Masteroppgave

NTNU
Norges teknisk-naturvitenskapelige universitet
Fakultet for økonomi
NTNU Handelshøyskolen

Jenny Fjerstad og Kristin Hamnes

Prediksjon av konkurs i bygg- og anleggsbransjen

En studie av variabler i konkursprediksjonsmodeller for SMB i bygg- og anleggsbransjen

Masteroppgave i Økonomi og administrasjon
Veileder: Ranik Raaen Wahlstrøm
Mai 2023

NTNU
Kunnskap for en bedre verden

Jenny Fjerstad og Kristin Hamnes

Prediksjon av konkurs i bygg- og anleggsbransjen

En studie av variabler i konkursprediksjonsmodeller for SMB i bygg- og anleggsbransjen

Masteroppgave i Økonomi og administrasjon
Veileder: Ranik Raaen Wahlstrøm
Mai 2023

Norges teknisk-naturvitenskapelige universitet
Fakultet for økonomi
NTNU Handelshøyskolen

NTNU
Kunnskap for en bedre verden

Forord

Denne oppgaven er det avsluttende arbeidet på vår mastergrad innen økonomi og administrasjon, ved NTNU Handelshøyskolen. Vi ser tilbake på fem lærerike og fine år på NTNU Handelshøyskolen. Masteroppgaven utgjør 30 studiepoeng innen hovedprofilen Business Analytics, og er skrevet våren 2023. Arbeidet med masteroppgaven har bydd på både opp- og nedturer. Det har i perioder vært krevende, men til gjengjeld veldig lærerrikt. Å få muligheten til å utforme sin egen oppgave og sette seg inn i temaet konkurspredikskjon har vært spennende.

Vi vil rette en stor takk til vår veileder, Ranik Raaen Wahlstrøm, for konstruktive tilbakemeldinger og gode råd, og ikke minst tiden han har brukt sammen med oss i prosessen. Vi vil også takke han for å ha bidratt med datagrunnlaget i oppgaven. Til slutt takker vi også familie og venner for støtte, samt medstudenter som har hjulpet oss gjennom prosessen.

Innholdet i denne oppgaven står for forfatterenes regning.

Trondheim, mai 2023

Jenny Fjerstad

Kristin Hamnes

Sammendrag

Formålet med oppgaven er å bidra med ny innsikt innen viktige variabler i konkursprediksjonsmodeller for SMB i bygg- og anleggsbransjen, og kunne si noe om nytten av å utvikle konkursprediksjonsmodeller for en spesifikk bransje. Vi ønsker å undersøke om variablene vi finner skiller seg ut for bransjen, og om modeller utviklet for bygg- og anleggsbransjen presterer bedre enn modeller utviklet for alle bransjer. Følgelig er vår problemstilling: *Hvilke variabler egner seg til å predikere konkurs for SMB i bygg- og anleggsbransjen, og forbedres modellene av å utvikles for en spesifikk bransje?*

Datagrunnlaget i oppgaven består av ukonsoliderte årsregnskaper for norske private og børsnoterte selskaper, fra perioden 2006 til 2020. For å finne hvilke variabler som egner seg best, benytter vi et variabelsett bestående av 160 variabler, der 155 er basert på regnskapstall som også ble benyttet i Paraschiv mfl. (2021), samt fem makroøkonomiske variabler. Først benytter vi metoden Least Absolute Shrinkage and Selection Operator (LASSO), til å velge variabler som inkluderes i logistiske regresjonsmodeller. Deretter utvikler vi Extreme Gradient Boosting (XGBoost)-modeller og undersøker viktigheten av variablene ved bruk av SHapley Additive exPlanations (SHAP). Variablene vi finner som viktige sammenlignes med tre eksisterende variabelsett, utviklet av Altman og Sabato (2007), Paraschiv mfl. (2021) og variablene i SEBRA-modellen. For å undersøke hvorvidt modellene forbedres når de utvikles for bygg- og anleggsbransjen, sammenligner vi modellenes evalueringsmål.

Vi finner flere variabler som egner seg til prediksjon av konkurs for SMB i bygg- og anleggsbransjen. Det er noe variasjon i hvilke variabler vi anser som viktige ved bruk av ulike modeller. Basert på både LASSO og XGBoost med SHAP, finner vi at variabler innen kategoriene belåning, likviditet og alder er viktige. Ved bruk av LASSO finner vi også en viktig variabel innen kategorien lønnsomhet. Innen kategorien soliditet anses flere variabler som viktige, basert på SHAP-verdiene. De makroøkonomiske variablene anses ikke som viktige. Vi finner flere viktige variabler som er en del av tidligere utviklede variabelsett, men også flere som skiller seg ut for bransjen. Ved å sammenligne evalueringsmålene for modellene, finner vi at modeller utviklet for bygg- og anleggsbransjen presterer noe bedre enn modeller utviklet for flere bransjer.

Abstract

The objective of our thesis is to contribute with new insight regarding important variables for bankruptcy prediction for SMEs in the construction industry, and comment on the value of developing bankruptcy prediction models for particular industries. We want to examine whether the variables we find stands out for the industry, and if the models perform better when they are developed specifically for the construction industry. Hence, the question we answer in our thesis is: *Which variables are suited to predict bankruptcy of SMEs in the construction industry, and are the models improved by being developed for a specific industry?*

The dataset in our thesis consists of unconsolidated annual Norwegian accounts from private and listed companies, from the period 2006 to 2020. To find the most suitable variables, we use a variable set consisting of 160 variables, where 155 of them are based on accounting figures from the study of Paraschiv mfl. (2021), as well as five macroeconomic variables. To begin with, we use the method Least Absolute Shrinkage and Selection Operator (LASSO) to select variables to include in logistic regression models. Further, we develop Extreme Gradient Boosting (XGBoost) models and examine the importance of the variables by using SHapley Additive exPlanations (SHAP). We compare the most important variable findings with three existing variable sets, developed by Altman og Sabato (2007), Paraschiv mfl. (2021) and the variables from the SEBRA-model. To examine whether models are improved when they are developed for the construction industry, we compare the evaluation metrics for the models.

We find a number of important variables for predicting bankruptcy for SMEs in the construction industry. However, which variables we regard as important, depends on the models used. Based on both LASSO and XGBoost with SHAP, we find that variables within the categories of leverage, liquidity, and age are important. When using LASSO, we also find an important variable within the profitability category. Within the category solidity, several variables are considered important, according to the SHAP-values. None of the macroeconomic variables are considered important. Several of the variables we find to be important, are also a part of previously developed variable sets, but some variables appear to be of greater importance for the construction industry. By comparing the evaluation metrics for the models, we find that models developed for the construction industry perform slightly better than models developed for all industries.

Innhold

Figurer

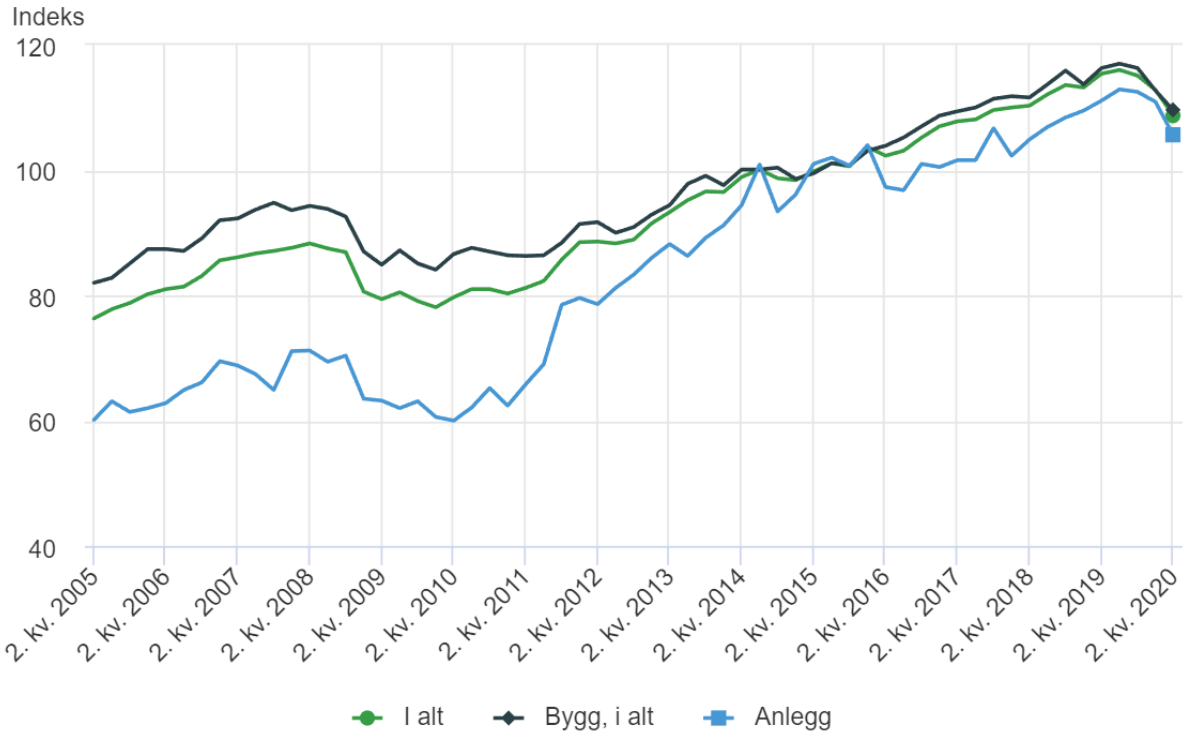
Tabeller

Kapittel 1

Innledning

I løpet av 2022 gikk 3 713 norske selskaper konkurs (Statistisk sentralbyrå, 2023c). Dersom et selskap går konkurs vil det berøre flere interessenter, deriblant kunder, kreditorer, investorer, aksjeeiere, leverandører og ansatte (Kumar og Ravi, 2007; Pelja og Wahlstrøm, 2021). Interessenter kan lide økonomisk tap, i form av blant annet å miste arbeidsplassen, ikke få levert varer som er bestilt eller ikke få dekt sine krav. Konkurser vil kunne ha negative virkninger i samfunnet. Det er derfor nyttig for mange å kunne forutsi hvilke selskaper som går konkurs, eller hvilke selskaper som står i fare for å gå konkurs.

Bygg- og anleggsvirksomhet er den næringen som hadde flest konkurser i Norge i 2022 (Statistisk sentralbyrå, 2023c). Historisk sett har næringen hatt en stor andel av konkursene i landet. Bygg- og anleggsbransjen er sentral for utviklingen av samfunnet og per 2021 var det 263 672 sysselsatte innen næringen (Statistisk sentralbyrå, 2022b). I 2021 var det 57 861 foretak i næringen og omsetningen var på 674 761 millioner kroner. Utviklingen i omsetning, antall sysselsatte og antall foretak har vært relativt jevn de siste årene (Statistisk sentralbyrå, 2022b). Bygg- og anleggsbransjen påvirkes i stor grad av makroøkonomiske forhold. Etter finanskrisen i 2008 og koronapandemien i 2020, var det et fall i aktiviteten i næringen (Haugen, 2020). Dette fremgår av figur 1.1, som viser utviklingen i aktiviteten i perioden 2005-2020, ved sesongjustert produksjonsindeks. De siste årene har bygg- og anleggsnæringen blant annet vært påvirket av koronapandemien, krigen i Ukraina, energikrise og høy inflasjon. Dette har i noen deler av næringen ført til mangel og økte priser på råvarer, og utsettelse av prosjekter. Prisøkning på varer og tjenester i 2022 skyldtes blant annet økte kostnader på råvarer, ferdigvarer og energi (BDO, 2022). Siden september 2021 har rentene steget (Norges Bank, 2023), noe som blant annet har påvirket etterspørselen i markedet (BDO, 2022). Bransjen påvirkes også av retningslinjer, lover og befolkningsutviklingen i landet, samt offentlige investeringer (BDO, 2022).



Figur 1.1: Produksjonsindeks for bygg- og anleggsbransjen i perioden 2005-2020. Det er en månedlig volumindeks, som baseres på timeverk innen næringen. Den er sesongjustert og fremstilt kvartalsvis (Haugen, 2020).

Mange ulike metoder er benyttet til å predikere konkurs, deriblant multivariat diskriminantanalyse, logistisk regresjon og nevrale nettverk (Kumar & Ravi, 2007). Konkursprediksjonsmodeller baseres også på ulike variabler, som er valgt ved bruk av ulike metoder. Imidlertid er det ingen konsensus blant hverken akademikere eller praktikere om hvilke variabler som skal benyttes i modeller for konkursprediksjon (Paraschiv mfl., 2021; Tian mfl., 2015). Chava og Jarrow (2004) peker på viktigheten av å inkludere industrieffekter i konkursprediksjonsmodeller. Alaka mfl. (2016) fokuserer i sin studie på kritiske faktorer som forklarer konkurs i bygg- og anleggsbransjen. De fant flere sentrale kategorier for bransjen, men kom ikke frem til sentrale variabler. Det kan derfor være interessant å se nærmere på hvilke variabler som er viktige for konkursprediksjon i bygg- og anleggsbransjen. I følge Pelja og Wahlstrøm (2021) er generelle konkursprediksjonsmodeller mindre nøyaktige ved konkursprediksjon for store, og SMB. Utvikling av konkursprediksjonsmodeller for spesifikke bedriftsstørrelser kan gjøre modellene mer nøyaktige. Altman og Sabato (2007) argumenterer også for nytten av å utvikle konkursprediksjonsmodeller for spesifikke størrelser. I Norge utgjør SMB omtrent 99% av norske selskaper (Statistisk sentralbyrå, 2023b), og de bidrar med en stor del av verdiskapingen i landet. Å avgrense konkursprediksjonsmodeller til å kun se på SMB, kan dermed være interessant. Vi ønsker

å bidra med ny innsikt innen hvilke variabler som egner seg innen konkurspredikSjon for SMB i bygg- og anleggsbransjen. I tillegg ønsker vi å si noe om nytten av å utvikle konkurspredikSjonsmodeller for en spesifikk bransje.

Vi har formulert følgende problemstilling:

Hvilke variabler egner seg til å predikere konkurs for SMB i bygg- og anleggsbransjen, og forbedres modellene av å utvikles for en spesifikk bransje?

For å besvare problemstillingen benytter vi to metodiske rammeverk. Først benytter vi Least Absolute Shrinkage and Selection Operator (LASSO), som er funnet å være godt egnet ved konkurspredikSjon (Paraschiv mfl., 2021; Tian mfl., 2015). Variablene benyttes i *discrete hazard models* med logistisk regresjon, som er en av de mest brukte metodene innen konkurspredikSjon (Paraschiv mfl., 2021). I tillegg predikerer vi konkurs med Extreme Gradient Boosting (XGBoost) og benytter SHapley Additive exPlanations (SHAP) for variabelseleksjon. XGBoost er en maskinlæringsteknikk som kan gi nøyaktige resultater ved prediksjon (Chen & Guestrin, 2016), og sammen med SHAP, har metoden vist seg å være egnet til variabelseleksjon (Jabeur mfl., 2022; Lin & Bai, 2022). For å avgjøre hvilke variabler som egner seg, ser vi på et utvalg av 160 variabler. 155 av variablene er basert på regnskapstall og hentet fra studien til Paraschiv mfl. (2021). Fem av variablene er makroøkonomiske variabler, som inkluderes med bakgrunn i at bygg- og anleggsbransjen er konjunktursensitiv, og at Hol (2007) peker på flere makroøkonomiske faktorer som viktig innen konkurspredikSjon.

Vi sammenligner variablene vi identifiserer som best egnet med tre eksisterende variabelsett fra litteraturen. Det første variabelsettet vi sammenligner med er utviklet av Altman og Sabato (2007), som er utviklet for å predikere konkurs hos SMB. Det andre variabelsettet er hentet fra SEBRA-modellen, og er utviklet av Norges Bank, spesifikt for norske selskaper (Bernhardsen og Larsen, 2007). Det tredje variabelsettet er variablene som Paraschiv mfl. (2021) fant at gir best prediksjonsevne blant norske SMB, ved bruk av LASSO. Vi velger å sammenligne med disse variabelsettene for å undersøke om vi finner de samme variablene, og om variablene egner seg bedre til å predikere konkurs for bygg- og anleggsbransjen, eller om tidligere variabelsett er minst like gode. De tidligere variabelsettene er ikke utviklet for en spesifikk bransje. Vi foretar også variabelseleksjon for alle bransjer ved bruk av XGBoost og SHAP, ettersom vi ikke har et tidligere variabelsett å sammenligne med, som er utviklet ved bruk av disse metodene.

For å undersøke i hvilken grad modellene forbedres av bransjespesifikke variabler, sammenligner vi evalueringsmål for modellene. Vi sammenligner logistiske regresjonsmodeller ved bruk

av variablene i tre eksisterende variabelsett og variablene vi finner med LASSO. I tillegg sammenligner vi prestasjonen til XGBoost-modeller utviklet med data som består av årsregnskaper for bygg- og anleggsselskaper med XGBoost-modeller utviklet med data som består av årsregnskaper for flere bransjer.

Vi finner flere viktige variabler for prediksjon av konkurs for SMB i bygg- og anleggsbransjen. Det er noe variajson i hvilke variabler som anses som viktige ved bruk av LASSO og XGBoost med SHAP, og hvilke kategorier de representerer. Fire variabler anses som viktigst basert på både LASSO og XGBoost med SHAP, og vi anser dermed disse variablene blant de best egnende for prediksjon av konkurs i bygg- og anleggsbransjen. De gir informasjon om selskapets likviditet, belåning og alder. Ved bruk av LASSO finner vi også en viktig variabel innen kategorien lønnsomhet. Av SHAP-verdiene ser vi at flere variabler innenfor kategorien soliditet er representert. Ved sammenligning med tre eksisterende variabelsett finner vi flere likheter. Noen variabler skiller seg riktignok ut for bygg- og anleggsbransjen. Evalueringsmålene for de logistiske regresjonsmodellene og XGBoost-modellene tyder også på at konkursprediksjonsmodeller forbedres til en viss grad, når de er utviklet for bygg- og anleggsbransjen.

Vår forskning vil kunne være nyttig for blant annet banker som tilbyr lån til SMB innen bygg- og anleggsbransjen, investorer, leverandører, samarbeidspartnere og andre som ønsker å utvikle konkursprediksjonsmodeller. For banker kan det være nyttig å vite hvilke variabler som er sentrale for at et selskap går konkurs, og for investorer dersom de skal investere i et selskap. Vår oppgave vil gi innsikt i nytten av å utvikle konkursprediksjonsmodeller for bygg- og anleggsbransjen. For banker som benytter konkursprediksjonsmodeller, vil en forbedring i prestasjon kunne gi store utslag på lønnsomheten til bankene (Paraschiv mfl., 2021). Innsikten vi bidrar med kan benyttes til å utvikle bransjespesifikke konkursprediksjonsmodeller. Modeller utviklet for spesifikke bransjer kan for eksempel benyttes til å gjøre en samlet kredittrisikovurdering av bransjer.

For å belyse problemstillingen vil vi først i kapittel 2 redegjøre for tidligere forskning, før vi gjennomgår datagrunnlaget for oppgaven i kapittel 3. Videre i kapittel 4 legges forskningsmetoden frem, og i kapittel 5 presenterer vi resultatene. Til slutt diskuterer vi resultatene og gir en konklusjon på problemstillingen, i henholdsvis kapittel 6 og 7.

Kapittel 2

Tidligere forskning

I dette kapittelet presenterer vi relevant litteratur for oppgaven. Først presenteres en overordnet oversikt over tidligere forskning innen konkursprediksjon. Deretter ser vi på forskning som retter seg mot konkursprediksjon for SMB og bygg- og anleggsbransjen, samt bransjespesifikasjon. Til slutt gjennomgår vi enkelte variabelseleksjonsmetoder i tidligere forskning.

Det har blitt gjennomført flere studier innen konkursprediksjon siden 1960-tallet (Kumar & Ravi, 2007). Enkelte studier har fokusert på ulike metoder og variabler for predikering av konkurs. Det har blitt utviklet modeller som senere har blitt benyttet i praksis, og flere har sett nærmere på spesifikke metoder eller sammenlignet metoder. De første studiene innen litteraturen benyttet univariat diskriminantanalyse. Deretter ble multivariat diskriminantanalyse benyttet av flere, før logistisk regresjon ble tatt i bruk. Maskinlæringsteknikker som nevrale nettverk, støttevektormaskiner og beslutningstrær er også benyttet av flere. Kumar og Ravi (2007) ga i sin artikkel en gjennomgang av tidligere forskning på intelligente og statistiske metoder for konkursprediksjon fra perioden 1968-2005. Ifølge Kumar og Ravi (2007) er kunstige nevrale nettverk den maskinlæringsteknikken som er mest brukt innen konkursprediksjon og statistiske metoder blir ikke benyttet alene i like stor grad lenger.

Beaver (1966) presenterte en av de første betydningsfulle forskningsartiklene innen konkursprediksjon. Artikkelen omhandler bruken av økonomiske nøkkeltall for prediksjon av konkurs. I studien ble det benyttet børsnoterte selskaper for å analysere forholdet mellom ulike økonomiske nøkkeltall og sannsynligheten for konkurs, ved bruk av univariat diskriminantanalyse. Han så på 79 selskaper som hadde gått konkurs og 79 selskaper som ikke hadde gått konkurs, innen 38 ulike bransjer. Selskapene hadde mellom 0,6 millioner og 45 millioner USD i eiendeler, og gjennomsnittet lå på 6 millioner. Årsregnskapene ble hentet fra Moody's. Han så på 30 ulike

variabler, som ble delt inn i følgende kategorier: kontantstrømforhold, lønnsomhet, belåning, likviditet og aktivitet. Analysen for å velge variabler ble delt inn i tre deler: (1) sammenligning av gjennomsnittsverdier, (2) dikotom klassifiseringstest, og (3) analyse av sannsynlighetsforhold. Seks variabler innenfor de nevnte kategoriene ble valgt.

Z-score modellen av Altman (1968) er sentral innen konkursprediksjon og bygger blant annet på forskningen til Beaver (1966). Altman (1968) tok i bruk multivariat diskriminantanalyse, som ikke tidligere hadde blitt benyttet innen konkursprediksjon. Datagrunnlaget bestod av årsrapporter fra 66 børsnoterte produksjonsbedrifter i perioden 1946-1966. Z-score modellen viste god prediksjonsevne ved prediksjon opp til to år frem i tid. Altman (1968) hadde 22 potensielle variabler i fem ulike kategorier: likviditet, lønnsomhet, belåning, soliditet og aktivitet. De relevante variablene ble hentet fra tidligere litteratur. Ved en samlet vurdering av variablenes statistiske signifikans, korrelasjon og prediktiv treffsikkerhet, ble til slutt inkludert fem variabler.

Altman mfl. (1977) utviklet Zeta-modellen, som har bedre prediksjonsevne for større selskaper, og ved lengre tidshorisont. Modellen ble utviklet med data fra perioden 1969-1975, som bestod av 53 selskaper som hadde gått konkurs, og 58 selskaper som ikke hadde gått konkurs. Multivariat diskriminantanalyse ble tatt i bruk. De hadde i utgangspunktet 27 variabler å velge mellom, og benyttet statistiske tester for å velge de syv variablene som ga best resultater. Treffsikkerheten til modellen var signifikant bedre enn tidligere modeller. Zeta-modellen viste seg å være mer presis enn Z-score modellen ved prediksjon lengre frem i tid. Ved prediksjon 1 år og 5 år frem i tid, hadde Zeta-modellen en nøyaktighet på henholdsvis 95% og 70%. Z-score modellen hadde kun 30% nøyaktighet ved prediksjon av konkurs 5 år frem i tid.

Senere ble Z-score-modellen revidert av Altman (2000) til to nye versjoner: Z'-score modellen og Z"-score modellen. Variabelen $x_4 = Markedsverdi\ egenkapital/Sum\ gjeld$ ble endret i førstnevnte modell, der markedsverdi på egenkapital ble byttet ut med bokført verdi. Dette gjorde modellen bedre egnet for ikke-børsnoterte selskaper. I Z"-score modellen ble ikke variabel $x_5 = Salgsinntekter/Sum\ eiendeler$ inkludert. Denne ble fjernet da det kan være store forskjeller mellom finansieringen av eiendelene for ulike selskaper. Modellen ble dermed mindre bransjespesifikk.

Ohlson (1980) introduserte O-score modellen, som er basert på logistisk regresjon. Han benyttet data fra perioden 1970 til 1976. Studien er basert på observasjoner fra 105 selskaper som hadde gått konkurs, og 2058 selskaper som ikke hadde gått konkurs. Nøkkeltallene i modellen ble valgt med bakgrunn i tidligere litteratur.

