT	0,040	0,327	0,014	0,000	0,021	0,680	0,998		0,045	0,802
DNNA	0,966	0,003	0,663	0,018	0,772	0,114	0,044	0,045		0,023
DNNB	0,020	0,472	0,007	0,000	0,010	0,500	0,808	0,802	0,023	

Tabell 30: *p*-verdier ved McNemars test ved sammenligning av modeller basert på variabelsett 2 og estimert ved ulike metoder.

	GLM	GAM	LDA	QDA	NN	KNN	SVM	CT	DNNA	DNNB
GLM		0,009	0,000	0,000	0,000	0,025	0,020	0,068	0,007	0,012
GAM	0,009		0,000	0,000	0,000	0,718	0,781	0,434	0,000	0,916
LDA	0,000	0,000		0,402	0,779	0,000	0,000	0,000	0,086	0,000
QDA	0,000	0,000	0,402		0,258	0,000	0,000	0,000	0,010	0,000
NN	0,000	0,000	0,779	0,258		0,000	0,000	0,000	0,154	0,000
KNN	0,025	0,718	0,000	0,000	0,000		0,941	0,681	0,000	0,805
SVM	0,020	0,781	0,000	0,000	0,000	0,941		0,621	0,000	0,871
CT	0,068	0,434	0,000	0,000	0,000	0,681	0,621		0,000	0,505
DNNA	0,007	0,000	0,086	0,010	0,154	0,000	0,000	0,000		0,000
DNNB	0,012	0,916	0,000	0,000	0,000	0,805	0,871	0,505	0,000	

Tabell 31: *p*-verdier ved McNemars test ved sammenligning av modeller basert på variabelsett 3 og estimert ved ulike metoder.

	GLM	GAM	LDA	QDA	NN	KNN	SVM	CT	DNNA	DNNB
GLM		0,409	0,748	0,015	0,905	0,903	0,714	0,886	0,685	0,627
GAM	0,409		0,247	0,001	0,339	0,338	0,653	0,327	0,681	0,741
LDA	0,748	0,247		0,036	0,848	0,850	0,486	0,867	0,461	0,414
QDA	0,015	0,001	0,036		0,021	0,021	0,005	0,023	0,004	0,003
NN	0,905	0,339	0,848	0,021		0,994	0,619	0,989	0,592	0,538
KNN	0,903	0,338	0,850	0,021	0,994		0,618	0,991	0,591	0,537
SVM	0,714	0,653	0,486	0,005	0,619	0,618		0,603	0,977	0,913
CT	0,886	0,327	0,867	0,023	0,989	0,991	0,603		0,576	0,522
DNNA	0,685	0,681	0,461	0,004	0,592	0,591	0,977	0,576		0,944
DNNB	0,627	0,741	0,414	0,003	0,538	0,537	0,913	0,522	0,944	

Tabell 32: p-verdier ved McNemars test ved sammenligning av modeller estimert ved GAM og basert på ulike variabelsett.

	Variabelsett 1	Variabelsett 2	Variabelsett 3
Variabelsett 1		0,338	0,495
Variabelsett 2	0,338		0,099
Variabelsett 3	0,495	0,099	

Tabell 33: p-verdier ved McNemars test ved sammenligning av modeller estimert ved DNNB og basert på ulike variabelsett.

	Variabelsett 1	Variabelsett 2	Variabelsett 3
Variabelsett 1		0,732	0,284
Variabelsett 2	0,732		0,155
Variabelsett 3	0,284	0,155	