I Norge er SEBRA-modellen sentral innen konkursprediksjon. Den er utviklet av Norges Bank som gjennom mange år har benyttet modellen til å analysere bankenes kredittrisiko i foretakssektoren, for å bidra til å sikre finansiell stabilitet i landet (Eklund mfl., 2001). Modellen ble utviklet i 2001 og inkluderte syv variabler som gir informasjon om selskapets inntjening, soliditet, likviditet, alder og størrelse. I 2007 ble det videreutviklet en basisversjon og en utvidet versjon av den opprinnelige modellen (Bernhardsen og Larsen, 2007). Den opprinnelige modellen ble utvidet med bakgrunn i ulike svakheter som ble avdekket. Svakhetene omfattet blant annet innføring av nye regnskapsregler og endringer i konkursregistreringen. Nyere og flere observasjoner bidro også til behovet for videreutvikling av modellen. I tillegg var det behov for å skille mellom risiko for tap og risiko for konkurs. Basismodellen inneholder de samme kategoriene som den opprinnelige SEBRA-modellen, men noen variabler ble endret. Variablene for inntjening og soliditet varierer mer over tid i basismodellen, og blir derfor estimert årlig i stedet for hele estimeringsperioden. Endringer i beskatning på utbytte gjorde at variabelen *utbetalt utbytte* ikke ble inkludert i SEBRA-basis og SEBRA-utvidet. Variabelen ga mer informasjon om skattetilpasninger, fremfor størrelse på selskapet som opprinnelig var hensikten. Den utvidede SEBRA-modellen inkluderer de samme nøkkeltallene som basismodellen, samt tre variabler som relateres til selskapenes størrelse. De tre variablene er *sum eiendeler i faste kroner*, *leverandørgjeld i prosent av totalkapital* og *skyldige offentlige avgifter i prosent av totalkapital*. Små foretak har historisk sett hatt flere konkurser sammenlignet med større foretak. Dette var noe av årsaken til inkludering av variablene i den utvidede modellen. Siden basismodellen ikke inkluderer størrelse-variablene er konkurssannsynligheten til større bedrifter høyere ved bruk av basismodellen, sammenlignet med den utvidede modellen. Den utvide modellen består av åtte variabler, som er presentert i tabell 3.3 i kapittel 3.

2.1 Små og mellomstore bedrifter

Flere studier har utviklet konkursprediksjonsmodeller for SMB. Edmister (1972) var en av de første som undersøkte nytten av å predikere konkurs for SMB. Han benyttet stegvis multivariat diskriminantanalyse til å velge variabler som ga best resultater. Det ble gjort en sammenligning av 19 variabler basert på tidligere forskning. Han kom frem til syv relevante variabler, og fant at modellen ikke presterte bedre av å inkludere flere variabler. Modellen ga en treffsikkerhet på 93%. Altman og Sabato (2007) utviklet også en konkursprediksjonsmodell for SMB. De tok utgangspunkt i arbeidet til Edmister (1972) og ønsket å finne de variablene som var viktigst for kredittverdigheten til selskapene. I tillegg ville de undersøke om modellen kunne senke kapitalkravene fra bankene og dermed også utlånsrenten på lån. Altman og Sabato (2007) utviklet

en prediksjonsmodell med et års tidshorisont basert på data fra omtrent 2000 selskaper i USA fra perioden 1994-2002. De benyttet logistisk regresjon, og tok utgangspunkt i variabler som var ansett som gode i tidligere forskning. Først valgte de ti variabler basert på nøyaktighetsmål og kom til slutt frem til fem variabler ved bruk av trinnvis regresjon. Modell for SMB var 30% mer nøyaktig sammenlignet med Z"-score modellen. De fant at logistisk regresjon klassifiserer konkurs bedre enn multivariat diskriminantanalyse ved bruk av de samme variablene. Til slutt påpekte de også at kapitalkravet kunne senkes med 0,5% ved bruk av en modell for SMB. Pelja og Wahlstrøm (2021) fant i sin studie at størrelse påvirker modellers evne til å predikere konkurs. Studien benyttet data med årsregnskaper fra ikke-børsnoterte selskaper fra perioden 2006-2016. De benyttet metodene logistisk regresjon og nevrale nettverk.

2.2 Bransjespesifikasjon

Chava og Jarrow (2004) bidrar med å vise viktigheten av å inkludere industrieffekter. De begrunnet viktigheten av industrieffekter med bakgrunn i ulike nivåer av konkurranse i de forskjellige bransjene, og derfor kan det være flere konkurser i noen bransjer sammenlignet med andre. I studien ble selskapene delt inn i fire grupper; (i) finans, forsikring og eiendom, (ii) transport, kommunikasjon og verktøy, (iii) produksjon og mineraler, og (iv) diverse industrier. De fant ved bruk av en kji-kvadrat-test at industrieffektene var statistisk signifikante. Alaka mfl. (2016) utviklet et teoretisk rammeverk med kritiske faktorer for konkursprediksjon innen bygg- og anleggsvirksomhet. Følgende tre tilnærminger ble benyttet for å komme frem til rammeverket: 1) identifisere eksisterende faktorer, 2) kartlegge hyppigheten av bruken av faktorene og nøyaktigheten ved modellene i tidligere studier og 3) spørreskjema besvart av ansatte innen bygg- og anleggsvirksomhet. De fant ved bruk av denne tilnærmingen hvilke kvantitative og kvalitative kategorier som var viktigst for konkursprediksjon innen bygg- og anleggsvirksomhet. De kvantitative kategoriene de fant var: lønnsomhet, likviditet, belåning, ledelseseffektivitet og kontantstrøm. Det ble konkludert med at de viktigste kvalitative kategoriene var: egenskaper ved ledelsen og eierne, intern strategi, ledelsesbeslutninger og makroøkonomiske forhold. Alaka mfl. (2016) påpekte følgende makroøkonomiske forhold som betydningsfulle: aktivitet innen næringen, antall tilgjengelige kontrakter i landet på et gitt tidspunkt, rentesatsen, svakheter i næringen og mulighet for nyetableringer. Hol (2007) påpekte også at makroøkonomiske faktorer kan være viktige innen konkursprediksjon. Hun analyserte hvordan konjunktursykluser, økonomiske nøkkeltall og makroøkonomiske variabler påvirker sannsynligheten for konkurs. Hun fant at de viktigste makroøkonomiske variablene var BNP-gapet, produksjonsindeks og pengemengden M1.

2.3 Variabelseleksjon

Vi ser videre på forskning med fokus på variabelseleksjon i konkursprediksjon. En sentral studie å trekke frem er Paraschiv mfl. (2021), som så på hvilke nøkkeltall og indikatorer som er relevante for å predikere konkurs i SMB. Ved bruk av data fra norske årsregnskap i perioden 2006-2017, sammenlignet de tre variabelseleksjonsmetoder. De benyttet 155 regnskapsvariabler basert på tidligere studier av blant annet Ohlson (1980), Zmijewski (1984), Altman og Sabato (2007), Kumar og Ravi (2007), Campbell mfl. (2008), og Härdle mfl. (2009), samt flere andre. Logistisk regresjon og dype nevrale nettverk ble benyttet til å predikere konkurs. Metoden som ga best prediksjonsevne ved valg av variabler, var LASSO. LASSO viste seg å gi stabile resultater over flere år, og krevde mindre datakraft, sammenlignet med andre metoder. Dype nevrale nettverk var bedre enn logistisk regresjon, men forskjellen var ikke statistisk signifikant. De undersøkte også hvordan prediksjonsmodellene påvirker kredittrisikoprising og beslutninger, som påvirker lønnsomheten til bankene. Det ble gjort ved å simulere et kredittmarked fra virkeligheten, basert på rammeverket til Blöchlinger og Leippold (2006). Variabelsettet valgt ut ved bruk av LASSO viste seg å være mest lønnsom for bankene, da de kunne gi ut lån til kundene som ikke går konkurs, samtidig som de styrer unna de kundene som går konkurs. Små forskjeller i målt prediksjonsevne kan gi store utslag i overskuddet til bankene. De fant at lønnsomheten til bankene kunne forbedres med 50% ved bruk av variablene som ga best resultater ved variabelseleksjonen, sammenlignet med variablene som ga nest best resultater. Det ble videre påpekt at bransjespesifikke variabelsett kan forbedre prediksjonsevnen ytterligere. Tian mfl. (2015) benyttet også LASSO til å velge variabler for prediksjon av konkurs. Ved bruk av LASSO undersøkte de viktigheten av variabler hentet fra tidligere litteratur. De benyttet data fra perioden 1980-2009. Resultatene viste at variablene valgt ved LASSO hadde best prediksjonsevne sammenlignet med andre variabelsett fra blant annet Campbell mfl. (2008).

Jabeur mfl. (2022) sammenlignet metodene stegvis logistisk regresjon, stegvis diskriminantanalyse, minste kvadraters diskriminantanalyse og XGBoost for variabelseleksjon ved prediksjon av konkurs. I analysen ble det benyttet data fra franske ikke-børsnoterte selskaper. XGBoost ga de mest nøyaktige prediksjonene. SHAP kan brukes til å tolke XGBoost, og dermed benyttes til å forstå hvilke variabler som er av størst betydning innen konkursprediksjon. Wahlstrøm (2023) viser i sin studie hvordan SHAP kan benyttes til å tolke maskinlæringsteknikker, ved å predikere konkurs for SMB med data fra perioden 2010-2020. I studien til Xiaomao mfl. (2019) ga også SHAP best resultater blant flere variabelseleksjonsmetoder for prediksjoner innen finans. Lin og Bai (2022) sammenlignet LASSO og XGBoost med SHAP som evalueringsmetoder, og fant at XGBoost ga best resultater, i sin studie av indikatorer for gjeldsfinansiering.

Kapittel 3

Data

I dette kapittelet ser vi på datagrunnlaget for oppgaven. Vi presenterer informasjon om selve datasettet, responsvariabelen og hvordan vi preprosesserer dataen. Videre presenterer vi de 160 variablene vi benytter i oppgaven. Resultatene i oppgaven skal sammenlignes med tre variabelsett fra eksisterende litteratur, og disse gjennomgår vi til slutt.

3.1 Datasettet

Datasettet danner grunnlaget for oppgaven, og inneholder alle ukonsoliderte årsregnskap for norske private og børsnoterte selskaper som har rapportert til norske myndigheter, i perioden 2006-2020 (Wahlstrøm, 2022). Brønnøysundregisteret har levert datasettets innhold. Det opprinnelige datasettet består 4 248 493 observasjoner og 287 variabler før preprosesseringen.

Responsvariabelen *konkurs*, er en dummy-variabel som tar verdien 0 eller 1. I oppgaven definerer vi selskapene som konkurs dersom årsregnskapet er det siste levert av et selskap, og selskapet er begjært konkurs i henhold til variabelen *konkursdato*. Et selskap kan begjæres konkurs om det ikke lenger er i stand til å innfri de økonomiske forpliktelsene, og er insolvent (Altinn, 2021). Et insolvent selskap vil ha langvarige betalingsproblemer og forpliktelser dekkes ikke av selskapets eiendeler. Hvis selskapet selv eller kreditor begjærer virksomheten konkurs, avgjør tingretten om konkurs åpnes. Gitt at konkurs åpnes, oppnevnes en bostyrer som skal sikre en rettferdig fordeling av verdiene til kreditorene, og avvikling av selskapet.

Det gjennomføres en preprosessering av datasettet for å hente ut relevant informasjon til oppgaven. Først fjerner vi alle selskapsformer utenom aksjeselskap. Dette er med bakgrunn i at en stor andel av SMB er aksjeselskap og våre funn vil være av større nytte om vi ser på en spesi-

fikk selskapsform. Paraschiv mfl. (2021) påpeker også nytten av å se på "AS", da en stor del av tidligere litteratur har fokus på børsnoterte selskaper. Alle selskaper med sum eiendeler under NOK 500 000 fjerner vi i likhet med Bernhardsen og Larsen (2007). EU sitt regelverk definerer SMB som bedrifter med antall ansatte under 250, en årsomsetning under 50 millioner euro eller totale eiendeler under 43 millioner euro (European Comission, 2003). I Norge er det vanlig å skille SMB fra store bedrifter på antall ansatte, hvor SMB har under 100 ansatte (Regjeringen, 1997). I hovedsak benytter også EU antall ansatte som definisjon på selskapers størrelse. Vi har ikke informasjon om antall ansatte for alle selskaper, og tar derfor utgangspunkt i EU sin definisjon og skiller på årsomsetning eller årsbalanse i euro. Denne definisjonen er benyttet i flere tidligere studier, blant annet av Paraschiv mfl. (2021) og Schalck og Yankol-Schalck (2021). Videre inkluderer vi kun selskaper med landskode "NO", altså norske selskaper med hovedkontor i Norge.

Vi utarbeider to datasett som skal benyttes i analysene. Ett der vi kun inkluderer bygg- og anleggsselskaper og ett der vi inkluderer selskaper innen alle næringer. I logistisk regresjon og XGBoost benytter vi datasettet som kun inkluderer bygg- og anleggsbransjen. Datasettet som inneholder alle næringer benytter vi kun i XGBoost-modellene. For datasettet som kun inneholder selskaper innen bygg- og anleggsbransjer, fjerner vi alle selskaper med næringskode ulik "F". Næringskode "F"står for bygg- og anleggsvirksomhet, og inkluderer oppføring av bygninger, anleggsvirksomhet og spesialisert bygge- og anleggsvirksomhet (Statistisk sentralbyrå, 2009). Oppføring av bygninger inkluderer utvikling av byggeprosjekt. Anleggsvirksomhet innebærer bygging av veier, jernbaner, vann- og kloakkanlegg, samt anlegg for elektrisitet og telekommunikasjon. Spesialisert bygge- og anleggsvirksomhet omhandler blant annet rivning og grunnarbeid, ferdiggjøring av bygninger og elektrisk installasjonsarbeid. For datasettet vi benytter til å se på flere bransjer, fjerner vi næringskodene for omsetning og drift av eiendom ("L"), finansierings- og forsikringsvirksomhet ("K"), vannforsyning, avløp og renovasjonsvirksomhet ("E"), elektrisitets-, gass- og varmtvannforsyning ("D"), offentlige virksomheter ("O") og holdingselskaper ("0"). Disse næringskodene er vanlig praksis å fjerne ved prediksjon av konkurs, og vi følger tidligere studier som Mansi mfl. (2012) og Paraschiv mfl. (2021). Selv om vi utelater noen bransjer, betegner vi selskapene i datasettet med flere bransjer, som *alle bransjer* videre i oppgaven. Tabell 3.1 viser en oversikt over antall observasjoner, konkurser, ikke-konkurser og konkurs i prosent, i datasettet ved preprosesseringen. Datasettet som inneholder alle bransjer, består av 2 824 140 observasjoner, hvor 27 501 er konkurs. Datasettet som kun inkluderer næringskode "F"består av 376 936 observasjoner, og 6041 konkurser.

Tabell 3.1: Oversikt over antall observasjoner og konkurser ved preprossessering av data for bygg- og anleggsbransjen og alle næringer.

Filtrering	Antall observasjoner	Antall ikke-konkurs	Antall konkurs	Antall konkurs i %
Før preprossesering	4 248 493	4 197 366	51 127	1,20%
Kun aksjeselskap	3 710 810	3 661 575	49 235	1,3%
Eiendeler < 500 000 NOK	2 839 739	2 812 209	27 530	0,97%
Omsetningen < 50 M/EURO eller eiendeler < 43 M/EURO	2 828 562	2 80 153	27 509	0,97%
Ekskludering av næringskode L, K, E, D, O og 0	2 824 140	2 796 639	27 501	0,97%
Kun næringskode F	376 936	370 895	6 041	1,60%

Tabell 3.2 viser en oversikt over antall observasjoner og konkurser per regnskapsår etter preprosessering av data. I tabell 3.2 ser vi at det er færre konkurser i 2020, sammenlignet med tidligere år. I 2020 brøt koronapandemien ut i Norge (Tjernshaugen mfl., 2023). Bedrifter som opplevde omsetningsfall grunnet pandemien fikk mulighet til å søke om kompensasjon fra staten (Altinn, 2021). Det kan være noe av årsaken til færre konkurser i 2020. Vi ser også færre konkurser i 2019, sammenlignet med tidligere år. Dette kan skyldes at noen selskaper begjæres konkurs to til tre år etter at selskapet leverer siste årsregnskap. Det kan dermed være at noen av selskapene som ikke leverte årsregnskap i 2018, 2019 eller 2020, vil få en konkursdato i perioden 2021-2023, og dermed ikke er klassifisert som konkurs enda.

Tabell 3.2: Oversikt over antall observasjoner og konkurser per år for alle bransjer og bygg- og anleggsbransjen, etter preprossesering av data.

	Alle bransjer		**Bygg- og anleggsbransjen**	
	Antall observasjoner	**Antall konkurser**	**Antall observasjoner**	**Antall konkurser**
2006	144 692	1 402	16 466	256
2007	157 178	2 205	18 676	472
2008	162 330	1 941	19 708	419
2009	163 782	1 783	20 081	368
2010	166 377	1 556	20693	360
2011	170 836	1 619	21 737	395
2012	177 464	1 688	23 189	400
2013	184 422	1 653	24 762	404
2014	190 994	1 600	26 030	413
2015	197 858	1 507	27 411	414
2016	204 929	1 725	28 859	466
2017	213 618	1 754	30 361	492
2018	221 520	2 018	31 835	577
2019	229 522	1 209	32 949	398
2020	238 618	573	34 179	207

Det er flere årsregnskaper som er klassifisert som ikke konkurs, enn konkurs. Vi har dermed

et ubalansert datasett som vi velger å ikke balansere. Enkelte studier påpeker at balansering av datasettet løser klassebalanseringsproblemet mellom konkurs og ikke-konkurs. Studier som Beaver (1966) og Altman (1968) har valgt å balansere datasettet. Zmijewski (1984) argumenterer for hvorfor datasett ikke skal balanseres. Dersom andelen mellom konkurs og ikke-konkurs balanseres, vil ikke forholdet nødvendigvis reflektere virkeligheten, som kan føre til forvrenging av konkursprediksjonsmodeller. Shumway (2001) argumenterer også for at den faktiske fordelingen mellom konkurs og ikke-konkurs skal benyttes i konkursprediksjon. Det er flere som støtter bruken av et ubalansert datasett, blant annet Ohlson (1980), Altman og Sabato (2007) og Paraschiv mfl. (2021). Vi velger derfor å ikke balansere datasettene vårt.

For å håndtere og minimere effekten av ekstremverdier er dataen transformert ved bruk av *winsorizing*. Det vil si at vi setter ekstremverdier lik en angitt persentil av dataen. Vi velger å *winsorize* de finansielle nøkkeltallene i variabelsettet på 160 variabler, på 1. og 99. persentilen. Dette er i tråd med tidligere forskning (Shumway, 2001; Chava og Jarrow, 2004;Paraschiv mfl., 2021.

3.2 Variabler

I oppgaven benytter vi LASSO til å velge variabler fra et utvalg på 160 variabler. En oversikt over variabelsettet finnes i vedlegg VI. 155 av disse er basert på regnskapsinformasjon og er de samme som benyttes i Paraschiv mfl. (2021). Av de 155 variablene, er 151 finansielle variabler, to er dummyvariabler, og de siste to er henholdsvis log (alder) på selskapet og log (sum eiendeler). Vi inkluderer også fem makroøkonomiske variabler: BNP, produksjonsindeks for bygge- og anleggsvirksomhet, samt styringsrenten (Statistisk sentralbyrå, 2022a, 2023a, 2023d). Hol (2007) påpekte BNP og produksjonsindeksen som viktige variabler innen konkursprediksjon. Makrovariablene inkluderes basert på regnskapsår. Siden styringsrenten kan endre seg flere ganger i året, har vi valgt å inkludere gjennomsnittlig styringsrente per år, styringsrenten ved slutten av året og endring av styringsrenten per år. Produksjonsindeks er en månedlig indeks, og vi benytter gjennomsnittlig produksjonsindeks per år.

3.3 Variabelsett til sammenligning

3.3.1 Variablene i SEBRA-modellen

Variablene i SEBRA-modellen er presentert i tabell 3.3. De beskriver selskapets inntjening, soliditet, likviditet, størrelse og alder.

Tabell 3.3: Variablene i SEBRA-modellen, samt tilhørende kategorier.

Variabeler	**Kategori**
Ordinært resultat før av- og nedskrivninger i prosent av total gjeld	Inntjening
Egenkapital i prosent av total gjeld	Soliditet
Innskutt egenkapital er mindre bokført egenkapital (dummyvariabel)	Soliditet
Likvider minus kortsiktig gjeld i prosent av omsetning	Likviditet
Log(antall år siden etablering) (dummyvariabel)	Alder
Log(sum eiendeler i faste kroner)	Størrelse
Leverandørgjeld i prosent av totalkapital	Likviditet
Skyldige offentlige avgifter i prosent av totalkapital	Likviditet

Inntjening er en av de sentrale kategoriene i SEBRA-modellen. Inntjeningen i en bedrift må i forhold til betalingsforpliktelsene være rimelig, for at likviditeten i selskapet skal være god (Eklund mfl., 2001). Dersom inntjeningen ikke er tilstrekkelig, kan det føre til at bedriften ikke får inn ny kapital fra investorer og banker. Innenfor denne kategorien er variabelen *ordinært resultat før av- og nedskrivninger etter skatt, i prosent av langsiktig gjeld*, inkludert i SEBRA-modellen. Et minimumskrav for inntjening er blant annet at avdrag og utbytte skal dekkes.

Likviditet gir informasjon om selskapers betalingsevne, og Eklund mfl. (2001) skriver om mangel på likviditet som en faktor for hvorfor bedrifter går konkurs. I SEBRA-modellen beskriver *skyldige offentlige avgifter som andel av totalkapital*, *leverandørgjeld som andel av totalkapital* og *likvider fratrukket kortsiktig gjeld, i prosent av driftsinntekter*, bedriftens likviditet. Sistnevnte er inkludert på bakgrunn av at forverret likviditet vises i form av nedsatt betalingsevne eller økning i den kortsiktige gjelden.

Soliditet gir informasjon om en bedrifts evne til å tåle tap, og det blir ofte målt ved å se på egenkapitalandelen (Eklund mfl., 2001). I kategorien soliditet inkluderes variablene *egenkapital som andel av totalkapital* og *bokført egenkapital < enn innskutt egenkapital (dummyvariabel)*. Eklund mfl. (2001) påpeker at høyere egenkapitalandel vil gjøre foretaket bedre rustet for tøffe perioder, samt gjøre det enklere å skaffe midler gjennom salg eller lån.

Kategorien størrelse er inkludert i den utvidede SEBRA-modellen ved variabelen *log (sum eiendeler i faste kroner)*. I følge Eklund mfl. (2001) går mindre selskaper oftere konkurs og det er derfor viktig å inkludere en variabel som sier noe om selskapets størrelse. Til slutt er variabelen *log (antall år siden etablering)* inkludert innen kategorien alder. Variabelen er inkludert på bakgrunn av at det er flere yngre selskaper som går konkurs (Eklund mfl., 2001).

3.3.2 Variablene i Altman og Sabato (2007)

Variabelsettet utviklet av Altman og Sabato (2007) er som nevnt utviklet spesifikt for SMB. Variablene er presentert i tabell 3.4, og gir informasjon om selskapets soliditet, lønnsomhet, belåning, likviditet og aktivitet.

Tabell 3.4: Variablene fra Altman og Sabato (2007), samt tilhørende kategorier.

Variabeler	**Kategori**
Opptjent egenkapital/ sum eiendeler	Soliditet
Driftsresultat/ sum eiendeler	Lønnsomhet
Kortsiktig gjeld/ bokført egenkapital	Belåning
Kontantbeholdning/ sum eiendeler	Likviditet
Driftsresultat/ rentekostnader	Aktivitet

Den første variabelen er *opptjent egenkapital / sum eiendeler*. Opptjent egenkapital er egenkapital som kommer av at det holdes igjen et eventuelt overskudd i selskapet og at det ikke tas ut som utbytte. *Opptjent egenkapital / sum eiendeler* gir viktig informasjon om selskapets soliditet. Variabelen er typisk lavere for yngre firma. Alderen til selskapet tas på denne måten med gjennom variabelen. Det er historisk sett flere yngre selskaper enn eldre som går konkurs (Altman & Sabato, 2007). Variabelen *driftsresultat / sum eiendeler* gir informasjon om selskapets lønnsomhet, og er derfor inkludert. De vurderte også andre variabler som *årsresultat/ sum eiendeler* og *opptjent egenkapital/sum eiendeler*, men valgte å inkludere *driftsresultat / sum eiendeler*. *Kortsiktig gjeld / bokført egenkapital* går under kategorien belåning, men kan også gi informasjon om selskapets likviditet. Selskapet kan få betalingsproblemer om kortsiktig gjeld er høy i forhold til egenkapitalen. De valgte denne variabelen fremfor *totale forpliktelser/ sum eiendeler* og *bokført egenkapital/ totale forpliktelser*. Variabelen *kontantbeholdning / sum eiendeler* gir informasjon om selskapets likviditet, og sier noe om andelen av et selskaps eiendeler som er i kontanter eller omsettelige verdipapirer. Variabelen *driftsresultat / rentekostnader* har Altman og Sabato (2007) klassifisert under kategorien aktivitet. Variabelen gir informasjon om hvor kapabel en bedrift er til å betale utestående rentekostnader på sine lån.

3.3.3 Variablene i Paraschiv mfl. (2021)

Variabelsettet utviklet av Paraschiv mfl. (2021) består av ti variabler og er presentert i tabell 3.5. Variablene er funnet ved bruk av LASSO fra et utvalg på 155 variabler, som vi også benytter for variabelseleksjon. Paraschiv mfl. (2021) benytter data for SMB.

Tabell 3.5: Variablene fra Paraschiv mfl. (2021), samt tilhørende kategorier.

Variabler	**Kategori**
Leverandørgjeld/ sum eiendeler	Likviditet
Dummy: 1 dersom totale forpliktelser overstiger totale eiendeler	Belåning
Kortsiktig gjeld- sum bankinnskudd, kontanter o.l./ sum eiendeler	Belåning
Årsresultat/ sum eiendeler	Lønnsomhet
Skyldige offentlige avgifter/ sum eiendeler	Likviditet
Rentekostnader/ sum eiendeler	Soliditet
Dummy: 1 dersom innbetalt egenkapital er mindre enn sum egenkapital	Soliditet
Log (alder i år)	Alder
Sum varer/ omløpsmidler	Likviditet
Sum bankinnskudd, kontanter ol./ omløpsmidler	Likviditet

Av de ti variablene som ble funnet i Paraschiv mfl. (2021), er fire av disse også i den utvidede SEBRA-modellen. Dette gjelder *leverandørgjeld / sum eiendeler*, *skyldig offentlige avgifter / sum eiendeler*, *dummy: sum innskudd egenkapital er mindre enn sum egenkapital* og *log (alder i år)*.

Variablene *(kortsiktig gjeld - sum bankinnskudd, kontanter og lignende) / sum eiendeler* og *dummy: 1 dersom sum gjeld er større enn sum eiendeler*, gir informasjon om selskapets belåning. Høye verdier av variablene øker sjansen for konkurs. Hvis *(kortsiktig gjeld - sum bankinnskudd, kontanter og lignende) / sum eiendeler* er positiv og høy, kan det indikere betalingsproblemer. Variabelen *årsresultat/ sum eiendeler* gir informasjon om selskapets lønnsomhet. Innen kategorien lønnsomhet benytter SEBRA-modellen variabelen *resultat før av- og nedskrivninger/ sum eiendeler* og Altman og Sabato (2007) benytter *driftsresultat/ sum eiendeler*. Variabelen *rentekostnader / sum eiendeler* gir informasjon om selskapets soliditet. Denne variabelen gir også informasjon om andelen gjeld selskapet har i forhold til selskapets totalkapital. *Sum varer / sum omløpsmidler* og *sum bankinnskudd, kontanter og lignende / sum omløpsmidler* går under kategorien likviditet.

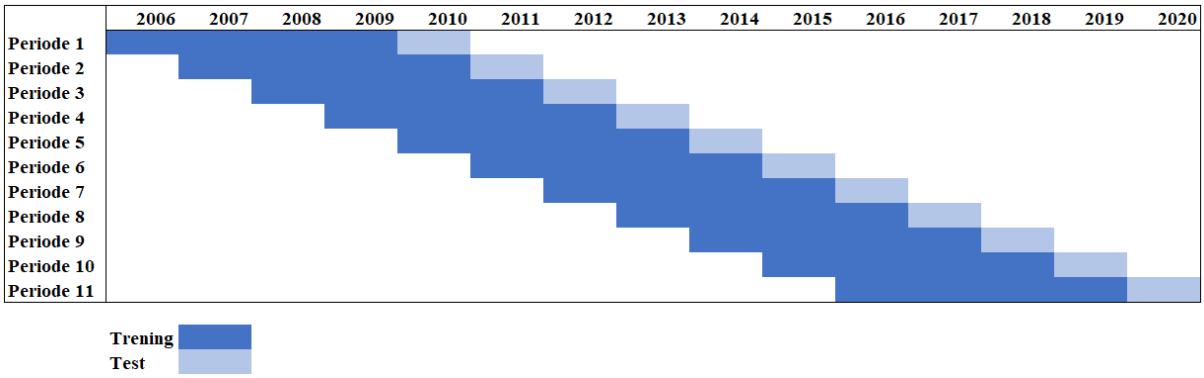
Kapittel 4

Metode

I dette kapittelet gjennomgår vi fremgangsmåten for å besvare problemstillingen. Vi ser først på inndelingen av datasettet i test- og treningssett. Videre presenterer vi logistisk regresjon og LASSO, før vi ser på XGBoost og SHAP. Deretter beskriver vi hvilke evalueringsmål vi benytter for å evaluere resultatene. Til slutt kommenterer vi oppgavens validitet, reliabilitet og repliserbarhet.

4.1 Trenings- og testsett

Dataen deles inn i test- og treningssett ved bruk av glidende vindu, også kalt *rolling window*. Vi følger flere selskaper over tid og ønsker å trene modellen på data som er relevant for dataen det testes på. Vi har valgt glidende vindu, da det gjør treningssettene like store, noe for eksempel *expanding window*, ikke gjør. Figur 4.1 viser fordelingen av test- og treningssett.



Figur 4.1: Oversikt over fordelingen av trenings- og testsett for periode 1 til 11.

En periode består av fire treningsår, og ett testår. Vi har totalt 11 perioder. Den første perioden er

trent på data fra årene 2006 til 2009, og det første teståret er 2010. Den siste perioden består av treningsperioden 2016 til 2019, og teståret 2020. Modellene blir altså trent på de fire foregående årene før teståret.

4.2 Logistisk regresjon

Logistisk regresjon er en statistisk læringsmetode som angir sannsynligheten for at en responsvariabel tilhører en binær gruppe (James mfl., 2013). Det er en klassifikasjonsmetode, som er basert på at sannsynlighetene kan benyttes til predikere et utfall mellom 0 eller 1. I vårt tilfelle predikerer vi om observasjonene tilhører gruppen 1 *konkurs*, eller 0 *ikke-konkurs*.

Vektoren for predikerte sannsynligheter for konkurs $\hat{y} = \{\hat{y_n}\}_{n=1,...,N} \in [0,1]^N$, i logistisk regresjon, er gitt ved (Paraschiv mfl., 2021):

$$\hat{\mathbf{y}} = \iota \oslash (\iota + exp(-\mathbf{X}\mathbf{w} - \iota\beta_0)), \tag{4.1}$$

hvor $\mathbf{X} = \{x_{(n,i)}\}_{n=1,...,N,i=1,...,I}$, er en matrise av verdier for variablene *i* og årsregnskap *n* (Paraschiv mfl., 2021). Vektoren av koeffisienter er gitt ved $\mathbf{w} = \{w_i\}_{i=1,...,I}$, hvor w_0 er konstantleddet. Her er ι en $N \times 1$-vektor bestående av enere, og $\oslash$ betegner Hadamard-divisjon. Koeffisientene $\mathbf{w}$ og w_0 estimeres ved å minimere den negative av logit-funksjonen $\ell(\mathbf{w}, w_0)$, som er gitt ved:

$$\ell(\mathbf{w}, w_0) = \sum_{n=1}^{N} [\mathbf{y} \odot (\mathbf{X}\mathbf{w} + \iota w_0) - log(\iota + exp(\mathbf{X}\mathbf{w} + \iota w_0))], \tag{4.2}$$

i ligning 4.2, er $y = \{y_n\}_{n=1,...,N} \in [0,1]^N$ vektoren for faktiske klassifiseringer av konkurs (1) og ikke-konkurs (0), og $\odot$ er betegnelsen på Hadamard-produktet. Regresjonskoeffisientene i logistiske regresjonsmodeller angir hvor mye log-oddsen endres, ved en endring i en gitt variabel, gitt at de andre variablene holdes konstant.

Logistisk regresjon er bedre egnet til å fange opp sannsynlighetsområdet enn lineære regresjonsmodeller, da sistnevnte kan gi meningsløse verdier mindre enn 0 og større enn 1 (James mfl., 2013). Metoden er mye brukt i tidligere studier, blant annet av Ohlson (1980), Shumway (2001) og Paraschiv mfl. (2021), da den er intuitiv å tolke. Vi trener logistiske regresjonsmodeller på flere årsregnskaper, og ikke kun de foregående årene. Modellene kan dermed omtales som

discrete hazard models (Shumway, 2001). Shumway (2001) påpeker at *discrete hazard models* tar hensyn til endringer i bedrifter over tid og justerer automatisk for risikoperioder. Alaka mfl. (2018) fant at kunstig nevrale nettverk er noe bedre enn logistisk regresjon ved konkursprediksjon i bygg- og anleggsvirksomhet. Dette kan tale for valg av andre metoder enn logistisk regresjon i denne oppgaven. Fra et annet synspunkt skal funnene i oppgaven sammenlignes med tre variabelsett som også benytter logistisk regresjon. Å benytte den samme metoden som de tre mener vi vil gi et bedre sammenligningsgrunnlag. Logistisk regresjon krever også mindre datakapasitet, enn for eksempel nevrale nettverk (Paraschiv mfl., 2021), og egner seg å benytte sammen med LASSO. Vi benytter derfor variabelsettene fra tabell 3.3, 3.4 og 3.5, samt variablene valgt ved LASSO, i logistiske regresjonsmodeller, for å se hvordan modellene med ulike variabler presterer.

4.3 LASSO

LASSO er en regulariseringsmetode introdusert av Tibshirani (1996). Regulariseringsmetoder kan benyttes til å forbedre tilpasningen til en modell ved å krympe estimatene til koeffisientene i modellen (James mfl., 2013). LASSO kan velge variabler som inkluderes i en modell ved at enkelte koeffisienter settes lik 0. Koeffisientene i LASSO, $\hat{\beta}_{\lambda}^{L}$, minimerer følgende, i treningssettet (Paraschiv mfl., 2021):

$$-\ell(\mathbf{w}, w_0) + \lambda||\mathbf{w}||_1, \tag{4.3}$$

hvor $\ell(\mathbf{w}, w_0)$ er lik logit-funksjonen gitt i funksjon 4.2, og $||\mathbf{w}||_1$ er lik summen av absoluttverdiene til w_i (Paraschiv mfl., 2021). Hyperparameteren λ, fungerer som et straffeledd, som krymper koeffisientene i modellen eller setter koeffisientene lik 0 (James mfl., 2013). Jo større λ er, desto strengere vurderer LASSO variablene. Vi benytter kryssvalidering med tre folder for å velge størrelsen på λ. Kryssvalideringsfeil beregnes for et sett med verdier av λ, og vi benytter AUC som mål i modellene, hvor det er ønskelig med så høy AUC som mulig. LASSO-modellen trenes med en valgt λ-verdi basert på kryssvalideringen. Ofte benyttes "ett-standardfeil-regelen " ved kryssvalidering (Hastie mfl., 2009), hvor det tillates en kryssvalideringsfeil innenfor ett standardfeil av minimum gjennomsnittlig kryssvalideringsfeil (Hastie mfl., 2021). For å begrense antall variabler i modellen velger vi enλ-verdi som tillater 1,5 standardfeil. Vi testet med både en standardfeil og 1,5 standardfeil, og ved førstnevnte fikk vi veldig mange variabler. For å øke tolkbarheten ved resultatene og få frem kun de viktigste variablene, valgte vi 1,5 standardfeil.

Vi benytter LASSO da den har gitt gode resultater ved variabelseleksjon innen konkursprediksjon. Tian mfl. (2015) og Paraschiv mfl. (2021) fant at variablene valgt ved LASSO ga best prediksjonsevne i sine studier. Det er også en metode som krever mindre datakapasitet enn andre variabelseleksjonsmetoder (Paraschiv mfl., 2021). James mfl. (2013) karakteriserer LASSO som en metode som scorer høy på tolkbarhet og lavt på fleksibilitet. Metoden er restriktiv ved estimering av koeffisientene, sammenlignet med andre metoder som for eksempel dyplæringsmodeller, ettersom LASSO setter noen av koeffisientene lik 0. På denne måten er LASSO mindre fleksibel enn andre metoder, men også enklere å tolke. Tian mfl. (2015) trekker også frem at LASSO kan gi stabile resultater ved små avvik i dataen, tar hensyn til multikollinearitet mellom variablene og ereffektiv ved seleksjon mellom mange variabler.

4.4 XGBoost

XGBoost er en skalerbar maskinlæringsteknikk utviklet av Chen og Guestrin (2016). Metoden er basert på beslutningstrær og *gradient boosting*, og kombinerer flere maskinlæringsalgoritmer. Vi har valgt å benytte XGBoost med bakgrunn i at metoden har blitt benyttet i flere tidligere studier og har gitt gode resultater ved prediksjon av konkurs (Jabeur mfl., 2022; Lin & Bai, 2022; Wahlstrøm, 2023). Å benytte flere metoder mener vi vil være med på å styrke funnene i oppgaven, ettersom vi kan få ulike resultater ved bruk av ulike metoder. XGBoost er utviklet for å kunne benyttes på en enkel måte og krever ikke like mye datakraft som for eksempel nevrale nettverk. XGBoost kan også fange opp ikke-lineære relasjoner mellom variabler (Wahlstrøm, 2023). Vi presenterer ikke alle formlene som er bakgrunnen for XGBoost i denne oppgaven, men beskriver noen av de essensielle delene ved modellen. En dypere beskrivelse av algoritmene og formlene finnes i Chen og Guestrin (2016).

Gradient boosting er en maskinlæringsteknikk eller algoritme, som sekvensielt bygger en modell basert på flere svake modeller, eller modeller som i seg selv har lav treffsikkerhet. Gradient boosting ble først introdusert av Friedman (2001). Boosting-aloritmer minimerer bias og varians i en modell. I XGBoost bygges beslutningstrær sekvensielt (Chen & Guestrin, 2016). For hvert tre som bygges reduseres feil fra forrige tre. XGBoost hindrer overtilpasning ved regularisering. Metoden kan håndtere manglende verdier i dataen, samt vektede data. En modell som består av flere tre-baserte modeller predikerer et utfall ved *K* additive funksjoner:

$$\hat{y_n} = \phi(x_i) = \sum_{k=1}^{K} g_j(x_n), f_k \in \mathcal{F}, \tag{4.4}$$

hvor $\mathcal{F} = \{f(x) = w_q(x)\}(q : \mathbb{R} \rightarrow T, q \in \mathbb{R})^T)$, angir formelen for klassifikasjons- og regresjonstrærne (Chen & Guestrin, 2016). *T* er antall grener på trærne, *q* er strukturen til hvert tre og w er vektene til grenene. f_k korresponderer til en trestruktur q. Følgende regulariserende mål minimeres for hver iterasjon *t*:

$$\mathcal{L}^{(t)} = \sum_{n=1}^{N} l(y_n, \hat{y_n}^{(t-1)} + f_t(x_i)) + \Omega(f_t), \tag{4.5}$$

modellen forbedres ved at f_t som forbedrer modellen mest, legges til. $\hat{y_n}^{(t-1)}$ er lik prediksjonen for hver observasjon *n* og iterasjon *t*. Ω er et straffeledd som gjør modellen mindre kompleks når den øker (Chen & Guestrin, 2016). l er en konveks tapsfunksjon som angir forskjellen mellom $\hat{y_i}$ og y_i.

Å sortere data i tre-baserte modeller kan være tidkrevende (Chen & Guestrin, 2016). Ved XGBoost lagres data i lagringsenheter, som kalles blokker. Dataen i blokkene beregnes kun en gang og kan hentes opp igjen. Dette er med på å optimalisere beregningene og gjør at beregningene krever mindre datakapasitet. XGBoost har en *cahche-awareness*- algoritme, som gjør metoden effektiv. *Cache*, eller et hurtigminne, er en midlertidig lagringsplass som gjør det mulig å hente data raskt. Kjøretiden reduseres, særlig ved data med mange rader, noe vi har i vårt datasett med mange selskaper. En annen egenskap ved XGBoost er beregning utenfor kjernen ved bruk av blokker, som Chen og Guestrin (2016) kaller *out-of-core computing*. Det muliggjør for skalerbar læring. Data lagres som nevnt i blokker som lagres på disken. Metoden benytter teknikker som blokk-komprimering og blokk-deling.

Optimalisering av hyperparameterne i XGBoost-modellen ble gjort ved bruk av *grid-search*. Vi angir et sett med verdier av hyperparametere modellene tester for å velge optimal. Tabell 4.1 viser en oversikt over de hyperparameterne vi optimaliserer. Vi velger disse på bakgrunn av Saraswat (2016) sin anbefaling, og Wahlstrøm (2023) sin studie. For hyperparameteren *n_estimator*, velger vi standardparameter som er lik 100. Dette er med bakgrunn i at vi testet med et grid på 25, 50 og 100 på flere modeller, og fant at modellen valgte 100 på alle. For å spare tid ved optimalisering av parameterne velger vi dette for alle. Hyperparameterne tunes kun for XGBoost-modellene utviklet for bygg- og anleggsselskaper, med bakgrunn i at det er en tidkrevende prosess og modellen for bygg- og anleggsbransjen er hovedfokuset i oppgaven.

Tabell 4.1: Optimaliserte hyperparametere, samt en beskrivelse av de.

Hyperparameter	Beskrivelse
Gamma	Bestemmer minimum tapsreduksjon ved regularisering.
Learning_rate	Forhindrer overtilpasning og gjør modellen mer kompleks ved å krympe vektene.
Subsample	Andel av treningssettet benyttet i trebygging. Verdi mellom 0 og 1.
Reg_lamda	Regularisering ved rigde regresjon.
Max_depth	Beslutningstrærnes maksimale dybde. Høye verdier gjør modellen mer kompleks.
N_estimator	Antall beslutningstrær i modellen.

Vi utvikler XGBoost-modeller basert på data med selskaper fra kun bygg- og anleggsbransjen, og med data fra selskaper innen alle næringer. Dette gjør vi for å se om vi ved bruk av SHAP finner ulike variabler som viktige. Vi sammenligner også hvordan modellene presterer. XGBoost-modellene med alle bransjer trenes på data med alle bransjer og testes med data for kun bygg- og anleggsbransjen. Dette er for å få innsikt i hvordan modeller som er utviklet for en spesifikk bransje presterer i forhold til modeller utviklet for alle bransjer. Variablene vi finner ved bruk av LASSO sammenligner vi med tidligere utviklede variabelsett. Siden vi ikke har informasjon om viktige variabler for alle bransjer, basert på XGBoost og SHAP, mener vi det er interessant å finne dette, for å sammenligne med modeller for bygg- og anleggsbransjen.

4.5 SHAP

XGBoost kan gi nøyaktige prediksjoner, men ulempen er at resultatene kan være vanskelig å tolke. En utfordring ved maskinlæringsteknikker er det såkalte *black box problem*, som omhandler vanskeligheter ved å se hvordan modellen kommer frem til resultatene (Tjoa & Guan, 2020). Det kan skape utfordringer knyttet til å stole på resultatene, samt tolkning av modellene. SHAP er et rammeverk som kan benyttes til å tolke resultater av maskinlæringsteknikker (Lundberg & Lee, 2017). Vi benytter SHAP til å tolke XGBoost-modeller og få innsikt i hvilke variabler som er viktige innen konkursprediksjon, basert på SHAP-verdier.

SHAP er en metode som bygger på spillteorien introdusert av Shapley (1953). Spillteorien handler om individers bidrag til en gruppe (Dubey, 1975). SHAP ble introdusert av Lundberg og Lee (2017), med formål om å lettere kunne tolke modeller med høy treffsikkerhet, og er en additiv attribusjonsmetode for funksjonsegenskaper. Ifølge Lundberg og Lee (2017) har attribusjonsmetoder en forklaringsmodell *g(z')*, som er gitt ved følgende formel:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z_i', \tag{4.6}$$

hvor $z' \in \{0,1\}^I$, *M* er antall forenklede variabler og $\phi_i \in \mathbb{R}$ (Lundberg & Lee, 2017). Hvis variabel *i* er observert, er z' lik 1, og 0 ellers.

Det er flere additive attribusjonsmetoder som kan beskrives ved modell *g(z')* (Lundberg & Lee, 2017). Eksempler er *LIME* (*local interpretable model-agnostic explenations*), *deepLIFT*, lagvis relevansproagering (*layer-wise relevance propagation*) og SHAP-verdi estimering. SHAP-verdier tilfredsstiller egenskaper som de andre metodene ikke gjør og kan benyttes til å gi informasjon om viktigheten av variabler i en modell. SHAP-verdiene beregnes ved formelen under:

$$\phi_i = \sum_{S \subseteq I\{i\}} \frac{|S|!(|I| - |S| - 1)!}{|I|!} [f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)], \tag{4.7}$$

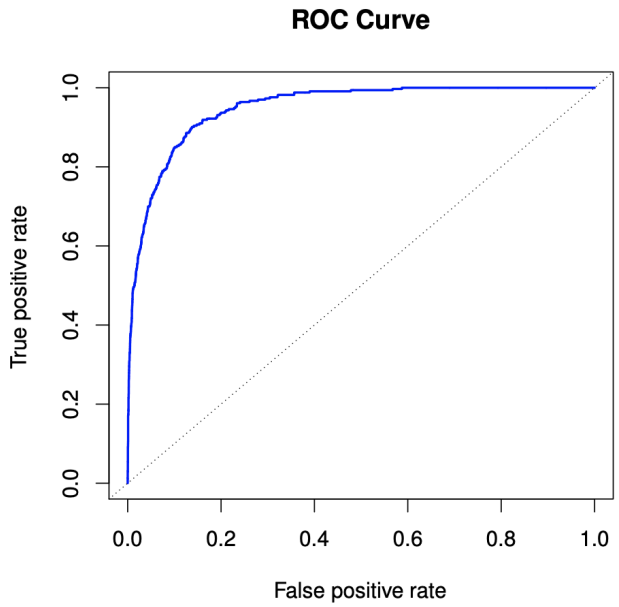
hvor ϕ_i er lik det forventede marginale bidraget til gruppe *S*, for hver variabel *i* (Lundberg & Lee, 2017). Variabel *i* er inkludert ved trening av modell $f_{S\cup\{i\}}$. I modell f_S er ikke variabel *i* inkludert. Prediksjonene for modellene sammenlignes ved følgende: $[f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)]$. Her er x_s gitt ved verdien til variabelen. Effekten av en variabel avhenger av de andre variablene. SHAP-verdiene er dermed lik et vektet gjennomsnitt av de mulige forskjellene gitt ved funksjonen over.

For alle XGBoost-modellene beregnes SHAP-verdien til alle variablene og sorteres etter høyest SHAP-verdi. Vi ser kun på de 10 viktigste variablene for hver treningsperiode, da vi mener disse gir tilstrekkelig informasjon om de viktigste variablene for hver periode.

4.6 Evaluering

For å evaluere modellene benytter vi målene AUC og Brier Score. Evalueringsmålene beregnes for både test- og treningssett. Vi ønsker å benytte mer en ett evalueringsmål for å styrke funnene i oppgaven, ettersom ulike mål gir informasjon om ulike forhold ved modellene.

The reciever operator characteristics curve (ROC) viser falsk positiv rate mot sann positiv rate for alle terskelverdier mellom 0 og 1 (James mfl., 2013). Area under the reciever operator characteristics curve (AUC) er arealet under ROC-kurven. Figur 4.2 viser et eksempel på en ROC-kurve.



Figur 4.2: Eksempel på hvordan en ROC-kurve kan se ut, med sann positiv rate på y-aksen og falsk positiv rate på x-aksen (James mfl., 2013).

Tabell 4.2 viser mulige utfall ved klassifisering av konkurs. Sann positiv rate, eller sensitivitet, er andelen korrekt klassifisert som konkurs av antall konkurser (James mfl., 2013), og er gitt i formel 4.8. Spesifisitet gir informasjon om hvor mange av de som ikke går konkurs som klassifiseres riktig. Falsk positiv rate er gitt ved 1−sensitivitet, og er angitt i formel 4.9. Falsk positiv rate er altså lik antall feilklassifisert som konkurs, som andel av alle selskapene som ikke gikk konkurs. En falsk positiv klassifikasjon gir type I-feil og en falsk negativ klassifikasjon gir type II feil. For eksempel for banker som gir ut lån vil type II-feil gjerne være mer kostbart enn type I-feil.

Tabell 4.2: Klassifikasjonsmatrise som viser en oversikt over mulige utfall ved prediksjon av konkurs, sett opp mot faktiske utfall.

	Faktisk: konkurs	**Faktisk: ikke konkurs**
Prediksjon: konkurs	Sann positiv	Falsk positiv (Type I feil)
Prediksjon: ikke konkurs	Falsk negativ (Type II feil)	Sann negativ

$$Sann\ positiv\ rate\ (sensitivitet) = \frac{Antall\ korrekt\ klassifisert\ som\ konkurs}{Antall\ konkurser\ i\ utvalget} \tag{4.8}$$

$$Falsk\ positiv\ rate\ (1 - spesifisitet) = \frac{Antall\ feilklassifisert\ som\ konkurs}{Antall\ ikke - konkurser\ i\ utvalget} \tag{4.9}$$

Det er ønskelig med en ROC-kurve opp mot øverste venstre hjørne i figur 4.2, slik at falsk positiv rate er lav samtidig som sann positiv rate er høy (James mfl., 2013). AUC er et tall mellom 0 og 1, hvor 1 er det beste. En AUC-verdi lik 0,5 kan tolkes som at en modell er like god som en tilfeldig modell. Basert på Hosmer Jr mfl. (2013) anses en AUC-verdi mellom 0,5 og 0,7 som dårlig, og en verdi mellom 0,7 og 0,8 som akseptabel. En utmerket modell har en AUC-verdi mellom 0,8 og 0,9, og en AUC-verdi over 0,9 tilsier at modellen er fremragende.

AUC er benyttet som evalueringsmål, og er brukt i flere tidligere studier innen konkursprediksjon (Paraschiv mfl., 2021). AUC er et evalueringsmål som egner seg å benytte ved ubalanserte datasett (Brzezinski & Stefanowski, 2017). Det ikke nødvendig å sette en terskel for hvor høy sannsynligheten må være for at en observasjon klassifiseres som konkurs ved bruk av AUC. Vi har som nevnt et ubalansert datasett og ved beregning av mål som treffsikkerhet, sensitivitet og spesifisitet vil resultatene påvirkes av hvilken terskel vi setter for prediksjonene. Siden det er skjev fordeling mellom konkurser og ikke-konkurser i datasettet, kan modellene ha god treffsikkerhet, men gir lite informasjon om hvor stor andel av konkursene som klassifiseres riktig og galt.