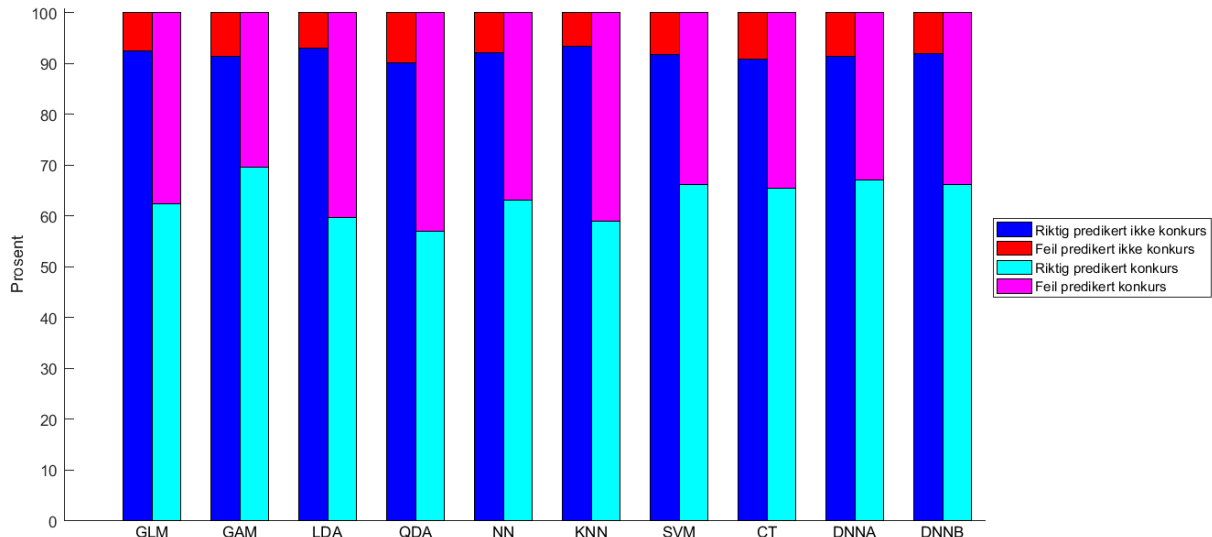
Appendiks 7: Resultater ved flere ikke-konkursselskaper

I dette appendikset presenteres prediksjonsevnen til modeller basert på variabelsett 3, ved et større datautvalg enn det som er benyttet i oppgaven. Som forklart i kapittel 3.9 gir et større datautvalg høyere treffprosent, men på bekostning av flere type 1-feil. ROC-verdi er tilnærmet uforandret. Det er gjennomført transformering av uavhengige variabler til standardisert verdi ved estimering av modellene i dette appendikset, slik det også er i oppgaven.

Dette appendikset illustrerer altså hvorfor det ikke er valgt flere ikke-konkursselskaper i datautvalget, samt hvorfor ROC-verdi er et bedre evalueringsmål enn treffprosent.

Datautvalg bestående av 41 580 selskaper

Her består datautvalget av samtlige 10 395 konkursselskaper og 31 185 tilfeldig valgte ikke-konkursselskaper.



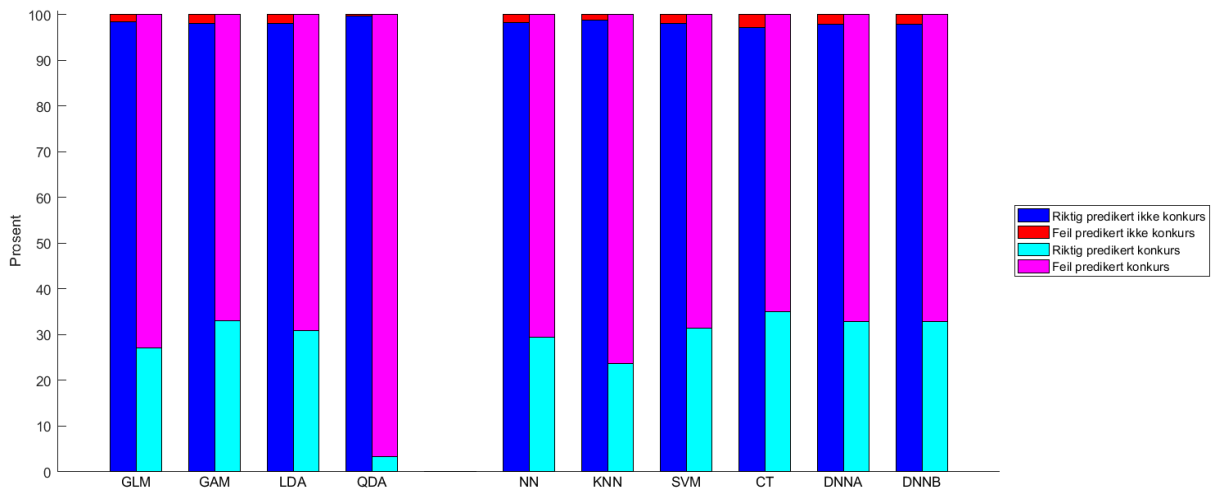
Figur 18: Feilklassifiseringsrate for modell basert på variabelsett 3, og med 41 580 selskaper i datautvalget.