Vi benytter også Brier (1950) Score til å måle nøyaktigheten ved prediksjonsmodellen. Brier Score er et mål som egner seg ved binære klassifikasjonsproblemer (Rufibach, 2010), og er gitt ved formelen:

$$Brier\ score(y, \hat{y}) = \frac{1}{N}\sum_{n=1}^{N}(\hat{y}_n - y_n)^2, \tag{4.10}$$

Brier Score er lik gjennomsnittlig kvadrert feil mellom predikert sannsynlighet for klasse 1, konkurs, og observert verdi på tvers av alle observasjonene *N* (Rufibach, 2010). Evalueringsmålet gir altså informasjon om treffsikkerheten til de predikerte sannsynlighetene, og dermed hvor godt modellen er kalibrert. Vi får en score mellom 0 og 1, hvor 0 er det beste. Jo nærmere 0, desto mer nøyaktig er modellen. Brier Score benyttes også som evalueringsmål i Tian mfl. (2015) og Paraschiv mfl. (2021).

4.7 Validitet og reliabilitet

For å vurdere kvaliteten på oppgaven ser vi videre på validiteten og reliabiliteten til resultatene som fremkommer. Validitet handler om hvorvidt det er mulig å trekke gyldige slutninger om det som skal undersøkes, basert på resultatene man finner (Dahlum, 2021). Det handler altså om i

hvilken grad vi kan besvare vår problemstilling med våre resultater. Det skilles mellom ytre og indre validitet.

Ytre validitet handler om hvorvidt resultatene kan generaliseres til et større utvalg enn dataen som blir benyttet. Vi baserer resultatene våre på alle norske årsregnskap som rapporterer til norske myndigheter. Vi skiller ut på blant annet størrelse, selskapsform og næringskode. Resultatene våre kan til en viss grad generaliseres til SMB i bygg- og anleggsbransjer. De minste bedriftene målt i eiendeler skilles ut. Vi baserer våre funn på historiske data og forholdene kan endre seg over tid. Det er ikke sikkert funnene gjelder for gruppen vi ser på flere år frem i tid.

Indre validitet går ut på om vi faktisk måler det vi ønsker å måle og innebærer å ha så lav bias som mulig (Dahlum, 2021). Vi deler datasettet inn i test- og treningsett for å unngå overtilpasning. LASSO og XGBoost er også metoder som på ulike måter skal hindre overtilpasning. Vi benytter også evalueringsmål for å vurdere prestasjonen til modellene. Dette er forhold som styrker oppgavens indre validitet. Indre validitet handler også om i hvilken grad indikatorene vi velger faktisk måler det vi ønsker å måle. Her er for eksempel definisjonen av konkurs relevant. Vi definerer årsregnskaper som konkurs dersom de har gått konkurs i henhold til variabel *konkursdato* og årsregnskapet er det siste levert av et selskapet (Wahlstrøm, 2022). Bernhardsen og Larsen (2007) og Paraschiv mfl. (2021) benytter samme definisjon av konkurs, men de velger å ikke predikere konkurs for de siste årene de har tilgang til, for å ta hensyn til at enkelte selskaper begjæres konkurs to til tre år etter at selskapet leverer siste årsregnskap. Vi valgte å inkludere alle år, for å benytte all tilgjengelig data. Vi ønsket også å undersøke om resultatene for 2020 ville skille seg ut fra andre år, ettersom at 2020 er et år hvor selskapene ble påvirket av makroøkonomiske forhold og det var færre konkurser. Ettersom vi presenterer resultater for elleve perioder, vi inkluderer de siste årene, kun påvirke resultatene for de to til tre siste periodene.

Reliabilitet handler om hvorvidt det er stabilitet i målingene som gjennomføres og om resultatene er pålitelige (Svartdal, 2020). Det er alltid fare for målefeil når tester gjennomføres. Systematiske målefeil svekker reliabiliteten. For å vurdere reliabiliteten for de logistiske regresjonsmodellene ser vi blant annet på signifikansnivå. Signifikansnivået gir informasjon om hvorvidt sammenhengen mellom variablene LASSO velger og responsvariabelen konkurs, er reell. Vi benytter flere metoder for å predikere konkurs, samt to ulike evalueringsmål for å styrke reliabiliteten. Evalueringsmålene vi benytter har betydning for hvordan modellene blir vurdert. Evalueringsmål kan gi ulike indikatorer på hvor gode modellene er, ettersom de kan måle ulike forhold ved modellene. Det at vi utarbeider flere modeller for flere perioder, er det også med på å styrke påliteligheten til resultatene.

4.7.1 Repliserbarhet

Dataen vi benytter har vi fått tilgang til for å besvare denne oppgaven, og kan ikke deles. Hvordan vi behandler dataen er beskrevet kapittel 3, Data. Vi benytter programmeringsspråkene Python og R til å behandle dataen og gjennomføre analyser. R benytter vi til å gjennomføre LASSO og logistisk regresjon, ellers benytter vi Python. I Python bruker vi blant annet pakkene *scikit-learn* (scikit-learn developers, 2007-2018), XGBoost (XGBoost developers, 2022) og SHAP (Lundberg, Scott, 2018). Pakken *scikit-learn* bruker vi blant annet til å finne AUC, og pakkene *XGBoost* og *SHAP* benytter vi henholdsvis til å utvikle XGBoost-modeller og beregne SHAP-verdier, samt fremstilling av resultater. I R benytter vi blant annet *glm* (Marschner & Donoghoe, 2018), *glmnet* (Hastie mfl., 2021), *pROC* (Robin mfl., 2011) og *ggplot2* (Wickham mfl., n.d). Vi benytter *glm* og *glmnet* til logistisk regresjon og LASSO. *pROC* og *ggplot2* benyttes henholdsvis til å beregne AUC og fremstille plot for LASSO. Koden vi benytter i oppgaven kan oppgis ved forespørsel.

Kapittel 5

Resultater

I dette kapittelet presenterer vi resultatene for oppgaven. Først gjennomgår vi resultatene fra LASSO og logistisk regresjon, før vi presenterer evalueringsmålene for de logistiske regresjonsmodellene. Deretter presenterer vi variablene med høyest SHAP-verdi i XGBoost-modellene, både for bygg- og anleggsbransjen og alle næringer, før vi legger frem evalueringsmålene for XGBoost-modellene.

5.1 LASSO og logistisk regresjon

5.1.1 Variabelseleksjon - bygg- og anleggsbransjen

Vedlegg I viser standardiserte koeffisienter og AUC, for de ulike verdiene av $log(\lambda)$. Figurene gir informasjon om hvordan AUC endrer seg for de ulike $log(\lambda)$-verdiene. Den striplede linjen lengst til venstre viser $log(\lambda)$-verdi for minimum gjennomsnittlig kryssvalideringsfeil. Den midterste striplede linjen viser den største verdien av $log(\lambda)$, hvor kryssvalideringsfeilen er innenfor ett standardfeil. Den tykkeste linjen viser valgt $log(\lambda)$-verdi, med 1,5 standardfeil. Siden vi benytter AUC som mål i modellen, er det ønskelig at AUC er så høy som mulig. Vi ser altså at AUC er noe lavere for 1,5 standardfeil sammenlignet med minimum gjennomsnittlig kryssvalideringsfeil og ett standardfeil. Valgt verdi av $log(\lambda)$ er med bakgrunn i at færre variabler gir økt tolkbarhet. Vi ser videre i figurene når $log(\lambda)$ øker, minker antall variabler LASSO velger. Variablene LASSO velger og som dermed er inkludert i logistiske regresjonsmodeller er presentert i figurene.

Tabell 5.1 og 5.2 viser en oversikt over hvilke variabler valgt ved bruk av LASSO med tilhørende regresjonskoeffisienter og signifikansnivå, samt kontantsleddet i modellene. Resultatene

for periode 1-6 er presentert i tabell 5.1 og resultatene for periode 7-11 er presentert i tabell 5.2. Det er til sammen 34 ulike variabler i tabellene og variablene øverst i tabellene er valgt flest ganger. Det er noe variasjon i hvilke variabler som velges og antall variabler fra periode til periode. Syv variabler velges av LASSO hver periode. Disse variablene er: *dummy: totale forpliktelser er større enn sum eiendeler*, *kortsiktig gjeld (KG) - sum bankinnskudd, kontanter ol./ sum eiendeler*, *årsresultat/ sum eiendeler*, *leverandørgjeld/ sum eiendeler*, *skyldige offentlige avgifter/ sum eiendeler*, *sum bankinnskudd, kontanter ol./ omløpsmidler*, og *log(alder i år)*. Variablene gir henholdsvis informasjon om selskapets belåning, lønnsomhet, likviditet og alder. Ingen av de makroøkonomiske variablene er inkludert. Regresjonskoeffisientene gir informasjon om endring i log-oddsen, ved en endring i den gitte variabelen, gitt at de andre variablene holdes konstant. De kan sier dermed noe om hvordan variablene påvirker oddsen eller sjansen for konkurs. Fortegnene til koeffisientene gir nyttig informasjon om hvilken retning variablene påvirker sjansen for konkurs. Flesteparten av variablene valgt av LASSO er signifikante på 0,1%-nivå.

Vi ser nærmere på de syv variablene valgt av LASSO alle periodene. Variabelen *dummy: totale forpliktelser er større enn sum eiendeler*, er signifikant på 0,1% nivå alle periodene, og har et positivt fortegn. Det vil si at hvis totale forpliktelser er større enn sum eiendeler, øker sjansen for konkurs. Denne variabelen tilhører kategorien belåning. Signifikansnivået for *årsresultat / sum eiendeler* varierer gjennom perioden, men er for de fleste perioder signifikant på 0,1%-nivå. Koeffisienten har negativt fortegn, med unntak av i periode 4, hvor den heller ikke er signifikant. Et negativt fortegn indikerer lavere sjanse for konkurs ved økning i verdien av variabelen. *Årsresultat / sum eiendeler* gir informasjon om selskapets lønnsomhet. *Leverandørgjeld / sum eiendeler* er også valgt alle periodene. Variabelen er signifikant på 0,1%-nivå i syv perioder. I periode 5 og 8 er den signifikant på 1%-nivå og i periode 4 og 7, er den ikke signifikant. Fortegnet er positivt alle periodene, og en økning i variabelen øker sjansen for konkurs. Denne variabelen tilhører kategorien likviditet.

Videre ser vi på variabelen *skyldig offentlige avgifter / sum eiendeler*, som er signifikant på 0,1%-nivå alle periodene. Regresjonskoeffisienten har et positivt fortegn og er høyest av koeffisientene til de syv variablene valgt alle perioder. Høye verdier av variabelen indikerer altså høyere sjanse for konkurs. Variabelen gir informasjon om selskapets likviditet. Variabelen *sum bankinnskudd, kontanter ol. / omløpsmidler* er også signifikant på 0,1%-nivå i alle periodene, og har et negativt fortegn. Koeffisientene er relativt høy i absoluttverdi i flere perioder. En økning i variabelen minker dermed sjansen for konkurs i stor grad, i forhold til de andre variablene med negativt fortegn. *Sum bankinnskudd, kontanter ol. / omløpsmidler* inngår i kategorien likviditet.

Tabell 5.1: Oversikt over regresjonkoeffisientene og signifikansnivået for variabler valgt ved LASSO i periode 1-6. Signifikansnivå er merket på følgende måte: ***: 0,1%, **: 1%, *: 5% og .: 10%. Konstantleddet for modellene er inkludert nederst i tabellen.

Variabelnavn	**Periode 1**	**Periode 2**	**Periode 3**	**Periode 4**	**Periode 5**	**Periode 6**
Dummy: totale forpliktelser > sum eiendeler	0,461***	0,324***	0,317***	0,617***	0,615***	1,089***
Årsresultat /sum eiendeler	-0,171	-1,000***	-0,996***	0,112	-0,874***	-0,765***
Leverandørgjeld / sum eiendeler	1,753***	1,856***	1,908***	0,456	0,466*	1,654***
Skyldig offentlige avgifter / sum eiendeler	4,767***	4,630***	4,347***	4,081***	4,16***	4,378***
Sum bankinnskudd, kontanter ol. / omløpsmidler	-1,583***	-1,581***	-1,382***	-1,200***	-1,302***	-1,501***
KG - sum bankinnskudd, kontanter ol. / sum eiendeler	0,084	0,224*	0,41***	0,990***	0,97***	0,506***
Log (alder i år)	-0,432***	-0,438***	-0,429***	-0,416***	-0,415***	-0,412***
Omsetning / sysselsatt kapital	0,009***			0,003	0,007***	0,003*
Omsetning / sum egenkapital				0,005**		0,007***
Omsetning / sum varer				2,183e-08	4,517e-08*	
Rentekostnader / sum eiendeler				4,744***	3,807***	
Dummy: 1 hvis innbetalt egenkapital < sum egenkapital	-0,397***	-0,424***	-0,475***	-0,356***	-0,234***	-0,152
Varekostnad / sum varer				3,393e-08	2,641e-08	7,991e-08***
Leverandørgjeld / kortsiktig gjeld				1,058***	0,995***	
Immaterielle eiendeler / sum eiendeler						
Sum bankinnskudd, kontanter ol./ sum eiendeler						
Sum egenkapital / sum eiendeler	-0,773***	-0,599***	-0,358***	0,066		
Likvide eiendeler / sum eiendeler						
Omsetning / omløpsmidler						
Omsetning / anleggsmidler				1,145e-08***		
Resultat før skatt / omsetning						
Log (sum eiendeler)						
Sum forpliktelser / sum eiendeler				0,034	0,074	
Egenkapital / sum eiendeler				-0,776***		
Omsetning / aksjonærenes egenkapital						
Driftsresultat / sum eiendeler				-1,163*		
Driftsresultat / omsetning						
Sum varer / arbeidskapital						
Årsresultat / innbetalt egenkapital						
Driftsresultat / sum inntekter						
Utbytte / årsresultat						
Resultat før skatt / sum eiendeler	-0,921					
Leverandørgjeld / sum varer						
Lønnskostnadsandel						
Konstantledd	-3,583	-3,520	-3,745	-4,631	-4,746	-4,503

Variabelen *(kortsiktig gjeld - sum bankinnskudd, kontanter ol.) / sum eiendeler* er signifikant på 0,1%-nivå alle perioder, bortsett fra i periode 2, hvor den er signifikant på 5%-nivå og periode 1, hvor den ikke er signifikant. Variabelen har positivt fortegn og gir informasjon om selskapets belåning. En økning i denne variabelen indikerer større sjanse for konkurs. Den siste variabelen som er valgt hver periode er *log (alder i år)*. Variabelen er signifikant på 0,1%-nivå for hele perioden, og fortegnet er negativt. Det negative fortegnet indikerer at desto yngre selskapet er, desto større sjanse er det for at det går konkurs. Denne variabelen inngår i kategorien alder.

Variabelen *omsetning / sysselsatt kapital* er valgt ved bruk av LASSO hele ni perioder. Signifikansnivået varierer noe. Variabelen er ikke signifikant i alle periodene og koeffisientene har positive fortegn. En økning i *omsetning / sysselsatt kapital* tyder på økt sjanse for konkurs. *Omsetning / sum egenkapital* og *omsetning / sum varer* er valgt i syv av periodene. Variabelen *omsetning / sum egenkapital* har et positivt fortegn og er signifikant på 0,1%-nivå de fleste peri-

Tabell 5.2: Oversikt over regresjonskoeffisientene og signifikansnivået for variabler valgt ved LASSO i periode 7-11. Signifikansnivå er merket på følgende måte: ***: 0,1%, **: 1%, *: 5% og .: 10%. Konstantleddet for modellene er inkludert nederst i tabellen.

Variabelnavn	**Periode 7**	**Periode 8**	**Periode 9**	**Periode 10**	**Periode 11**
Dummy: totale forpliktelser > sum eiendeler	1,053***	1,081***	0,88***	0,944***	0,937***
Årsresultat /sum eiendeler	-0,33*	-0,379**	-0,312***	-0,322**	-0,316***
Leverandørgjeld / sum eiendeler	0,414	0,544*	0,852***	0,776***	1,818***
Skyldig offentlige avgifter / sum eiendeler	4,139***	4,492***	4,576***	4,630***	4,515***
Sum bankinnskudd, kontanter ol. / omløpsmidler	-0,749***	-0,764***	-0,675***	-0,814***	-1,318***
KG - sum bankinnskudd, kontanter ol. / sum eiendeler	0,708***	0,704***	0,675***	0,738***	0,48***
Log (alder i år)	-0,45***	-0,439***	-0,435***	-0,453***	-0,397***
Omsetning / sysselsatt kapital	0,003	0,002	8,028e-04	0,002	0,003*
Omsetning / sum egenkapital	0,006***	0,006***	0,006***	0,006***	0,006***
Omsetning / sum varer	-1,091e-08	1,619e-08***	7,449e-09**	6,322e-09***	3,933e-09*
Rentekostnader / sum eiendeler	4,976***	3,486***	2,491**	2,833***	3,117***
Dummy: 1 hvis innbetalt egenkapital < sum egenkapital					
Varekostnad / sum varer	2,128e-08			7,734e-09	1,203r-08*
Leverandørgjeld / omløpsmidler	1,077***	1,081***	0,872***	0,878***	
Immaterielle eiendeler / sum eiendeler	1,187*	1,166*	3,052***	1,586**	2,827***
Sum bankinnskudd, kontanter ol./ sum eiendeler	-1,182***	-1,105***	-1,381***	-0,837***	
Sum egenkapital / sum eiendeler					
Likvide eiendeler / sum eiendeler	-	-	-	-	
Omsetning / omløpsmidler	-0,004	-0,009	-0,012		
Omsetning / anleggsmidler			1,646e-08***		1,1993-08***
Resultat før skatt / omsetning	-4,088e-07	-5,569e-07	-3,992e-07***		
Log (sum eiendeler)		-0,085**	-0,12***		
Sum forpliktelser / sum eiendeler					
Egenkapital / sum eiendeler	-0,551**				
Omsetning / aksjonærenes egenkapital	4,836e-04		6,221e-04**		
Driftsresultat / sum eiendeler					
Driftsresultat / omsetning	-1,705e-07				
Sum varer / arbeidskapital	-0.052**				
Årsresultat / innbetalt egenkapital	0,009*				
Driftsresultat / sum inntekter				-6,869e-04***	
Utbytte / årsresultat					-0,525***
Resultat før skatt / sum eiendeler					
Leverandørgjeld / sum varer	1,912e-07*				
Lønnskostnadsandel	0,576**				
Konstantledd	-4,826	-3,440	-3,027	-4,746	-4,848

odene. Variabelen *omsetning / sum varer* har et positivt fortegn for alle perioder, med unntak av periode 7. Signifikansnivået varierer noe. De tre nevnte variablene kan sies å gi informasjon om selskapets soliditet og aktivitet.

Andre variabler som også er valgt flere perioder av LASSO er: *rentekostnader / sum eiendeler*, *dummy: 1 hvis innbetalt egenkapital < sum egenkapital*, *varekostnad / sum varer* og *leverandørgjeld / kortsiktig gjeld*. Variablene *rentekostnader / sum eiendeler*, *varekostnad / sum varer* og *leverandørgjeld / kortsiktig gjeld*, har alle positivt fortegn for periodene de er valgt. De gir informasjon om henholdsvis selskapets belåning og likviditet. Signifikansnivået varierer noe for variablene *rentekostnader / sum eiendeler* og *varekostnad / sum varer*. Variabelen *leverandørgjeld / kortsiktig gjeld* er signifikant på 0,1%-nivå for alle periodene den er inkludert. *Dummy: 1 hvis innbetalt egenkapital < sum egenkapital*, har et negativt fortegn og er signifikant på 0,1%-nivå fem av de seks periodene den er valgt. Variabelen gir informasjon om selskapets soliditet.

5.1.2 Evalueringsmål

Tabell 5.3 viser en oversikt over AUC og Brier score i test- og treningssett for de logistiske regresjonsmodellene hvor vi har inkludert variablene valgt ved LASSO, og variablene fra Paraschiv mfl. (2021). Vi sammenligner prestasjonen på disse modellene, da variablene fra Paraschiv mfl. (2021) også er valgt ved bruk av LASSO. Vi benytter også samme datasett som Paraschiv mfl. (2021), og tar flere like valg ved preprosesseringen av dataen. Den største forskjellen er at vi kun inkluderer næringskode "F". Å sammenligne modeller med dette variabelsett, kan dermed gi oss informasjon om nytten av å velge variabler og utvikle modeller for en spesifikk bransje.

AUC på treningssettene til de logistiske regresjonsmodellene med variablene valgt med LASSO varierer mellom 0,871 og 0,888. På testsettene varierer AUC mellom 0,852 og 0,894. AUC-verdiene på testsettene tilsier at dette er utmerkede modeller (Hosmer Jr mfl., 2013). For de logistiske regresjonsmodellene med variablene til Paraschiv mfl. (2021), ligger AUC på treningssettene mellom 0,868 og 0,881, og for testsettene er verdiene mellom 0,856 og 0,911. Dette er også utmerkede modeller. I de fleste periodene er AUC-verdien på treningssettene høyere enn på testsettene, med unntak av i enkelte perioder. I periode 11 er AUC lavest for den logistiske regresjonsmodellen med variablene valgt av LASSO, og AUC på den logistiske regresjonsmodellen med variablene fra Paraschiv mfl. (2021) er høyest for periode 11. Brier Score er gjennomgående lav for alle periodene i modellene med variabler valgt ved LASSO, og variablene fra Paraschiv mfl. (2021), med unntak av i periode 11 for modellene med LASSO-variablene. Brier Score for modellene med variabler fra Paraschiv mfl. (2021) er lavere i periode 11, sammenlignet med andre perioder.

Tabell 5.3: Evalueringsmål for logistiske regresjonsmodeller med variablene valgt av LASSO, og de logistiske regresjonsmodellene der variablene av Paraschiv mfl. (2021) er implementert.

	AUC treningssett		**Brier score treningssett**		**AUC testsett**		**Brier score testsett**	
	LASSO	**Paraschiv mfl. (2021)**	**LASSO**	**Paraschiv mfl. (2021)**	**LASSO**	**Paraschiv mfl. (2021)**	**LASSO**	**Paraschiv mfl. (2021)**
Periode 1	0,875	0,873	0,018	0,018	0,857	0,856	0,016	0,016
Periode 2	0,875	0,872	0,018	0,018	0,866	0,864	0,016	0,016
Periode 3	0,871	0,871	0,017	0,017	0,865	0,866	0,016	0,016
Periode 4	0,872	0,868	0,016	0,016	0,872	0,876	0,015	0,015
Periode 5	0,872	0,868	0,016	0,016	0,894	0,891	0,014	0,014
Periode 6	0,877	0,876	0,015	0,015	0,876	0,881	0,014	0,014
Periode 7	0,888	0,879	0,014	0,015	0,875	0,874	0,015	0,015
Periode 8	0,888	0,881	0,014	0,014	0,882	0,879	0,015	0,014
Periode 9	0,888	0,880	0,014	0,014	0,879	0,875	0,016	0,016
Periode 10	0,884	0,877	0,015	0,015	0,872	0,872	0,011	0,011
Periode 11	0,882	0,877	0,014	0,014	0,852	0,911	0,031	0,006

Tabell 5.4 gir en oversikt over de gjennomsnittlige evalueringsmålene i alle periodene for logistiske regresjonsmodeller når variablene fra Altman og Sabato (2007), SEBRA-modellen (Eklund mfl., 2001), Paraschiv mfl. (2021) og variablene valgt ved LASSO inkluderes. Eva-

lueringsmålene for alle periodene for modellene ved bruk av variablene fra Altman og Sabato (2007) og variablene i SEBRA-modellen, er presentert i vedlegg II. Evalueringsmålene i alle periodene for variablene valgt ved LASSO og variablene fra Paraschiv mfl. (2021), er presentert i tabell 5.3.

Gjennomsnittlig AUC i treningssettene er høyest for variablene valgt ved LASSO. På testsettene er gjennomsnittlig AUC høyest for modellene ved bruk av variablene til Paraschiv mfl. (2021). Modellene ved bruk av de valgte LASSO-variablene har nest høyest gjennomsnittlig AUC på testsettene. Gjennomsnittlig Brier Score på treningssettene er lavest for modellene ved bruk av variablene fra SEBRA-modellen, Paraschiv mfl. (2021) og LASSO. Modellene med variablene fra Paraschiv mfl. (2021) har lavest gjennomsnittlig Brier Score, og modellene med LASSO-variablene har høyest gjennomsnittlig Brier Score på testsettet.

Tabell 5.4: Gjennomsnittsverdier for evalueringsmål for logistiske regresjonsmodeller der variablene av Altman og Sabato (2007), variablene i SEBRA-modellen, Paraschiv mfl. (2021) og variablene valgt ved LASSO er inkludert.