Tabell 34: ROC-verdi og treffprosent for modell basert på variabelsett 3 og med 41 580 selskaper i datautvalget.

Tradisjonelle statistiske metoder	ROC-verdi	Treffprosent
Generell Lineær Modell (GLM)	89,6%	85,0%
Generell Additiv Modell (GAM)	91,1%	86,0%
Lineær diskriminantanalyse (LDA)	89,2%	84,7%
Kvadratisk diskriminantanalyse (QDA)	85,2%	81,8%
Maskinlæringsteknikker	ROC-verdi	Treffprosent
Tolags forovermatet nevralt nettverk (NN)	89,3%	84,9%
<i>k</i> -nærmeste naboer (KNN)	89,6%	84,8%
Støttevektormaskin (SVM)	89,9%	85,4%
Klassifiseringstrær (CT)	89,3%	84,5%
Dypt nevralt nettverk skapt av to autoencodere (DNNA)	90,1%	85,3%
Dypt nevralt nettverk med back-propagation (DNNB)	90,0%	85,4%

Datautvalg bestående av 114 345 selskaper

Her består datautvalget av samtlige 10 395 konkurselskaper og 103 950 tilfeldig valgte ikke-konkurselskaper.



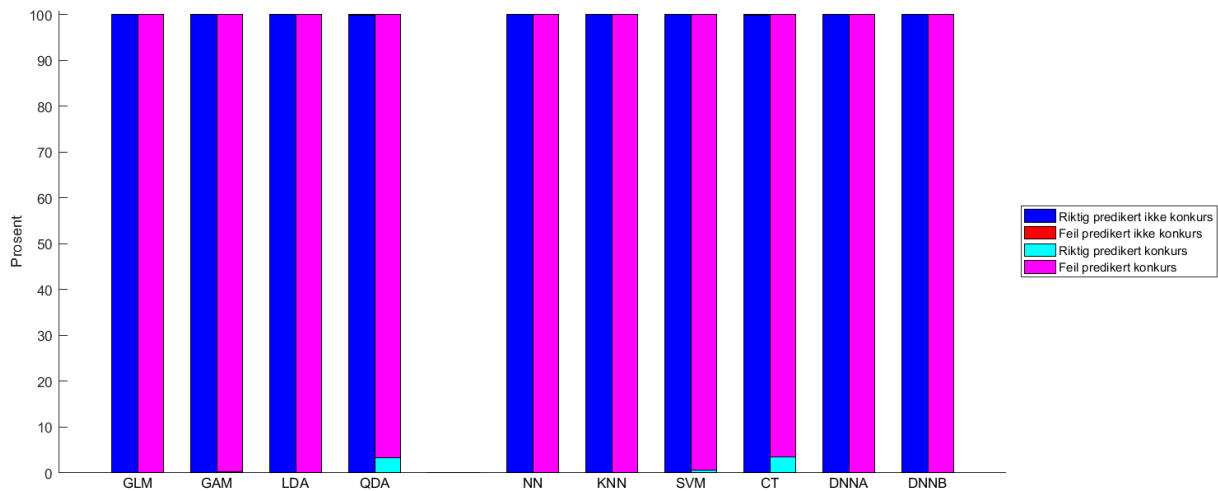
Figur 19: Feilklassifiseringsrate for modell basert på variabelsett 3, og med 114 345 selskaper i datautvalget.