	Altman og Sabato (2007)	Variabler SEBRA-modellen	Paraschiv mfl. (2021)	Variabler valgt ved LASSO
AUC trening	0,784	0,862	0,875	0,879
Brier score trening	0,017	0,016	0,016	0,016
AUC test	0,778	0,863	0,877	0,872
Brier score test	0,015	0,015	0,014	0,016

5.2 XGBoost og SHAP

5.2.1 Variabelseleksjon - bygg- og anleggsbransjen

Tabell 5.5 viser en oversikt over de ti variablene med høyest SHAP-verdi i XGBoost-modellene, og som dermed har størst betydning for prediksjonene. Modellene er utviklet med data fra selskaper innen bygg- og anleggsbransjen, og ved bruk av optimaliserte hyperparametere, som ligger i vedlegg III. Figur 5.5 viser skalerte SHAP-verdier. De opprinnelige gjennomsnittlige SHAP-verdiene for hver variabel i hver periode, blir skalert fra 0 til 100. Variabelen av størst betydning for hver periode har i tabell 5.5 verdien 100. Resterende variabler for den gitte perioden har en verdi mellom 0 og 100, som andel av SHAP-verdien til den viktigste variabelen. De variablene som ikke er blant de 10 med høyest SHAP-verdi for hver periode har fått verdien 0. Den høyeste SHAP-verdien er representert ved den mørkeste blåfargen, og lavere SHAP-verdier har en svakere blåfarge. Variablene med høyest SHAP-verdier tilsammen for de elleve periodene, står øverst i tabellen. Det er tilsammen 18 variabler i tabellen.

I periode 3 og 5, har *log (alder i år)* høyest SHAP-verdi og tilhører kategorien alder. *Skyldige*

offentlige avgifter / sum eiendeler er den viktigste variabelen i periode 1 og 2, basert på SHAP-verdier. Den variabelen som har høyst SHAP-verdi i de resterende periodene, er *(kortsiktig gjeld - sum bankinnskudd, kontanter ol.) / sum eiendeler*. *Skyldige offentlige avgifter / sum eiendeler* og *(kortsiktig gjeld - sum bankinnskudd, kontanter ol.) / sum eiendeler*, sier noe om selskapets likviditet og belåning. Variabelen *sum egenkapital/omsetning* kan si noe om selskapets soliditet og har høye SHAP-verdier hver periode. *Likvide midler / kortsiktig gjeld*, har også høye SHAP-verdier hver periode og tilhører kategorien likviditet. *Leverandørgjeld / omsetning* er representert ved SHAP-verdier alle perioder sett bort i fra periode 8, og gir også informasjon om selskapets likviditet.

Det er flere variabler med som er blant de ti med høyest SHAP-verdi, men som ikke er representert i like mange perioder som variablene nevnt ovenfor. Dette inkluderer blant annet *opptjent egenkapital / kortsiktig gjeld*, *(egenkapital (EK)- immaterielle eiendeler (IE)) / (sum eiendeler - IE - sum bankinnskudd, kontanter ol.)* og *leverandørgjeld / sum eiendeler*. De resterende variablene i tabellen er de ni variablene med høyest SHAP-verdi fire perioder eller færre. Ingen av de makroøkonomiske variablene er en av de variablene med høyest SHAP-verdi.

Tabell 5.5: De ti variablene med høyest gjennomsnittlig SHAP-verdi i XGBoost-modellene i periode 1-11. Modellene er basert på data fra bygg- og anleggsbransjen.

Variabelnavn/ Periode	1	2	3	4	5	6	7	8	9	10	11
(KG- sum bankinnskudd, kontanter ol.) / Sum eiendeler	56,7	72,4	78,6	100,0	92,6	100,0	100,0	100,0	100,0	100,0	100,0
Skyldige offentlige avgifter/ Sum eiendeler	100,0	100,0	92,9	92,9	85,2	86,7	97,1	88,2	82,5	69,0	82,9
Log (Alder i år)	96,7	96,6	100,0	85,7	100,0	90,0	88,6	76,5	67,5	57,1	70,7
Sum egenkapital/ Omsetning	66,7	75,9	64,3	78,6	88,9	93,3	88,6	70,6	62,5	59,5	53,7
Likvide midler/ Kortsiktig gjeld	40,0	62,1	71,4	42,9	44,4	43,3	88,6	29,4	27,5	26,2	39,0
Leverandørgjeld/ Omsetning	40,0	58,6	39,3	46,4	44,4	36,7	31,4	0,0	22,5	28,6	26,8
Opptjent egenkapital/ Kortsiktig gjeld	83,3	79,3	64,3	46,4	29,6	0,0	0,0	29,4	0,0	0,0	0,0
(EK- IE) / (Sum eiendeler- IE- Sum bankinnskudd, kontanter ol.)	36,7	58,6	35,7	57,1	63,0	26,7	25,7	0,0	0,0	0,0	0,0
Leverandørgjeld/ Sum eiendeler	33,3	24,1	39,3	28,6	0,0	30,0	0,0	41,2	25,0	0,0	29,3
Renteinntekter/ Rentekostnader	0,0	0,0	0,0	0,0	0,0	0,0	0,0	47,1	37,5	38,1	41,5
Utbytte/ Omsetning	53,3	41,4	0,0	35,7	0,0	0,0	0,0	0,0	0,0	0,0	26,8
Omsetning/ Sum bankinnskudd, kontanter ol.	0,0	0,0	0,0	0,0	0,0	43,3	28,6	0,0	27,5	14,3	0,0
Log (Sum eiendeler)	0,0	0,0	0,0	0,0	0,0	0,0	0,0	20,6	32,5	26,2	29,3
Renteinntekter/ Sum eiendeler	0,0	0,0	0,0	0,0	29,6	30,0	28,6	0,0	0,0	0,0	0,0
Leverandørgjeld/ Omløpsmidler	0,0	0,0	0,0	0,0	0,0	0,0	28,6	0,0	0,0	19,0	0,0
Opptjent egenkapital/ Omsetning	0,0	0,0	35,7	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Sum bankinnskudd, kontanter ol./ Omsetning	0,0	0,0	0,0	0,0	0,0	0,0	0,0	26,5	0,0	0,0	0,0
Sum bankinnskudd, kontanter ol/ Omløpsmidler	0,0	0,0	0,0	0,0	25,9	0,0	0,0	0,0	0,0	0,0	0,0

I vedlegg IV er *bee swarm*-plot for XGBoost-modellene utviklet for bygg- og anleggsbransjen vedlagt. Bee swarm-plottene gir informasjon om hvordan de ti variablene med høyest SHAP-verdi påvirker responsvariabelen *konkurs*. Den horisontale aksen representerer SHAP-verdien (Lundberg, 2018a). Hver observasjon i datasettet angis som et punkt i plottet med tilhørende

SHAP-verdi på den horisontale aksen, og ved høy tetthet står punktene vertikalt. Fargene gir informasjon om de opprinnelige verdiene til variablene for de ulike observasjonene, hvor rød farge gir informasjon om en høy verdi og blå gir informasjon om en lav verdi. I vedlegg IV ser vi at formen og fargen i bee swarm-plottene er ganske like for hver enkelt variabel fra periode til periode. Det er naturligvis noe variasjon i formen, da variabelverdien for hver observasjon varierer.

Vi ser nærmere på bee swarm-plot for periode 11, angitt i vedlegg IV. Variablene og SHAP-verdiene samsvarer med de skalerte SHAP-verdiene i tabell 5.5 for periode 11. Variabelen *kortsiktig gjeld- sum bankinnskudd, kontanter ol. / sum eiendeler*, har basert på SHAP-verdiene, størst påvirkning på responsvariabelen. Variabelen er rød ved positive SHAP-verdier, som indikerer at høyere verdier av *kortsiktig gjeld- sum bankinnskudd, kontanter ol. / sum eiendeler*, indikerer større sjanse for konkurs. Variabelen med nest størst påvirkning er *skyldige offentlige avgifter/ sum eiendeler*. Høye verdier av variabelen indikerer også større sjanse for konkurs. Videre har *log (alder i år)* høy SHAP-verdi og høyere alder indikerer lavere sjanse for konkurs. *Sum egenkapital/ omsetning* har noen røde prikker for negative SHAP-verdier. En økning i variabelen har dermed sammenheng med lavere sjanse for konkurs. I hvilken retning de resterende variablene påvirker sjansen for konkurs, er som forventet.

Vedlegg V viser *dependence scatter plot* av to valgte variabler for periode 11. *Dependence scatter plot* viser hvordan en enkelt variabel påvirker modellenes prediksjoner, ved at hver prikk angir en prediksjon (Lundberg, 2018b). Y-aksen viser SHAP-verdien for variabelen og x-aksen viser variabelen sin faktiske verdi. Variabelen *skyldige offentlige avgifter / sum eiendeler* er trukket frem med bakgrunn i at den er valgt av LASSO og har høye SHAP-verdier i alle periodene, og *likvide midler / kortsiktig gjeld* er trukket frem da den har høye SHAP-verdier og ikke valgt av LASSO. Vi ønsker å se om det er forskjeller mellom variablene som velges av LASSO og som har høye SHAP-verdier. Scatter plottet for *skyldige offentlige avgifter / sum eiendeler* viser en lineær sammenheng mellom variabelen og modellens prediksjoner. Vi ser ikke det samme lineære forholdet mellom *likvide midler / kortsiktig gjeld* og prediksjonene.

5.2.2 Variabelseleksjon - alle bransjer

Tabell 5.6 viser SHAP-verdier for variablene i XGBoost-modellene som er utviklet basert på selskaper innen alle bransjer. Modellene er også tilpasset ved bruk av hyperparameterne i vedlegg III. Det er til sammen 18 variabler i tabellen og de som står øverst har høyest SHAP-verdi i flest perioder.

Fire variabler har høyest SHAP-verdi i en eller flere perioder. Variabelen *log (alder i år)* står øverst i tabellen og har høyest SHAP-verdi i periode 5, 6 og 10. *Leverandørgjeld / sum eiendeler* har høyest SHAP-verdi de fire første periodene, og tilhører kategorien belåning. I periodene 7 og 11 har *skyldige offentlige avgifter / sum eiendeler* høyest SHAP-verdi, og den tilhører kategorien likviditet. Variabelen *(kortsiktig gjeld- sum bankinnskudd, kontanter ol.) / sum eiendeler*, har høyest SHAP-verdi i periode 8, 9 og 10. De fire nevnte variablene står øverst i tabellen og har følgelig høye SHAP-verdier gjennom hele perioden. I flere perioder har også variablene *log (sum eiendeler)* og *sum bankinnskudd, kontanter ol. / kortsiktig gjeld*, høye SHAP-verdier. Andre viktige variabler basert på SHAP-verdier er blant annet *sum egenkapital/ omsetning*, *leverandørgjeld / omsetning*, *utbytte/ omsetning*, *(egenkapital (EK) - immaterielle eiendeler (IE)) / (sum eiendeler - IE - sum bankinnskudd, kontanter ol.)* og *opptjent egenkapital / omsetning*. Resterende variabler i tabellen har høye SHAP-verdier i fire perioder eller færre.

Tabell 5.6: De ti variablene med høyest gjennomsnittlig SHAP-verdi i XGBoost-modellene i periode 1-11. Modellene er basert på data for alle bransjer.

Variabelnavn/ Periode	1	2	3	4	5	6	7	8	9	10	11
Log (Alder i år)	91,3	79,2	73,1	79,2	100,0	100,0	78,1	82,8	75,8	100,0	96,6
Leverandørgjeld/ Sum eiendeler	100,0	100,0	100,0	100,0	87,5	95,7	56,3	86,2	69,7	76,9	69,0
Skyldige offentlige avgifter/ Sum eiendeler	69,6	66,7	65,4	83,3	76,2	91,3	100,0	86,2	60,6	84,6	100,0
(KG - sum bankinnskudd, kontanter ol.) / Sum eiendeler	43,5	0,0	0,0	50,0	81,0	87,0	96,9	100,0	100,0	100,0	79,3
Log (Sum eiendeler)	43,5	37,5	50,0	45,8	52,4	52,2	59,4	55,2	57,6	69,2	75,9
Sum bankinnskudd, kontanter ol./ Kortsiktig gjeld	52,2	70,8	80,8	83,3	81,0	47,8	0,0	31,0	30,3	46,2	51,7
Sum egenkapital/ Omsetning	0,0	0,0	0,0	62,5	57,1	78,3	37,5	62,1	51,5	65,4	69,0
Leverandørgjeld/ Omsetning	60,9	54,2	53,8	0,0	0,0	0,0	40,6	31,0	33,3	42,3	55,2
Utbytte/ Omsetning	60,9	62,5	46,2	50,0	47,6	39,1	0,0	0,0	0,0	0,0	44,8
(EK- IE)/ (Sum eiendeler- IE- Sum bankinnskudd, kontanter ol.)	0,0	37,5	46,2	54,2	76,2	56,5	43,8	0,0	0,0	0,0	0,0
Opptjent egenkapital/ Kortsiktig gjeld	52,2	66,7	65,4	0,0	0,0	0,0	0,0	0,0	27,3	46,2	0,0
Likvide midler/ Kortsiktig gjeld	0,0	41,7	0,0	50,0	52,4	0,0	34,4	0,0	0,0	0,0	0,0
Sum bankinnskudd kontanter ol. / Omsetning	0,0	0,0	0,0	0,0	0,0	0,0	0,0	37,9	0,0	65,4	48,3
Opptjent egenkapital/ Omsetning	0,0	0,0	0,0	0,0	81,0	0,0	0,0	0,0	0,0	0,0	0,0
Renteinntekt / Rentekostnader	0,0	0,0	0,0	0,0	0,0	0,0	0,0	37,9	27,3	0,0	0,0
Årsresultat/ Sum eiendeler	0,0	0,0	46,2	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
(Driftskostnader - lønnskostnader) / Sum eiendeler	43,5	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Renteinntekter/ Sum eiendeler	0,0	0,0	0,0	0,0	0,0	39,1	0,0	0,0	0,0	0,0	0,0

5.2.3 Evalueringsmål

Tabell 5.7 viser oversikt over AUC og Brier Score for XGBoost-modellene for bygg- og anleggsbransjen og alle næringer. XGBoost-modellene som er utviklet og trent med data for selskaper innen alle næringer, er testet med data for bygg- og anleggsselskaper. Dette er for å kunne si noe om hvorvidt modellene utviklet for bygg- og anleggselskaper presterer bedre ved prediksjon av konkurs for bygg- og anleggselskaper, og dermed si noe om nytten av å utvikle modeller for spesifikke bransjer.

AUC på treningssettene for bygg- og anleggsbransjen varierer mellom 0,915 og 0,995, og for testsettene varierer AUC mellom 0,876 og 0,931. XGBoost-modellene for alle næringene har en AUC-verdi på treningssettene mellom 0,903 og 0,925. For testsettene er AUC mellom 0,872 og 0,928. XGBoost-modellene for bygg- og anleggsbransjen kan sies å være utmerket i de fleste periodene og fremragende i periode 5, 6 og 11, basert på AUC. XGBoost-modellene for alle næringer er også fremragende i periodene 5, 6, 7, 9 og 11, og utmerket i de resterende periodene. I de fleste periodene er AUC på testsettene noe høyere for XGBoost-modellene for bygg- og anleggsbransjen, sammenlignet med modellene for alle bransjer. Alle modellene har en Brier Score under 0,2 på test- og treningssettene. Periode 11 skiller seg ut både for bygg- og anleggsbransjen og for alle næringer, ved at AUC på treningssettene og testsettene er høyere, og Brier Score er lavere, sammenlignet med de andre periodene.

Tabell 5.7: Evalueringsmål for XGBoost-modellene for bygg- og anleggsbransjen (BA) og for alle næringer (Alle) i periode 1-11.

	AUC treningssett		**Brier score treningssett**		**AUC testsett**		**Brier score testsett**	
Periode	**BA**	**Alle**	**BA**	**Alle**	**BA**	**Alle**	**BA**	**Alle**
Periode 1	0,923	0,903	0,017	0,012	0,876	0,876	0,016	0,016
Periode 2	0,918	0,904	0,017	0,012	0,889	0,885	0,016	0,016
Periode 3	0,917	0,907	0,017	0,011	0,887	0,886	0,016	0,015
Periode 4	0,915	0,909	0,015	0,010	0,897	0,888	0,014	0,015
Periode 5	0,915	0,909	0,015	0,010	0,914	0,912	0,014	0,014
Periode 6	0,922	0,910	0,014	0,009	0,905	0,901	0,014	0,014
Periode 7	0,995	0,922	0,013	0,009	0,895	0,903	0,014	0,015
Periode 8	0,927	0,913	0,013	0,009	0,896	0,899	0,014	0,014
Periode 9	0,926	0,913	0,013	0,008	0,898	0,901	0,016	0,016
Periode 10	0,926	0,914	0,014	0,009	0,896	0,893	0,011	0,011
Periode 11	0,940	0,925	0,013	0,008	0,931	0,930	0,006	0,006

Kapittel 6

Diskusjon

I dette kapittelet diskuterer vi resultatene presentert i kapittel 5. Først ser vi på variablene vi anser som viktige ved bruk av LASSO og SHAP. Deretter sammenligner vi variablene vi anser som viktige, med de tre eksisterende variabelsettene fra Altman og Sabato (2007), variablene i SEBRA-modellen og Paraschiv mfl. (2021). Videre sammenligner vi variablene med høye SHAP-verdier for bygg- og anleggsbransjen, med alle bransjer. Til slutt diskuterer vi modellenes evalueringsmål.

6.1 Variabelseleksjon

Det er variasjon i hvilke variabler som velges av LASSO og hvilke variabler som har høye SHAP-verdier. Samtidig er det også flere likheter. Variablene *log (alder i år)*, *skyldige offentlige avgifter/sum eiendeler* og *(kortsiktig gjeld- sum bankinnskudd, kontanter ol.)/sum eiendeler*, er valgt alle periodene av LASSO og har høyeste SHAP-verdier alle perioder. Disse variablene inngår henholdsvis i kategoriene alder, likviditet og belåning. Variabelen *leverandørgjeld/sum eiendeler* har høy SHAP-verdi i åtte av periodene og er valgt ved LASSO alle perioder. Variabelen gir også informasjon om selskapets likviditet. Vi anser dermed disse fire variablene som viktige for prediksjon av konkurs i bygg- og anleggsbransjen, basert på både LASSO og SHAP.

Log (alder i år) har basert på SHAP-verdiene størst betydning de første periodene. De logistiske regresjonskoeffisientene varierer mellom -0,41 og -0,45. Fortegnet på koeffisientene og fargen på *bee swarm*-plottet er i samsvar med tidligere litteratur, som påpeker at yngre selskaper oftere går konkurs (Altman, 1968; Bernhardsen & Larsen, 2007). Dette gir mening da det kan være krevende å etablere seg i markedet. I bygg- og anleggsbransjen krever oppstart av et firma gjerne investeringer i anleggsmidler. Det kan være krevende å få tilgang på nok kapital og kunder i en

oppstartsperiode. Bransjen består av mange ulike typer selskaper. For eksempel trenger ikke rådgivere og arkitekter nødvendigvis å investere i anleggsmidler, men å etablere seg i markedet kan likevel være krevende. *Log (alder i år)* er også inkludert i variabelsettet i SEBRA-modellen og variabelsettet i Paraschiv mfl. (2021). SEBRA-modellen og variabelsettet til Paraschiv mfl. (2021) er ikke utviklet spesifikt for bygg- og anleggsbransjen. Dette kan tyde på at *log (alder i år)* er en sentral variabel innen konkursprediksjon uavhengig av bransje.

Ved bruk av LASSO finner vi at *skyldige offentlige avgifter/ sum eiendeler* er signifikant, og har en positiv og relativt høy regresjonskoeffisient alle perioder. Koeffisienten varierer mellom 4,08 og 4,77. Denne variabelen er blant de 10 viktigste variablene i XGBoost og har høy SHAP-verdi alle perioder. Høye og positive regresjonskoeffisienter samsvarer med fargene på bee swarm-plottene, der høyere verdier av variabelen indikerer høyere sjanse for konkurs. Skyldige offentlige avgifter er en del av selskapets kortsiktige gjeld, og innebærer blant annet arbeidsgiveravgift og merverdiavgift. Skyldige offentlige avgifter påvirkes av lønnskostnadene, samt kostnader til kjøp av varer og tjenester. Det kan tenkes at variabelen er viktig da sum eiendeler burde være høy i forhold til skyldige offentlige avgifter, for at selskapet skal ha god likviditet. Denne variabelen er også en del av variablene i SEBRA-modellen og variablene fra Paraschiv mfl. (2021). Dette peker i retning av at *skyldige offentlige avgifter / sum eiendeler* er viktig i både bygg- og anleggsbransjen og flere næringer.

I periodene 6-11, samt periode 4, har variabelen *(kortsiktig gjeld- sum bankinnskudd, kontanter ol.)/ sum eiendeler*, høyest SHAP-verdi av alle variablene. Regresjonskoeffisienten varierer noe, men er for de fleste perioder mellom 0,5-0,9. Positive fortegnet er som forventet, og samsvarer med fargen på bee swarm-plottene. Hvis kortsiktig gjeld er større enn sum bankinnskudd, kontanter ol., er variabelen positiv, og dette øker sjansen for konkurs. Denne sammenhengen ser vi også i bee swarm-plottene. Det gir mening at denne variabelen kan være viktig for selskaper innen bygg- og anleggsbransjen. Det er ikke uvanlig at selskaper innen bransjen tar opp kortsiktig gjeld i forbindelse med for eksempel kjøp av varer og tjenester. Kortsiktig gjeld kan innebære leverandørgjeld, skyldige offentlige avgifter, kassakreditt og skatt. Denne variabelen er også en del av variabelsettet til Paraschiv mfl. (2021), og kan dermed være en viktig variabel for flere bransjer. Basert på SHAP-verdiene i tabell 5.5, er *(kortsiktig gjeld- sum bankinnskudd, kontanter ol.)/ sum eiendeler*, viktigere enn *skyldige offentlige avgifter/ sum eiendeler*. Hvis vi ser på regresjonskoeffisientene har *skyldige offentlige avgifter/ sum eiendeler* en større koeffisientverdi i absoluttverdi, sammenlignet med *(kortsiktig gjeld- sum bankinnskudd, kontanter ol.)/ sum eiendeler*. En endring i *skyldige offentlige avgifter/ sum eiendeler* gir dermed større utslag på sjansen for konkurs. Vi ser altså noen forskjeller i viktigheten av variablene ved bruk av ulike

metoder.

Regresjonskoeffisienten til *leverandørgjeld/sum eiendeler* varierer mellom 0,456 og 1,907, og variabelen har høy SHAP-verdi flere perioder. Fortegnet er som forventet positivt og samsvarer med fargen på bee swarm-plottene. Det er ikke uvanlig at selskaper som kjøper varer og tjenester har leverandørgjeld. Store deler av utgiftene til for eksempel byggefirma er gjerne leverandørgjeld fra kjøp av byggematerialer. *leverandørgjeld/sum eiendeler* er også inkludert i SEBRA-modellen og variabelsettet utviklet av Paraschiv mfl. (2021).

Vi ser videre på forskjeller i hvilke variabler som velges ved LASSO og som har høy SHAP-verdi. Det er variasjon i variablene vi finner. Det kan skyldes at metodene fungerer på ulike måter. LASSO er en regulariseringsmetode som vi benytter til variabelseleksjon, og SHAP er en metode vi benytter til å tolke XGBoost, ved å se på SHAP-verdiene til variablene.

Variabler med høy SHAP-verdi over flere perioder, som ikke er valgt ved LASSO, er *sum egenkapital/ omsetning*, *egenkapital- immaterielle eiendeler)/(sum eiendeler- immaterielle eiendeler-sum bankinnskudd, kontanter ol.)*, *likvide midler/kortsiktig gjeld*, *leverandørgjeld/omsetning* og *opptjent egenkapital/kortsiktig gjeld*. Ingen av disse variablene er representert i de tre variabelsettene vi sammenligner med. De nevnte variablene gir informasjon om selskapets soliditet, likviditet og belåning. Fargen på bee swarm-plottene for variablene virker fornuftige i forhold til hvilken retning de påvirker sjansen for konkurs. En av årsakene til at disse variablene har høy SHAP-verdi, men ikke velges ved bruk av LASSO, kan være at XGBoost fanger opp eventuelle ikke-lineære forhold ved variablene, noe logistisk regresjonsmodeller ikke kan i like stor grad. I vedlegg V ser vi at det er en ikke-lineær sammenheng mellom variabelen *likvide midler/kortsiktig gjeld* og prediksjonene til modellene. Variabelen *skyldige offentlige avgifter/ sum eiendeler* har som nevnt høye SHAP-verdier og er valgt ved bruk av LASSO alle perioder. For denne variabelen ser vi en lineær sammenheng mellom variabelen og modellenes prediksjoner. Ikke-lineære sammenhenger kan altså være en del av forklaringen til at vi finner ulike variabler som viktige ved bruk av LASSO og SHAP. *Likvide midler/ kortsiktig gjeld* gir informasjon om midlene selskapet har til å dekke den kortsiktig gjelden. Det gir dermed mening at dette kan være et viktig mål for prediksjon av konkurs.

Det er to variabler som er valgt av LASSO alle perioder, men ikke noen av periodene basert på SHAP-verdiene. Det er variablene *dummy: totale forpliktelser > sum eiendeler* og *årsresultat/ sum eiendeler*. De gir henholdsvis informasjon om selskapets soliditet og lønnsomhet. Fortegnene til regresjonskoeffisientene til variablene er som forventet. De to nevnte variablene er også ansett som viktige av Paraschiv mfl. (2021), noe som peker i retning av at de er egnet for flere

bransjer, og ikke bare bygg- og anleggsbransjen. Det tyder også på at de kan være gode mål å benytte i konkursprediksjon.