Tabell 35: ROC-verdi og treffprosent for modell basert på variabelsett 3 og med 114 345 selskaper i datautvalget.

Tradisjonelle statistiske metoder	ROC-verdi	Treffprosent
Generell Lineær Modell (GLM)	89,5%	91,9%
Generell Additiv Modell (GAM)	91,1%	92,2%
Lineær diskriminantanalyse (LDA)	89,6%	91,9%
Kvadratisk diskriminantanalyse (QDA)	85,4%	90,9%
Maskinlæringsteknikker	ROC-verdi	Treffprosent
Tolags forovermatet nevralt nettverk (NN)	89,5%	92,0%
<i>k</i> -nærmeste naboer (KNN)	89,7%	91,9%
Støttevektormaskin (SVM)	79,1%	92,0%
Klassifiseringstrær (CT)	88,4%	91,5%
Dypt nevralt nettverk skapt av to autoencodere (DNNA)	90,0%	92,0%
Dypt nevralt nettverk med back-propagation (DNNB)	89,8%	92,0%

Datautvalg bestående av 738 045 selskaper

Her består datautvalget av samtlige 10 395 konkurs-selskaper og 727 650 tilfeldig valgte ikke-konkurs-selskaper.



Figur 20: Feilklassifiseringsrate for modell basert på variabelsett 3, og med 738 045 selskaper i datautvalget.

Tabell 36: ROC-verdi og treffprosent for modell basert på variabelsett 3 og med 738 045 selskaper i datautvalget.

Tradisjonelle statistiske metoder	ROC-verdi	Treffprosent
Generell Lineær Modell (GLM)	88,9%	98,6%
Generell Additiv Modell (GAM)	91,0%	98,6%
Lineær diskriminantanalyse (LDA)	89,5%	98,6%
Kvadratisk diskriminantanalyse (QDA)	84,4%	98,4%
Maskinlæringsteknikker	ROC-verdi	Treffprosent
Tolags forovermatet nevralt nettverk (NN)	86,3%	98,6%
<i>k</i> -nærmeste naboer (KNN)	88,4%	98,6%
Støttevektormaskin (SVM)	51,3%	98,6%
Klassifiseringstrær (CT)	84,9%	98,5%
Dypt nevralt nettverk skapt av to autoencodere (DNNA)	89,0%	98,6%
Dypt nevralt nettverk med back-propagation (DNNB)	84,7%	98,6%

Appendiks 8: Antall årsregnskaper i datautvalget

Tabell 37: Antall årsregnskaper fra aksje- og allmennaksjeselskaper i datautvalget, fordelt per regnskapsår og per del ved den fem-delte kryssvalideringen.

	Del 1		Del 2	
	Konkurs	Ikke konkurs	Konkurs	Ikke konkurs
2005	169	338	169	338
2006	195	391	195	391
2007	354	708	354	708
2008	301	601	301	601
2009	290	580	290	580
2010	245	490	245	490
2011	253	507	253	507
2012	272	543	272	543
Sum	2 079	4 158	2 079	4 158

	Del 3		Del 4	
	Konkurs	Ikke konkurs	Konkurs	Ikke konkurs
2005	169	338	169	338
2006	195	391	195	391
2007	354	708	354	708
2008	301	601	301	601
2009	290	580	290	580
2010	245	490	245	490
2011	253	507	253	507
2012	272	543	272	543
Sum	2 079	4 158	2 079	4 158

	Del 5		Sum alle deler	
	Konkurs	Ikke konkurs	Konkurs	Ikke konkurs
2005	169	338	845	1 690
2006	195	391	975	1 955
2007	354	708	1 770	3 540
2008	301	601	1 505	3 005
2009	290	580	1 450	2 900
2010	245	490	1 225	2 450
2011	253	507	1 265	2 535
2012	272	543	1 360	2 715
Sum	2 079	4 158	10 395	20 790