Vi ser nærmere på de ulike kategoriene variablene representerer. De syv viktigste variablene valgt av LASSO hver periode, gir informasjon om selskapets belåning, lønnsomhet, likviditet og alder. Dette kan tyde på at disse kategoriene er viktig for bygg- og anleggsbransjen. Ingen av de syv variablene som ble valgt i alle periodene av LASSO gir informasjon om selskapets soliditet. Siden tidligere utviklede variabelsett består av variabler innen kategorien soliditet, trodde vi at LASSO også ville velge en eller flere variabler innen denne kategorien hver periode. Det kan argumenteres for at ikke alle momenter under kategorien soliditet alltid er like relevant ved vurdering av om et selskap vil gå konkurs (Eklund mfl., 2001). Eklund mfl. (2001) begrunner inkludering av soliditetsmål blant annet i at soliditet gir informasjon om bedriftens akkumulerte historiske inntjening. Det gir nyttig informasjon om selskapets evne til å tåle tap, men dårlig soliditet kan også komme av at selskapet er i en vekstfase. LASSO velger riktignok flere variabler som gir informasjon om selskapets soliditet, hvis vi ser på de som er valgt i færre perioder. De syv variablene med høyest SHAP-verdi flest perioder gir informasjon om selskapets belåning, likviditet, alder og soliditet. Det er ingen av variablene med høyest SHAP-verdi som inngår i kategorien lønnsomhet. Dette er ikke som forventet, ettersom at lønnsomhet er inkludert i tidligere utviklede variabelsett, samt gir viktig informasjon om selskapets evne til å skape resultater. Vi ser altså at metodene vi benytter utelater en kategori hver, hvis vi ser på de viktigste variablene. Variabler innen kategoriene belåning, likviditet og alder inkluderes alle periodene av LASSO og har høye SHAP-verdier flere perioder. De er altså viktige kategorier innen bygg- og anleggsbransjen. Det at variabler innen soliditet og lønnsomhet ikke velges ved begge metodene, kan tyde på at de ikke nødvendigvis er like viktige ved prediksjon av konkurs i bygg- og anleggsbransjen. Det kan også gi informasjon om at å benytte en modell til variabelseleksjon ikke er tilstrekkelig for valg av variabler. Å benytte flere tilnærminger i variabelseleksjon kan gi et mer helhetlig bilde på hvilke variabler som egner seg til konkursprediksjon for SMB i bygg- og anleggsbransjen. Alaka mfl. (2016) trekker frem likviditet, belåning og lønnsomhet blant de viktigste innen konkursprediksjon for bygg- og anleggsbransjen. Funnene fra LASSO og SHAP bekrefter at kategoriene likviditet og belåning er viktige. Vi ser også at basert LASSO er kategorien lønnsomhet viktig, men ifølge SHAP-verdiene, ser soliditet ut til å være en viktigere kategori enn lønnsomhet.

Ingen av de makroøkonomiske variablene ble valgt som viktig ved LASSO og SHAP. Det er nok mange variabler som kunne gitt gode modeller, men når kun et fåtall av de 160 variablene blir valgt, er det noen som er viktigere enn andre. De variablene som ble valgt som viktige i mo-

dellene ser ut til å være bedre egnet til å predikere konkurs for SMB i bygg- og anleggsbransjen, enn de makroøkonomiske variablene. Det kan være at de finansielle nøkkeltallene gjenspeiler makroøkonomiske hendelser. LASSO og XGBoost tar hensyn til variabler som korrelerer med hverandre. De logistiske regresjonsmodellene omtales også som *discrete hazard models*, som hensyntar risikoperioder. Det kan tenkes at hvis vi hadde trent modellene på flere år, kunne de makroøkonomiske variablene vært av større betydning. Det at modellene ikke velger makroøkonomiske variabler kan være nyttig å vite ved utvikling av nye konkursprediksjonsmodeller. Noen makroøkonomiske variabler kan være tidkrevende å samle inn, og det er derfor verdt å vurdere nytten av å inkludere de. Det kan dog være at andre metoder og inkludering av andre makroøkonomiske variabler gir andre resultater.

6.1.1 Sammenligning av variabelsett og XGBoost-modellene

For å undersøke om de variablene vi finner at egner seg for selskaper i bygg- og anleggsbransjen er unike for denne bransjen, sammenligner vi variablene vi finner som viktige, med tidligere utviklede variabelsett. Vi sammenligner også variabler med høye SHAP-verdier i XGBoost-modellene utviklet for bygg- og anleggsbransjen, med variabler i XGBoost-modellene utviklet for alle bransjer.

Det er kun variabelen *sum bankinnskudd, kontanter ol./ omløpsmidler* som er inkludert i variabelsettet til Altman og Sabato (2007), som også velges av LASSO. Vi trodde det skulle være flere likheter mellom variablene vi fant og variabelsettet til Altman og Sabato (2007), ettersom at de kun ser på SMB. Samtidig baseres Altman og Sabato (2007) sin studie på selskaper i USA, og variablene er valgt basert på tidligere forskning og trinnvis regresjon. Vi finner flere likheter med variablene i SEBRA-modellen. Variablene *log (alder i år)*, *leverandørgjeld / sum eiendeler* og *skyldige offentlige avgifter / sum eiendeler*, er felles for SEBRA-modellen og variablene funnet ved LASSO. De har også høye SHAP-verdier. Det virker fornuftig at vi finner noen likheter med SEBRA-modellen, da datagrunnlaget i oppgaven er basert på norske selskaper og SEBRA-modellen er utviklet for norske selskaper (Bernhardsen & Larsen, 2007).

Vi finner flest likheter mellom variablene vi anser som viktig og variabelsettet til Paraschiv mfl. (2021). De syv variablene som er valgt av LASSO alle perioder er også en del av variabelsettet til Paraschiv mfl. (2021). Fire av de syv variablene, nemlig *log (alder i år)*, *skyldige offentlige avgifter/sum eiendeler*, *(kortsiktig gjeld- sum bankinnskudd, kontanter ol.)/sum eiendeler* og *leverandørgjeld/ sum eiendeler*, har også høye SHAP-verdier. I tillegg er variablene *dummy: 1 hvis innbetalt egenkapital < sum egenkapital* og *rentekostnader/ sum eiendeler*, som er inkludert i variabelsettet fra Paraschiv mfl. (2021), valgt ved LASSO i henholdsvis seks og syv perioder.

Likhetene mellom de variablene vi finner som viktige og variabelsettet til Paraschiv mfl. (2021), kan skyldes at vi også benytter LASSO til variabelseleksjon og logistisk regresjon til å predikere konkurs. Vi benytter likt datagrunnlag, har tatt flere av de samme valgene ved preprossesering av data og ser kun på SMB. Den største forskjellen er at vi kun ser på bygg- og anleggsbransjen. Dette kan tyde på at det ikke nødvendigvis er store forskjeller i hvilke variabler som egner seg til å predikere konkurs i bygg- og anleggsbransjen, sammenlignet med alle bransjer. Det at vi finner mange like variabler tyder på at det kanskje ikke er nødvendig å finne variabler for spesifikke bransjer. Det kan likevel være at regresjonskoeffisientene i logistiske regresjonsmodeller hadde hatt ulike verdier for en modell utviklet for alle næringer, sammenlignet med våre modeller utviklet for bygg- og anleggsbransjen. Ved sammenligning av variablene fra Paraschiv mfl. (2021) med variablene med høy SHAP-verdier, er det også større forskjeller.

Selv om vi finner mange like variabler som Paraschiv mfl. (2021) ved bruk av LASSO, er det noen variabler LASSO velger, som ikke er en del av variabelsettet til Paraschiv mfl. (2021). Hvis vi ser på de 10 øverste variablene i tabell 5.1 og 5.2 er variablene *omsetning/sysselsatt kapital*, *omsetning/ sum egenkapital* og *omsetning/ sum varer*, valgt i syv eller flere perioder. Variablene kan gi informasjon om selskapets soliditet og aktivitet. De er ikke en del av variablene til Paraschiv mfl. (2021) og kan dermed være viktigere for bygg- og anleggsbransjen enn andre bransjer. De er basert på LASSO viktigere enn variablene *dummy: 1 hvis innbetalt egenkapital < sum egenkapital* og *rentekostnader/ sum eiendeler*, som er felles for våre valgte variabler ved bruk av LASSO og variablene funnet i Paraschiv mfl. (2021). *Omsetning/sum egenkapital* er signifikant i alle periodene den er valgt av LASSO, og kan være et godt mål på selskapets soliditet. Fortegnet til koeffisienten er positiv. Det vil si at store verdier av omsetning i forhold til egenkapitalen, øker sjansen for konkurs. Det virker fornuftig, da selskaper med høy egenkapital, ikke er like sårbar ved endringer i omsetning. Ettersom at alle de tre variablene inkluderer omsetning, kan det tyde på at omsetning er et viktig mål innen bygg- og anleggsbransjen. Om variablene skal benyttes i en modell, er det ikke sikkert at det er ønskelig å inkludere alle tre variablene ettersom at de kan gi informasjon om noe av det samme, samt at variablene kan korrelere.

For å undersøke om de variablene som egner seg for bygg- og anleggsbransjen er unike for denne bransjen, sammenligner vi videre variablene i tabell 5.5 med variablene i tabell 5.6. Tabellene viser skalerte SHAP-verdier for modeller utviklet med data for selskaper i bygg- og anleggsbransjen og alle bransjer. 15 av variablene er felles for begge tabellene, men det er forskjeller i hvor viktige variablene er i de ulike periodene, basert på SHAP-verdiene.

Kortsiktig gjeld - sum bankinnskudd, kontanter ol. / sum eiendeler og *skyldige offentlige avgifter/ sum eiendeler*, ser ut til å være noe viktigere for bygg- og anleggsbransjen, sammenlignet med alle bransjer, basert på SHAP-verdiene. De er henholdsvis nummer en og to i tabell 5.5, som viser skalerte SHAP-verdier for modellene for bygg- og anleggsbransjen. De er nummer fire og tre i tabell 5.6, for alle bransjer. *Log (alder i år)* har høyest SHAP-verdi for modellen utviklet basert på data fra alle bransjer. Den er nummer tre for modellene utviklet med data for bygg- og anleggsbransjen. Det ser dermed ut som *log (alder i år)* er noe viktigere for alle bransjer, sammenlignet med bygg- og anleggsbransjen.

Sum egenkapital/ omsetning, *likvide midler/ kortsiktig gjeld*, *opptjent egenkapital/ kortsiktig gjeld* og *(EK- IE) / (sum eiendeler- IE- sum bankinnskudd, kontanter ol.)*, har høyere SHAP-verdier for modellene utviklet basert på data fra bygg- og anleggsselskaper, sammenlignet med modellene utviklet for alle bransjer. Disse variablene ser dermed ut til å være viktigere for prediksjon av konkurs i bygg- og anleggsselskaper. I tabellen med skalerte SHAP-verdier for bygg- og anleggsbransjen er de henholdsvis nummer fire, fem, syv og åtte, men i tabellen for alle bransjer er de nummer syv, tolv, elleve og ti. *Likvide midler/ kortsiktig gjeld* gir informasjon om selskapets likviditet, som er en viktig kategori innen bygg- og anleggsbransjen i følge Alaka mfl. (2016). Det er riktignok andre variabler som gir informasjon om selskapets likviditet, blant annet *sum bankinnskudd, kontanter ol./ kortsiktig gjeld*, som har høyere SHAP-verdier for XGBoost-modellene i alle bransjer. *Opptjent egenkapital/ kortsiktig gjeld*, *sum egenkapital/ omsetning* og *(EK- IE) / (sum eiendeler- IE- sum bankinnskudd, kontanter ol.)* kan si noe om selskapets soliditet. Selv om disse variablene er inkludert i både bygg- og anleggsbransjen og alle bransjer, anses flere variabler inne kategorien soliditet som viktigere i bygg- og anleggsbransjen enn for alle bransjer.

Variabelen *log (sum eiendeler)* har høyere SHAP-verdi i XGBoost-modellene utviklet med data for alle bransjer, sammenlignet med modellene utviklet basert på bygg- og anleggsbransjen. *Log (sum eiendeler)* gir informasjon om selskapets størrelse. I modellene for alle bransjer er flere selskaper inkludert. Det kan tenkes at det er større forskjeller mellom selskapene for alle bransjer, og at størrelse dermed har mer å si for hvorvidt et selskap går konkurs, sammenlignet med bygg- og anleggsbransjen.

Leverandørgjeld/ sum eiendeler har høyest SHAP-verdi flere perioder for alle bransjer. Sammenlagt for alle periodene har denne nest høyest SHAP-verdi. For modellen utviklet for bygg- og anleggsbransjen er denne variabelen nummer ni i tabell 5.5, men variabelen *leverandørgjeld/ omsetning*, som også har leverandørgjeld som teller, er riktignok nummer seks i tabll 5.5. *Leve-*

randørgjeld/ sum eiendeler og *leverandørgjeld/ omsetning*, gir begge informasjon om selskapets likviditet, men på ulike måter. Det tyder på at for bygg- og anleggsselskaper kan forholdet mellom leverandørgjeld og omsetning være viktigere for å predikere konkurs, sammenlignet med forholdet mellom leverandørgjeld og sum eiendeler. Det er riktignok flere variabler med høye SHAP-verdier som beskriver selskapets likviditet.

Tabell 5.6, som viser SHAP-verdier for modellene for alle bransjer, inkluderer *årsresultat/ sum eiendeler*, noe tabell 5.5, for modellene med bygg- og anleggsbransjen ikke gjør. Denne variabelen blir også valgt ved bruk av LASSO. Variabelen er en av de 10 variablene med høyest SHAP-verdi i tabell 5.6, men kun i periode 3. Vi ser altså at lønnsomhetsmål ikke ser ut til å være blant de viktigste, basert på SHAP-verdier, både for XGBoost-modellene for bygg- og anleggsbransjen og alle bransjer.

Det er ikke alle variablene vi finner ved bruk av LASSO og SHAP som er kjente økonomiske nøkkeltall eller som er implementert i tidligere konkursprediksjonsmodeller. De kan likevel være gode mål til å predikere konkurs for selskaper i bygg- og anleggsbransjen. Hvorvidt variablene vi finner er gode mål til å predikere konkurs vil variere noe for ulike selskapene. Bygg- og anleggsbransjen består av mange ulike typer selskaper og tallene i regnskapet kan fremstille selskapenes økonomiske tilstand på ulike måter. Siden dataen vi benytter baseres på regnskapsinformasjon vil dette være av betydning. Det er verdt å merke seg at det er nyttig å vurdere flere enn noen få variabler eller nøkkeltall, hvis en ønsker å vurdere ett enkelt selskap. Det er mange forhold som spiller inn for at et selskap går konkurs. Innen bygg- og anleggsbransjen er det for eksempel mange selskaper som har prosjekter som går over lang tid, hvor det er store investeringer involvert. Hvis det er noe som ikke går som planlagt og etterspørselen ikke blir som forventet, kan dette ha konsekvenser for selskaper som involvert i prosjektene. Det er altså mange faktorer som spiller inn, men å kunne si noe om hvilke variabler som er viktige for konkursprediksjon i bygg- og anleggsbransjen kan likevel være nyttig. Vi ser at flere av variablene vi finner er de samme som er inkludert i tidligere utviklede variabelsett, som tyder på at de kan være gode mål til å predikere konkurs. Innsikten vi bidrar med kan også benyttes til å utvikle eventuelle nye konkursprediksjonsmodeller, eller benyttes av interessenter til å vurdere selskaper innen bransjen.

6.2 Modellenes prestasjon

Tabell 5.3 viser en oversikt over evalueringsmålene for de logistiske regresjonsmodellene hvor variablene valgt ved LASSO og Paraschiv mfl. (2021) er benyttet. I de aller fleste perioder er AUC på treningssettene høyere enn på testsettene, som forventet, sett bort fra noen enkelte perioder. I periode 1 og periode 11 er AUC-verdien på testsettene, med variablene valgt ved LASSO, lavest. AUC på testsettene for modellene med variablene valgt av Paraschiv mfl. (2021), er høyest i periode 11, og lavest i periode 1. I periode 11 er 2020 teståret, som er året koronapandemien kom til Norge. Det året var det færre konkurser sammenlignet med årene før. Det kan ha vært med på å påvirke hvordan den logistiske regresjonsmodellen med variablene valgt ved LASSO presterer.

De logistiske regresjonsmodellene med variablene valgt ved LASSO har bedre AUC-verdier på treningssettene alle periodene, sammenlignet med modellene med variablene fra Paraschiv mfl. (2021). I de fleste periodene er AUC på testsettene for modellene med LASSO-variablene, høyere enn for modellene med variablene fra Paraschiv mfl. (2021). Unntakene er i periode 3, 4, 6 og 11. Selv om modellene med variablene fra Paraschiv mfl. (2021) presterer bedre noen perioder, er det generelt de logistiske regresjonsmodellene med variablene valgt av LASSO som har høyest AUC de fleste periodene. Det tyder på at konkursprediksjonsmodeller utviklet for en spesifikk bransje presterer bedre enn modeller utviklet for alle bransjer. Dette samsvarer med Chava og Jarrow (2004) sitt funn om viktigheten av å inkludere industrieffekter i konkursprediksjonsmodeller. Brier Score på trenings- og testsettene er nesten identisk for modellene med variablene valgt av LASSO og variablene til Paraschiv mfl. (2021). Brier Score på testsettene er noe lavere for de logistiske regresjonsmodellene med variablene valgt ved LASSO, med unntak av i periode 7, 8, 9 og 11. Ettersom at det er ønskelig at Brier Score er så lav som mulig, tyder det også på at modellene utviklet for en spesifikk bransje presterer bedre enn modeller utviklet for alle bransjer. Det er riktignok ikke veldig store forskjeller i AUC og Brier Score for modellene hvor variablene fra LASSO benyttes og variablene til Paraschiv mfl. (2021). Det at det ikke er så store forskjeller kan skyldes at variablene som velges av LASSO er mange av de samme som variablene i Paraschiv mfl. (2021). Hvorvidt det er nyttig å utvikle konkursprediksjonsmodeller, må derfor vurderes i forhold til ressursene som kreves for å utvikle modeller for en spesifikk bransje og nytten man får ved de. Konsekvensene for de som berøres av konkurser kan være store og innebære store kostnader. En liten bedring i AUC vil dermed kunne være verdt ressursene det krever å utvikle modellene. Paraschiv mfl. (2021) påpeker for eksempel at en liten forbedring i modellen kan gi store utslag på lønnsomheten til banker.

Tabell 5.4 viser gjennomsnittsverdiene av evalueringsmålene for logistiske regresjonsmodeller,

ved bruk av variablene fra Altman og Sabato (2007), variablene i SEBRA-modellen, Paraschiv mfl. (2021) og variablene valgt ved bruk av LASSO. Vi ser at modellene som inkluderer variablene valgt ved bruk av LASSO og Paraschiv mfl. (2021) presterer best. Gjennomsnittlig AUC på treningssettene er best for modellen med variablene valgt ved LASSO. Gjennomsnittlig AUC på testsettene er høyest for modellen med variablene fra Paraschiv mfl. (2021). Vi så i tabell 5.3 at periode 11 trekker gjennomsnittverdien for modellen med variablene fra LASSO ned, og modellen for variablene fra (Paraschiv mfl., 2021) opp. Modellene med variablene fra Altman og Sabato (2007) presterer dårligst. Det kan skyldes at variablene fra Altman og Sabato (2007) ikke ble utviklet spesifikt for norske selskaper. Ulik regnskapspraksis og kjennetegn ved bedriftene kan ha betydning for hvilke variabler som egner seg best. Modellene ved bruk av variablene vi finner ved LASSO, presterer bedre enn variablene fra Altman og Sabato (2007) og variablene i SEBRA-modellen, bekrefter også at det kan være nyttig å utvikle modeller for spesifikke bransjer. Det at modellene ved bruk av variablene vi finner ved LASSO og variablene fra Paraschiv mfl. (2021) presterer best, kan også skyldes at variablene er valgt ved bruk av datasett med kun SMB. Dette stemmer overens med funnene til Pelja og Wahlstrøm (2021), som påpeker at størrelse påvirker en modell sin evne til å predikere konkurs.

Tabell 5.7 gir informasjon om hvordan XGBoost-modellene for bygg- og anleggsbransjen og for alle næringer presterer basert på evalueringsmålene AUC og Brier Score. AUC på treningssettene og testsettene er bedre for modellene utviklet med selskaper innen bygg- og anleggsbransjen, sammenlignet med modellene utviklet med alle bransjer, med unntak av periode 7, 8 og 9 på testsettene. Brier Score på treningssettene er lavere for alle bransjer, sammenlignet med bygg- og anleggsbransjen. På testsettene er Brier Score lik for bygg- og anleggsbransjen og alle bransjer, de fleste periodene.

Ved å se på AUC og Brier Score på testsettene ser vi hvordan modellene som er trent på data for alle bransjer, presterer når de testes på data for bygg- og anleggsbransjen. AUC på testsettene er bedre flest perioder for modellene trent med data for bygg- og anleggsbransjer. Dette tyder på at en modell utviklet for bygg- og anleggsbransjen presterer bedre enn en modell utviklet for alle bransjer, ved prediksjon av konkurs for bygg- og anleggsselskaper. Basert på Brier Score på testsettene presterer modellene nærmest likt. For ni av periodene er Brier Score lik for modellene utviklet for bygg- og anleggsbransjen og alle bransjer. I periode 4 og 7 er Brier Score 0,001 lavere for bygg- og anleggsbransjen, og i periode 3 er Brier Score 0,001 lavere for alle bransjer. Vi ser altså ikke like stor forskjell i hvordan modellene presterer basert på Brier Score, som ved AUC. Ulike evalueringsmål kan gi ulike resultater, ettersom de ikke måler det samme. Brier Score egner seg å benytte om formålet er å undersøke om de predikerte sannsynlighetene

samsvarer med de faktiske utfallene. Brier Score er relativt lik for modellene utviklet for bygg- og anleggsbransjen og alle bransjer, som tyder på at det ikke er store forskjeller i hvor godt modellene er kalibrert. AUC gir informasjon om sann positiv rate og falsk positiv rate, og gir altså mer informasjon om prediksjonene for ulike prediksjonsterskler. Det er et mål som egner seg til å evaluere modeller som er utviklet med et ubalansert datasett, noe vi har i denne oppgaven. Hvis vi ser på AUC tyder det på at modellen utviklet for bygg- og anleggsbransjen er bedre til å klassifisere riktig på tvers av terskelverdiene, sammenlignet med modellen for alle bransjer. Vi ser altså at modellene forbedres noe ved å utvikles for alle bransjer. Her er det også relevant å vurdere nytten av å utvikle modeller for en spesifikk bransje opp mot ressursene det krever.

Til slutt sammenligner vi prestasjonsmålene til de logistiske regresjonsmodellene utviklet med variabler for bygg- og anleggsbransjen, med XGBoost-modellene for bygg- og anleggsbransjen. XGBoost-modellene presterer stort sett bedre basert på evalueringsmålene sammenlignet med de logistiske regresjonsmodellene. AUC er best for XGBoost-modellene de aller fleste periodene. Brier Score er lik de fleste periodene, og bedre for XGBoost-modellene noen få perioder. Dette er ikke overraskende i forhold til hva tidligere studier har funnet. Lin og Bai (2022) fant i sin studie at XGBoost presterte bedre enn logistisk regresjon. Evalueringsmålene i periode 11 skiller seg ut ved at AUC og Brier Score for testsettet er dårligere for logistiske regresjonsmodellen og bedre for XGBoost-modellen, sammenlignet med de andre periodene. Som nevnt er teståret i periode 11, 2020, hvor det var det færre konkurser sammenlignet med andre år. Dette kan ha utspilt seg forskjellig i de ulike modellene. XGBoost og logistisk regresjon fungerer på ulike måter og vi har sett at vi finner ulike variabler som viktige ved bruk av metodene. XGBoost er en maskinlæringsteknikk som kombinerer flere metoder og er kjent for å oppnå høy treffsikkerhet (Wahlstrøm, 2023). Det at XGBoost-modellene presterer bedre enn de logistiske regresjonsmodellene, kan tale for at vi burde legge større vekt på funnene ved bruk av XGBoost. På den andre siden er funnene fra de logistiske regresjonsmodellene også viktig, da variablene gir et godt sammenligningsgrunnlag med tidligere studier. LASSO velger også en variabel innen lønnsomhet og flere av de samme variablene flere perioder. Det er også interessant å se hvilke variabler vi finner ved bruk av både LASSO og XGBooost med SHAP.

Metodene vi benytter fungerer på ulike måter. Logistisk regresjon benyttes sammen med variabelseleksjonsmetoden LASSO. XGBoost-modellene tilpasser en modell ved bruk av alle de 160 variablene og SHAP-verdiene gir informasjon om viktigheten av hver variabel. XGBoost er en metode som krever tuning av flere hyperparametere. Ved LASSO velger vi kun hyperparameteren λ. XGBoost påvirkes dermed i større grad av våre valg når det kommer til tuning. XGBoost kan også fange opp eventuelle ikke-lineære forhold ved variablene, noe logistisk regresjons-

modeller ikke gjør i like stor grad. Det kan være noen av årsakene til at vi ikke får de samme variablene ved bruk av logistisk regresjon og XGBoost. Hvis formålet er å implementere variablene i en modell som skal benyttes til å predikere konkurs, kan man vurdere hvilke variabler som skal vektlegges, samt hvilken modell man skal benytte. Det er mulig å benytte de variablene vi presenterer med andre metoder for prediksjon av konkurs. Variablene og evalueringsmålene vi presenterer kan være nyttig ved valg av variabler og metode i konkursprediksjonsmodeller.

Kapittel 7

Konklusjon

I denne masteroppgaven ønsker vi å bidra med ny innsikt innen viktige variabler for konkursprediksjon hos SMB i bygg- og anleggsbransjen, og si noe om nytten av å utvikle konkursprediksjonsmodeller for en spesifikk bransje. Vi forsøker å besvare følgende problemstilling: *Hvilke variabler egner seg til å predikere konkurs for SMB i bygg- og anleggsbransjen, og forbedres modellene av å utvikles for en spesifikk bransje?* Vi benytter et variabelsett på 160 variabler, samt metodene logistisk regresjon med LASSO og XGBoost med SHAP. De variablene vi finner som viktige, har vi sammenlignet med tre eksisterende variabelsett. For å undersøke i hvilken grad modellene forbedres av å utvikles for en spesifikk bransje, har vi sammenlignet evalueringsmål for modellene.

Vi finner vi flere viktige variabler for prediksjon av konkurs for SMB i bygg- og anleggsbransjen. LASSO velger syv variabler alle periodene innenfor kategoriene belåning, lønnsomhet, likviditet og alder. De seks variablene med høyest SHAP-verdi gir informasjon om selskapets belåning, likviditet, alder og soliditet. Fire variabler er valgt av LASSO alle periodene og har høye SHAP-verdier. De anses dermed som blant de best egnede variablene for prediksjon av konkurs i bygg- og anleggsbransjen. De gir informasjon om selskapets belåning, likviditet og alder, og er følgelig sentrale kategorier for bygg- og anleggsbransjen. De største forskjellene vi finner ved LASSO og XGBoost med SHAP, er at lønnsomhet ser ut til å være en viktigere kategori ved bruk av LASSO, og soliditet en viktigere kategori basert på SHAP-verdier. Det at kategoriene belåning, likviditet og lønnsomhet er viktige innen bygg- og anleggsbransjen, stemmer overens med funnene i Alaka mfl. (2016). Ingen av de makroøkonomiske variablene viser seg å være viktige. Innsikten i hvilke variabler som egner seg for konkursprediksjon for bygg- og anleggsbransjen, kan være nyttig for blant annet banker, investorer, leverandører og selskapene selv, eller ved utvikling av nye konkursprediksjonsmodeller.

Det er flere likheter mellom variablene vi finner som viktige i bygg- og anleggsbransjen, sammenlignet med tidligere utviklede variabelsett. Det tyder på at variablene vi finner, egner seg til prediksjon av konkurs, samt at flere variabler egner seg for både bygg- og anleggsbransjen og andre bransjer. Vi finner riktignok variabler som skiller seg ut for bygg- og anleggsbransjen. XGBoost-modellene utviklet basert på bygg- og anleggsselskaper har en bedre AUC på testsettet de fleste periodene, sammenlignet med XGBoost-modellene utviklet for alle bransjer. De logistiske regresjonsmodellene som inkluderer variablene vi fant ved LASSO, presterer også bedre i de fleste periodene, sammenlignet med modellene hvor de tidligere utviklede variabelsettene benyttes. Vi ser altså at modeller utviklet for bygg- og anleggsbransjen presterer bedre enn modeller utviklet for flere bransjer de fleste periodene. Det at vi finner noen variabler som skiller seg ut for bransjen og at modellene presterer noe bedre, tyder på at det kan være nyttig å utvikle konkursprediksjonsmodeller for en spesifikk bransje.

7.1 Svakheter ved oppgaven

I oppgaven forsøker vi å noe om forskjeller i hvilke variabler som er viktige i bygg- og anleggsbransjen, sammenlignet med alle bransjer. For å kunne si noe om dette benytter vi XGBoost-modeller både for bygg- og anleggsbransjen og alle bransjer. Vi kunne også benyttet LASSO til variabelseleksjon for alle bransjer. Ettersom at det er gjennomført i lignende studier, hvor en av de benyttet samme variabelseleksjonsmetode, valgte vi heller å sammenligne med flere tidligere utviklede variabelsett, som er utarbeidet på ulike grunnlag.

Vi finner mange forskjellige variabler som viktige ved bruk av metodene LASSO og XGBoost med SHAP. Det er også noe variasjon i hvilke variabler LASSO velger, og hvilke variabler som har høye SHAP-verdier, fra periode til periode. Det at det er mye variasjon i hvilke variabler vi finner, kan gjøre det vanskelig å stole på resultatene våre og påstå at de variablene vi finner, er viktige for prediksjon av konkurs hos SMB i bygg- og anleggsbransjen. Metodene vi benytter fungerer dog på ulike måter og kan forklare at de indikerer at ulike variabler er viktige.

Ved tolkning av XGBoost-modellen valgt å kun se på de 10 variablene med høyest SHAP-verdi, og her kunne vi ha inkludert flere variabler. Vi kunne også inkludert flere variabler med LASSO, ved å velge en mindre streng λ. Dette ville gitt mer informasjon, men ikke nødvendigvis en bedre forståelse av de best egnede variablene. Vi inkluderer også fem makroøkonomiske variabler, og en av de var BNP. Hol (2007) nevnte BNP-gapet som en viktig variabel innen konkursprediksjon. Vi kunne ha valgt BNP-gapet, fremfor BNP som en variabel. Det ville gitt et tall som gir mer informasjon om BNP i forhold til andre perioder.

Ved bruk av XGBoost-modellen valgte vi selv hvilke hyperparametre vi ønsket å tune, samt ett sett med verdier i tuningen. Vi kunne ha tunet flere hyperparametre og valgt flere verdier i tuningen for en enda bedre modell. Dette var en avveining på hvor lang tid tuningen ville ta. Vi kunne også ha tunet hyperparametere for modellen utviklet for alle bransjer.

7.2 Videre forskning

I denne oppgaven bidrar vi med innsikt innen et avgrenset området innenfor konkursprediksjon. Det finnes mange interessante vinklinger innenfor forskningsfeltet og det er mye som kan være nyttig å forske videre på.

Det hadde vært interessant å benytte andre metoder og tilnærminger for variabelseleksjon og prediksjon av konkurs. Eventuelle funn kan være med å bekrefte våre resultater eller gi andre resultater.

Videre er det mulig å inkludere flere variabler for å se om de er viktige ved prediksjon av konkurs for SMB i bygg- og anleggsbransjen. Eksempler på flere variabler er ikke-finansielle variabler, variabler som gir informasjon om ledelseseffektivitet, som Alaka mfl. (2016) påpeker som viktig innen bygg- og anleggsbransjen og flere markedsvariabler. Slike variabler kan være tidkrevende å samle inn og en avveining av hvorvidt variablene gir merverdi til analysen må dermed vurderes. Andre makroøkonomiske variabler kan også være interessant å inkluderes i en konkursprediksjonsmodell.

Denne oppgaven fokuserer kun på en bransje. Det hadde vært interessant å se på flere bransjer for å finne ut om det det er forskjeller mellom hvilke variabler som er viktige for bransjene, eller få bedre innsikt i en enkelt bransje. Det er mulig å utvikle en modell som kan benyttes til å predikere konkurs for selskaper innen bygg- og anleggsbransjen. Ved utvikling av en ny modell kan våre funn benyttes. Det må vurderes hvilken modell som egner seg og hvilke variabler som skal inkluderes.

Til slutt hadde det vært interessant å se på om ulike variabler er viktige for selskaper av ulik størrelse. Det kan bli utviklet modeller kun for store selskaper, og sammenlignet med modeller for SMB i bygg- og anleggsbransjen. Da kunne vi ha sett på ulikhetene ved størrelse på bedrift for bransjen, og hvilke variabler som er viktige.

Referanseliste

Alaka, H., Oyedele, L., Owolabi, H., Akinade, O., Bilal, M., & Ajayi, S. (2018). A big data analytics approach for construction firms failure prediction models. *IEEE Transactions on Engineering Management*, *66*(4), 689–698. https://doi.org/10.1109/TEM.2018.2856376

Alaka, H., Oyedele, L. O., Owolabi, H. A., Oyedele, A. A., Akinade, O. O., Bilal, M., & Ajayi, S. O. (2016). Critical factors for insolvency prediction: towards a theoretical model for the construction industry. *International Journal of Construction Management*, *17*(1), 25–49. https://doi.org/10.1080/15623599.2016.1166546

Altinn. (2021, 18. november). *Konkurs i aksjeselskap*. Hentet 21. februar 2023, fra https://www.altinn.no/starte-og-drive/avvikling-sletting-og-konkurs/konkurs/konkurs-i-aksjeselskap/

Altman, E. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, *23*(4), 589–609. https://doi.org/https://doi.org/10.2307/2978933

Altman, E. (2000). Revisiting the Z-Score and ZETA®. *Predicting Financial Distress of Companies*, 2–16. https://doi.org/10.4337/9780857936097.00027

Altman, E., Haldeman, G., Robert, & Narayanan, P. (1977). Zeta analysis: A new model to identify bankruptcy risk of corporations. *Journal of Banking and Finance*, *10*, 29–54. https://doi.org/https://doi.org/10.1016/0378-4266(77)90017-6

Altman, E., & Sabato, G. (2007). Modelling credit risk for SMEs: Evidence from the US market. *Abacus*, *43*(3), 332–357. https://doi.org/https://doi.org/10.1111/j.1467-6281.2007.00234.x

BDO. (2022 oktober). *Bygg og anleggsanalysen 2022*. Hentet 28. februar 2023, fra https://www.bdo.no/nb-no/bransjer-nb/bygg-og-anlegg/bransjeanalysen

Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of Accounting Research*, 71–111. https://doi.org/https://doi.org/10.2307/2490171

Bernhardsen, E., & Larsen, K. (2007). Modelling credit risk in the enterprise sector-further development of the SEBRA model. *Norges Bank. Economic Bulletin*, *78*(3), 102.

Blöchlinger, A., & Leippold, M. (2006). Economic benefit of powerful credit scoring. *Journal of Banking & Finance*, *30*(3), 851–873. https://doi.org/10.1016/j.jbankfin.2005.07.014

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, *78*(1), 1–3. https://doi.org/https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

Brzezinski, D., & Stefanowski, J. (2017). Prequential AUC: properties of the area under the ROC curve for data streams with concept drift. *Knowledge and Information Systems*, *52*, 531–562. https://doi.org/https://doi.org/10.1007/s10115-017-1022-8

Campbell, J. Y., Hilscher, J., & Szilagyi, J. (2008). In search of distress risk. *The Journal of Finance*, *63*(6), 2899–2939. https://doi.org/https://doi.org/10.1111/j.1540-6261.2008.01416.x

Chava, S., & Jarrow, R. A. (2004). Bankruptcy prediction with industry effects. *Review of Finance*, *8*(4), 537–569. https://doi.org/https://doi.org/10.1093/rof/8.4.537

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. http://dx.doi.org/10.1145/2939672.2939785

Dahlum, S. (2021, 9. mars). *Validitet*. Hentet 13. mars 2023, fra https://snl.no/validitet

Dubey, P. (1975). On the uniqueness of the Shapley value. *International Journal of Game Theory*, *4*(3), 131–139. https://link.springer.com/article/10.1007/BF01780630

Edmister, R. O. (1972). An Empirical Test of Financial Ratio Analysis for Small Business Failure Prediction. *The Journal of Financial and Quantitative Analysis*, *7*(2), 1477–1493. https://doi.org/https://doi.org/10.2307/2329929

Eklund, T., Larsen, K., & Berhardsen, E. (2001). Model for analysing credit risk in the enterprise sector. *Norges Bank. Economic Bulletin*, *72*(3), 99. http://hdl.handle.net/11250/2480734

European Comission. (2003, 6. mai). *Internal Market, Industry, Entrepreneurship and SMEs*. Hentet 21. februar 2023, fra https://single-market-economy.ec.europa.eu/smes/sme-definition_en

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232. http://www.jstor.org/stable/2699986

Hastie, T., Qian, J., & Tay, K. (2021). An Introduction to glmnet. *CRAN R Repositary*. https://glmnet.stanford.edu/articles/glmnet.html

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Bd. 2). Springer.

Haugen, A. K. L. (2020, 13. august). *Koronarelatert fall i bygge- og anleggsaktiviteten i 2. kvartal*. Hentet 28. februar 2023, fra https://www.ssb.no/bygg-bolig-og-eiendom/artikler-og-publikasjoner/koronarelatert-fall-i-bygge-og-anleggsaktiviteten-i-2.kvartal

Hol, S. (2007). The influence of the business cycle on bankruptcy probability. *International transactions in operational research*, *14*(1), 75–90. https://doi.org/https://doi.org/10.1111/j.1475-3995.2006.00576.x

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Bd. 398). John Wiley & Sons.

Härdle, W., Lee, Y.-J., Schäfer, D., & Yeh, Y.-R. (2009). Variable selection and oversampling in the use of smooth support vector machines for predicting the default risk of companies. *Journal of Forecasting*, *28*(6), 512–534. DOI:10.1002/for.1109

Jabeur, S. B., Stef, N., & Carmona, P. (2022). Bankruptcy prediction using the XGBoost algorithm and variable importance feature engineering. *Computational Economics*, 1–27. https://doi.org/https://doi.org/10.1007/s10614-021-10227-1

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Bd. 112). Springer.

Kumar, P. R., & Ravi, V. (2007). Bankruptcy prediction in banks and firms via statistical and intelligent techniques–A review. *European journal of operational research*, *180*(1), 1–28. https://doi.org/https://doi.org/10.1016/j.ejor.2006.08.043

Lin, B., & Bai, R. (2022). Machine learning approaches for explaining determinants of the debt financing in heavy-polluting enterprises. *Finance Research Letters*, *44*, 102094. https://doi.org/https://doi.org/10.1016/j.frl.2021.102094

Lundberg, S. (2018a). *beeswarm plot*. Hentet 11. mai 2023, fra https://shap.readthedocs.io/en/latest/example_notebooks/api_examples/plots/beeswarm.html

Lundberg, S. (2018b). *scatter plot*. Hentet 16. mai 2023, fra https://shap.readthedocs.io/en/latest/example_notebooks/api_examples/plots/scatter.html#Simple-dependence-scatter-plot

Lundberg, S., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. https://arxiv.org/abs/1705.07874

Lundberg, Scott. (2018). *Welcome to the SHAP documentation*. Hentet 12. mai 2023, fra https://shap.readthedocs.io/en/latest/index.html

Mansi, S. A., Maxwell, W. F., & Zhang, A. (2012). Bankruptcy prediction models and the cost of debt. *Journal of Fixed Income*. https://doi.org/10.3905/jfi.2012.21.4.025

Marschner, I., & Donoghoe, M. W. (2018). *glm2: Fitting Generalized Linear Models*. Hentet 12. mai 2023, fra https://CRAN.R-project.org/package=glm2

Norges Bank. (2023, 20. januar). *Styringsrenten*. Hentet 26. februar 2023, fra https://www.norges-bank.no/tema/pengepolitikk/Styringsrenten/

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of accounting research*, 109–131. https://doi.org/https://doi.org/10.2307/2490395

Paraschiv, F., Schmid, M., & Wahlstrøm, R. R. (2021). Bankruptcy prediction of privately held SMEs using feature selection methods. *SSRN Electronic Journal*. https://doi.org/http://dx.doi.org/10.2139/ssrn.3911490

Pelja, I., & Wahlstrøm, R. R. (2021). Hvordan påvirker bedriftens størrelse predikering av konkurs? *Magma Tidsskrift for Økonomi Og Ledelse, 7, 82–91*. https://nye.econa.no/faglig-oppdatering/medlemsbladet-magma/7-2021/hvordan-pavirker-bedriftens-storrelse-predikering-av-konkurs/

Regjeringen. (1997). *Statlige anskaffelser - utfordringer for næringslivet, herunder små og mellomstore bedrifter*. Hentet 21. februar 2023, fra https://www.regjeringen.no/no/dokumenter/nou-1997-21/id141007/?ch=7

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC bioinformatics*, *12*(1), 1–8.

Rufibach, K. (2010). Use of Brier score to assess binary predictions. *Journal of clinical epidemiology*, *63*(8), 938–939. https://doi.org/https://doi.org/10.1016/j.jclinepi.2009.11.009

Saraswat, M. (2016). Beginners tutorial on xgboost and parameter tuning in r. https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/beginners-tutorial-on-xgboost-parameter-tuning-r/tutorial/

Schalck, C., & Yankol-Schalck, M. (2021). Predicting French SME failures: new evidence from machine learning techniques. *Applied Economics*, *53*(51), 5948–5963. https://doi.org/https://doi.org/10.1080/00036846.2021.1934389

scikit-learn developers. (2007-2018). *scikit-learn- machine learning in python*. Hentet 12. mai 2023, fra https://scikit-learn.org/stable/install.html

Shapley, L. S. (1953). *A value for n-person games*. Princeton University Press Princeton. https://apps.dtic.mil/sti/citations/AD0604084

Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. *The journal of business*, *74*(1), 101–124. https://doi.org/https://doi.org/10.1086/209665

Statistisk sentralbyrå. (2009, 31. januar). *Standard for næringsgruppering (SN)*. Hentet 31. januar 2023, fra https://www.ssb.no/klass/klassifikasjoner/6

Statistisk sentralbyrå. (2022a). *Renter i banker og kredittforetak*. Hentet 15. mars 2023, fra https://www.ssb.no/statbank/table/09381

Statistisk sentralbyrå. (2022b, 25. oktober). *Næringenes økonomiske utvikling*. Hentet 1. februar 2023, fra https://www.ssb.no/statbank/table/12817/tableViewLayout1/

Statistisk sentralbyrå. (2023a, 5. januar). *Fakta om norsk økonomi*. Hentet 26. februar 2023, fra https://www.ssb.no/nasjonalregnskap-og-konjunkturer/faktaside/norsk-okonomi

Statistisk sentralbyrå. (2023b, 5. januar). *Virksomheter*. Hentet 21. februar 2023, fra https://www.ssb.no/virksomheter-foretak-og-regnskap/virksomheter-og-foretak/statistikk/virksomheter

Statistisk sentralbyrå. (2023c, 25. januar). *Oppna konkurser*. Hentet 1. februar 2023, fra https://www.ssb.no/virksomheter-foretak-og-regnskap/konkurser/statistikk/opna-konkursar

Statistisk sentralbyrå. (2023d, 28. februar). *Produksjonsindeks for bygge- og anleggsvirksomhet*. Hentet 28. februar 2023, fra https://www.ssb.no/bygg-bolig-og-eiendom/bygg-og-anlegg/statistikk/produksjonsindeks-for-bygge-og-anleggsvirksomhet

Svartdal, F. (2020, 3. april). *Reliabilitet*. Hentet 14. mars 2023, fra https://snl.no/reliabilitet

Tian, S., Yu, Y., & Guo, H. (2015). Variable selection and corporate bankruptcy forecasts. *Journal of Banking & Finance*, *52*, 89–100. https://doi.org/https://doi.org/10.1016/j.jbankfin.2014.12.003

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tjernshaugen, A., Hiis, H., Bernt, J. F., Braut, G. S., Bahus, V. B., & Simonsen, M. (2023, 31. januar). *Koronapandemien*. Hentet 7. mai 2023, fra https://sml.snl.no/koronapandemien

Tjoa, E., & Guan, C. (2020). A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE transactions on neural networks and learning systems*, *32*(11), 4793–4813. https://doi.org/10.1109/TNNLS.2020.3027314

Wahlstrøm, R. R. (2022, 7. november). *Financial statements of companies in Norway*. https://doi.org/https://doi.org/10.48550/arXiv.2203.12842

Wahlstrøm, R. R. (2023). Explainable Artificial Intelligence (xAI) for Interpreting Machine Learning Methods and Their Individual Predictions. *Available at SSRN 4321303*. https://doi.org/http://dx.doi.org/10.2139/ssrn.4321303

Wickham, H., Chang, W., Henry, L., Pedersen, T., Takahashi, K., Wilke, Woo, K., Yutani, H., & Dunningston, D. (n.d). *ggplot2*. Hentet 12. mai 2023, fra https://ggplot2.tidyverse.org/

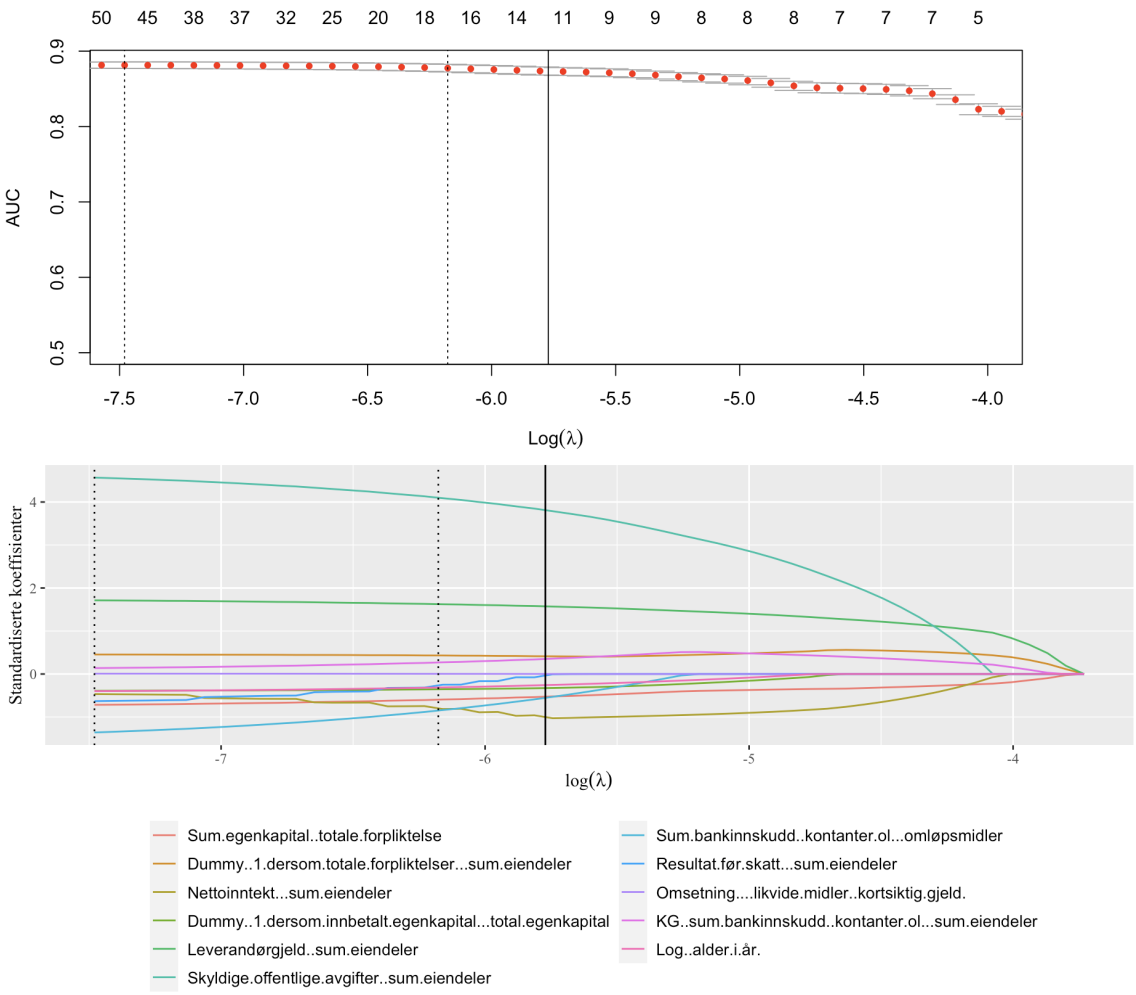
XGBoost developers. (2022). *XGBoost Python Package*. Hentet 12. mai 2023, fra https://xgboost.readthedocs.io/en/stable/python/index.html

Xiaomao, X., Xudong, Z., & Yuanfang, W. (2019). A comparison of feature selection methodology for solving classification problems in finance. *Journal of Physics: Conference Series*, *1284*(1), 012026. https://doi.org/10.1088/1742-6596/1284/1/012026

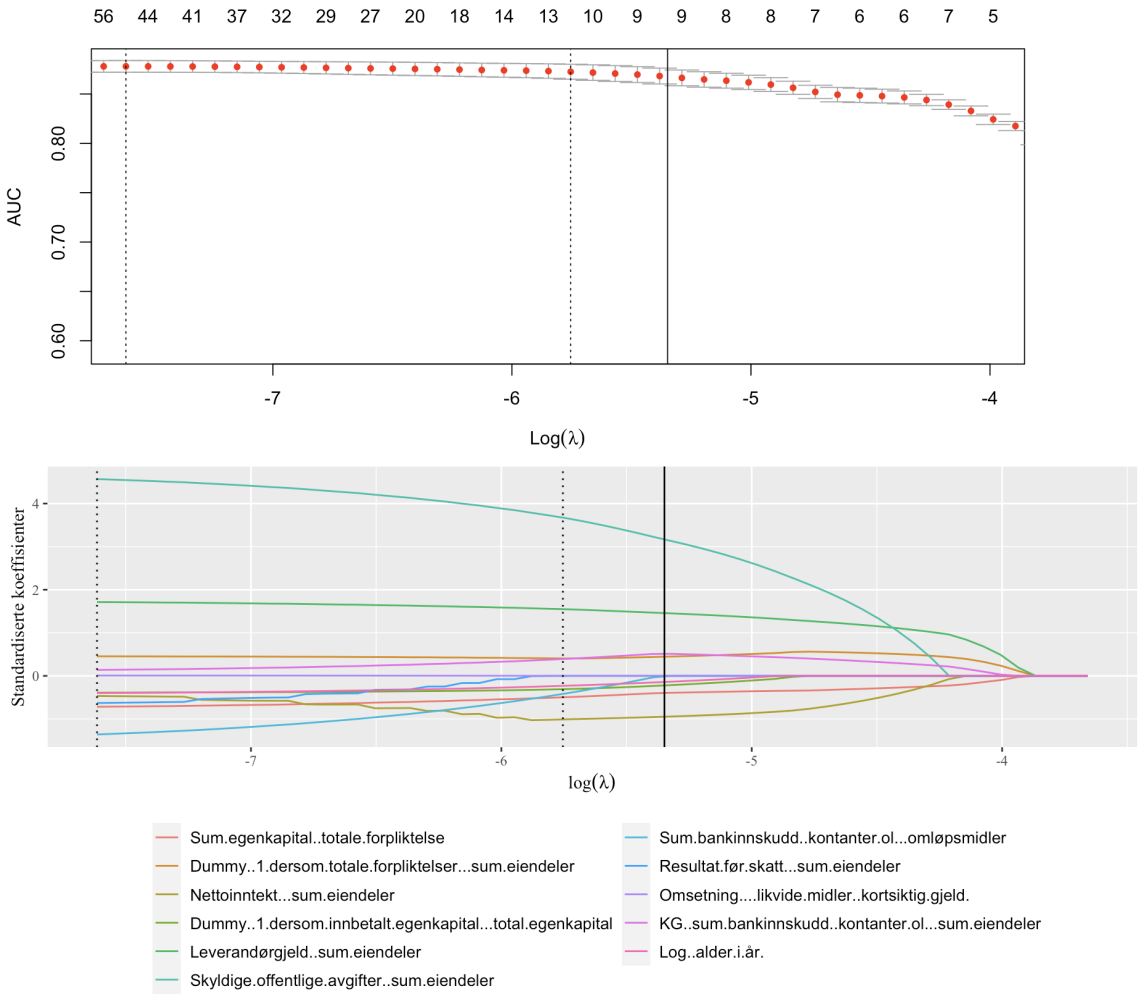
Zmijewski, M. E. (1984). Methodological issues related to the estimation of financial distress prediction models. *Journal of Accounting research*, 59–82. https://doi.org/https://doi.org/10.2307/2490859

Vedlegg

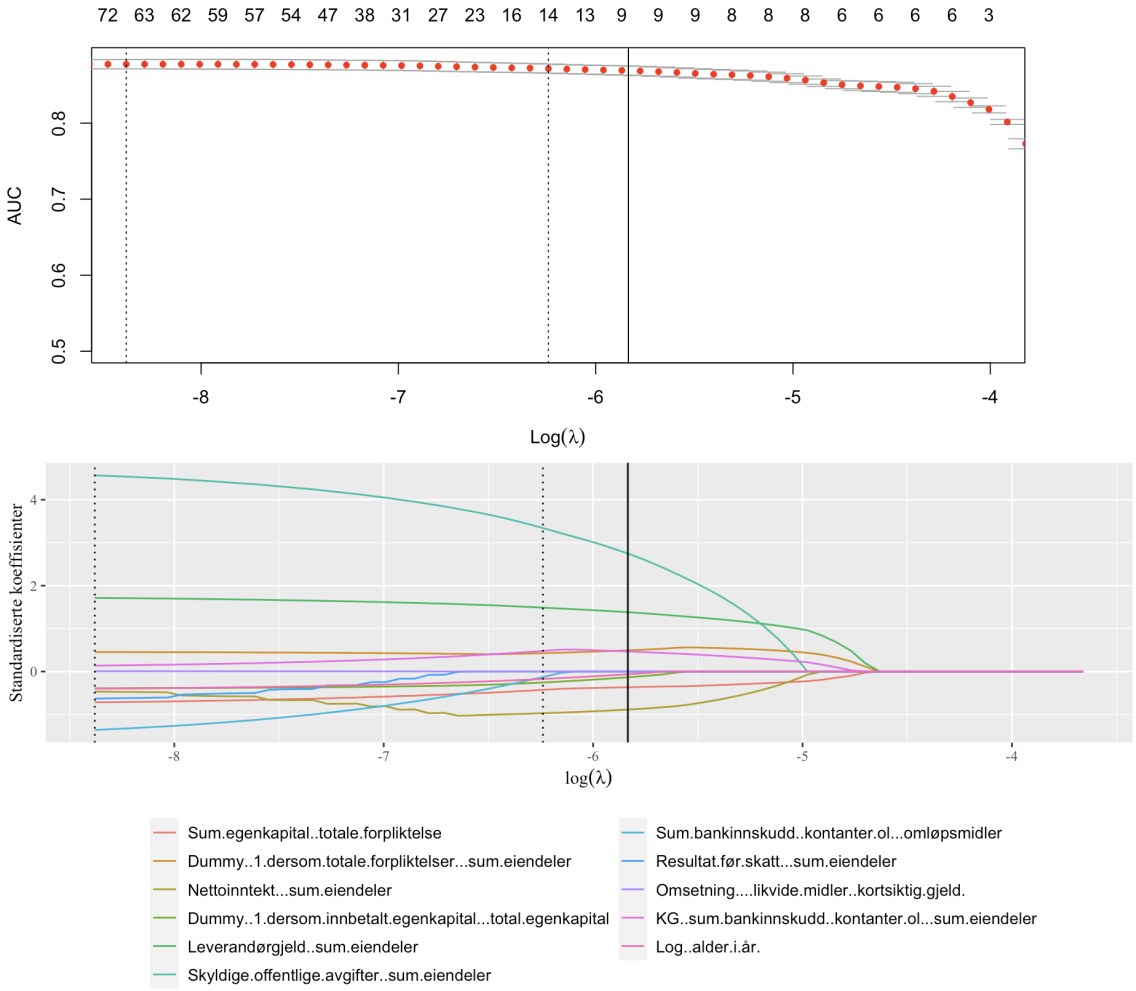
I AUC- og LASSO-plot



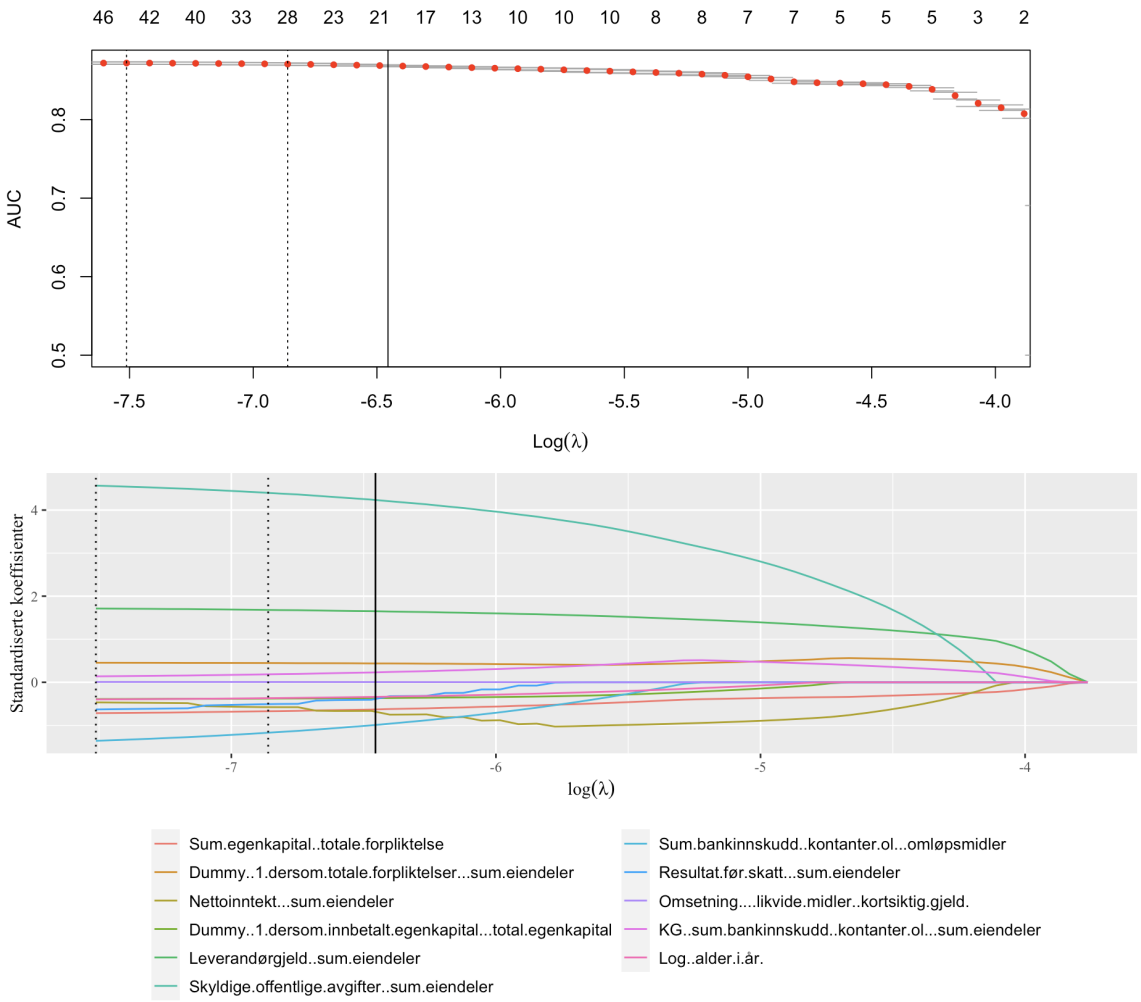
Periode 1. AUC og Standardiserte koeffisienter, over log(λ)-verdier. Den tykkeste linjen er lik log(λ) hvor vi tillater 1,5 standardfeil. Variablene til høyre for linjen velges av LASSO.



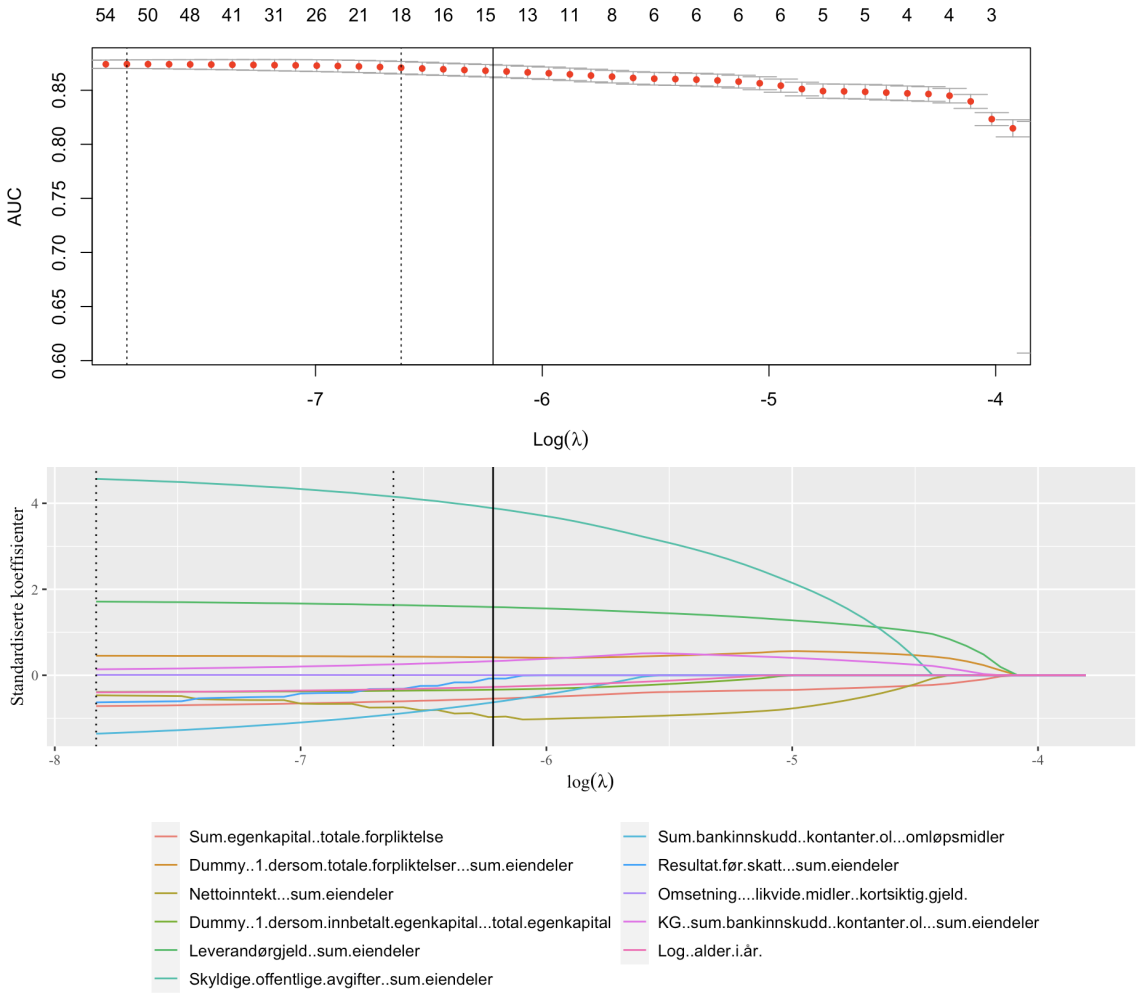
Periode 2. AUC og Standardiserte koeffisienter, over log(λ)-verdier. Den tykkeste linjen er lik log(λ) hvor vi tillater 1,5 standardfeil. Variablene til høyre for linjen velges av LASSO.



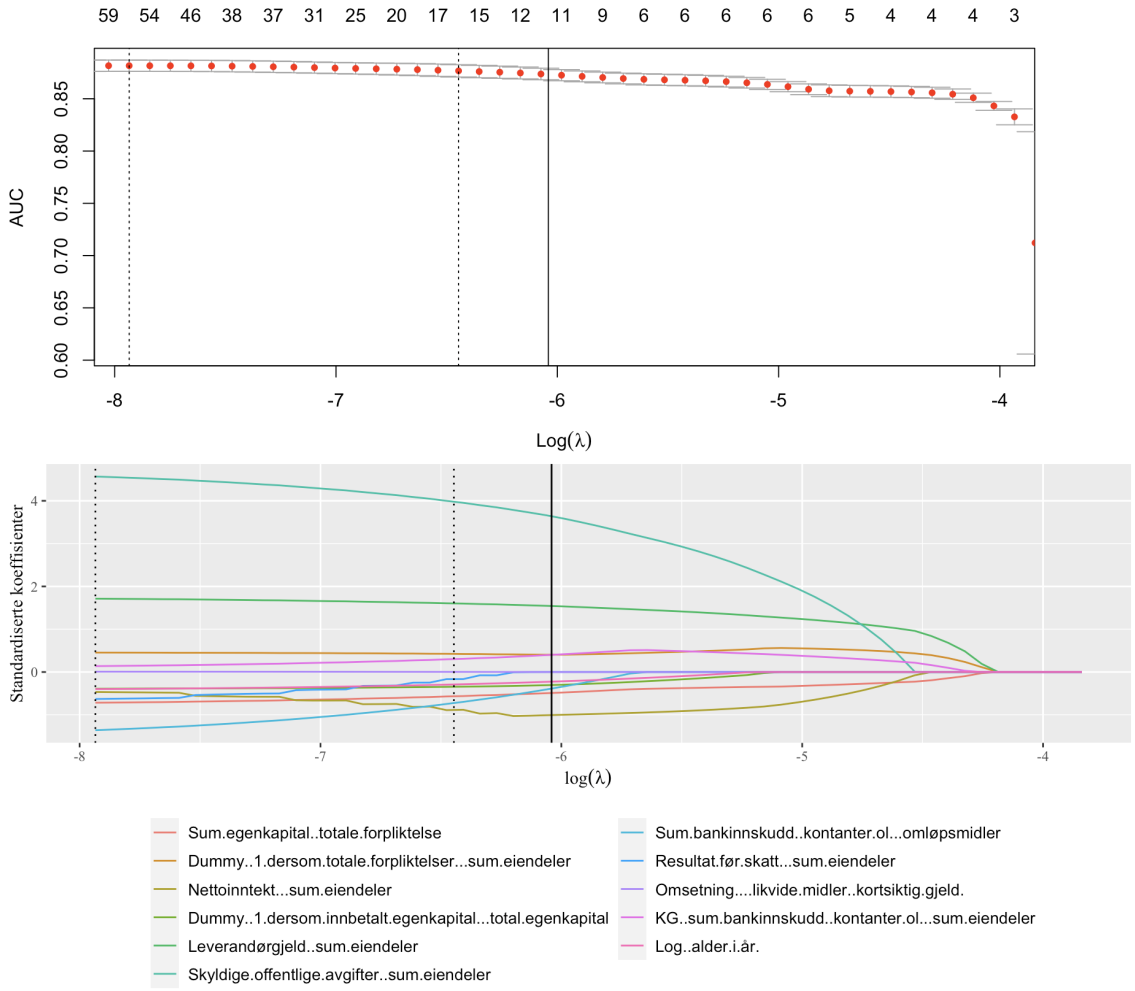
Periode 3. AUC og Standardiserte koeffisienter, over log(λ)-verdier. Den tykkeste linjen er lik log(λ) hvor vi tillater 1,5 standardfeil. Variablene til høyre for linjen velges av LASSO.



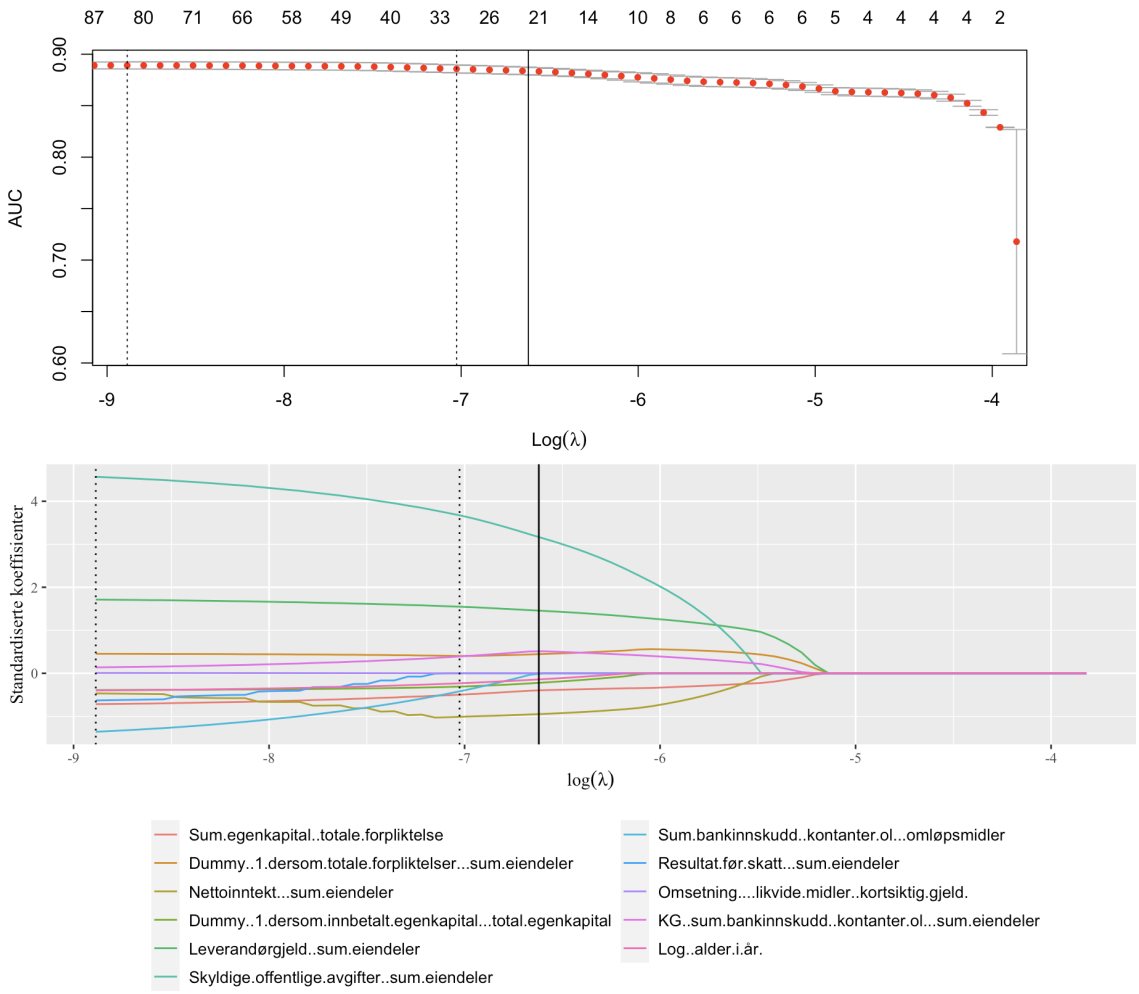
Periode 4. AUC og Standardiserte koeffisienter, over log(λ)-verdier. Den tykkeste linjen er lik log(λ) hvor vi tillater 1,5 standardfeil. Variablene til høyre for linjen velges av LASSO.



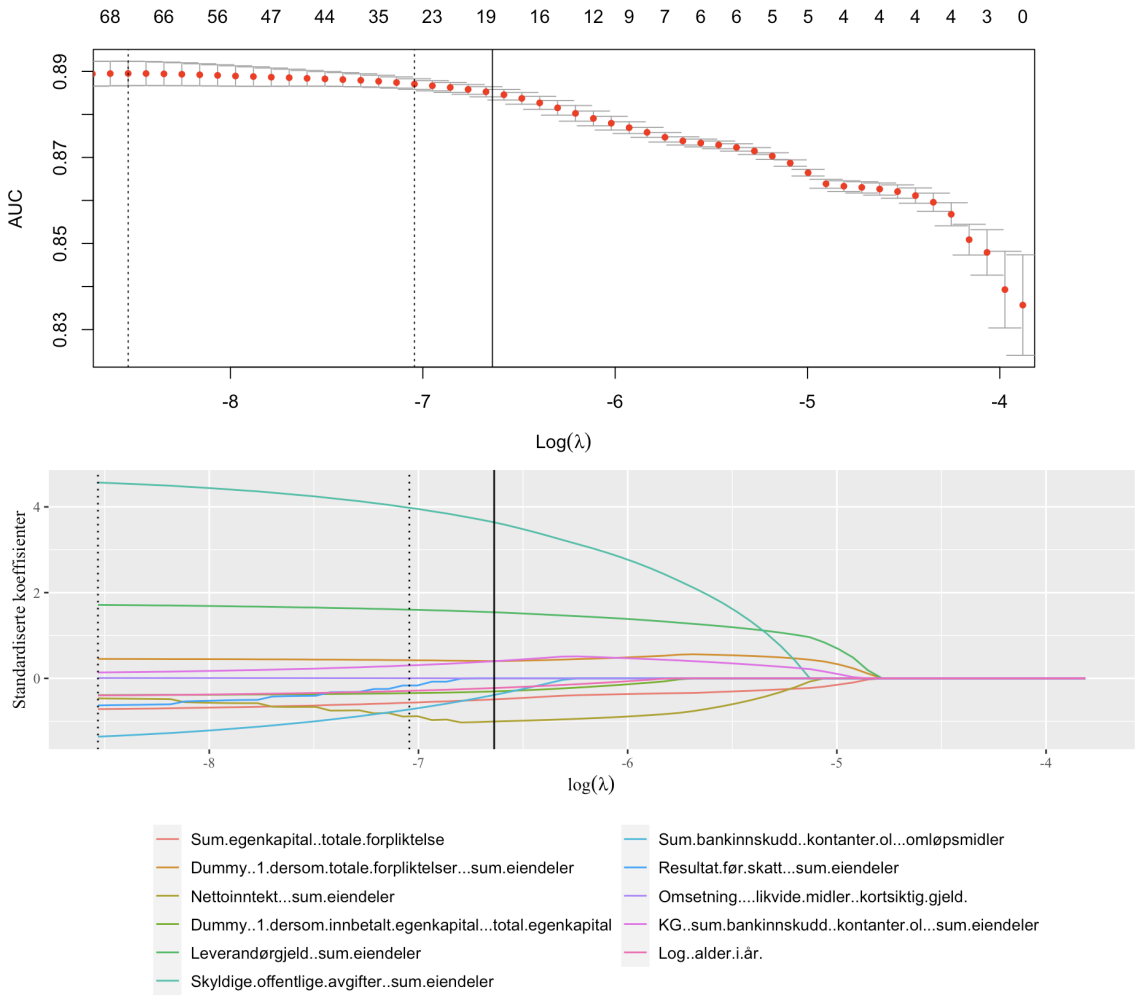
Periode 5. AUC og Standardiserte koeffisienter, over log(λ)-verdier. Den tykkeste linjen er lik log(λ) hvor vi tillater 1,5 standardfeil. Variablene til høyre for linjen velges av LASSO.



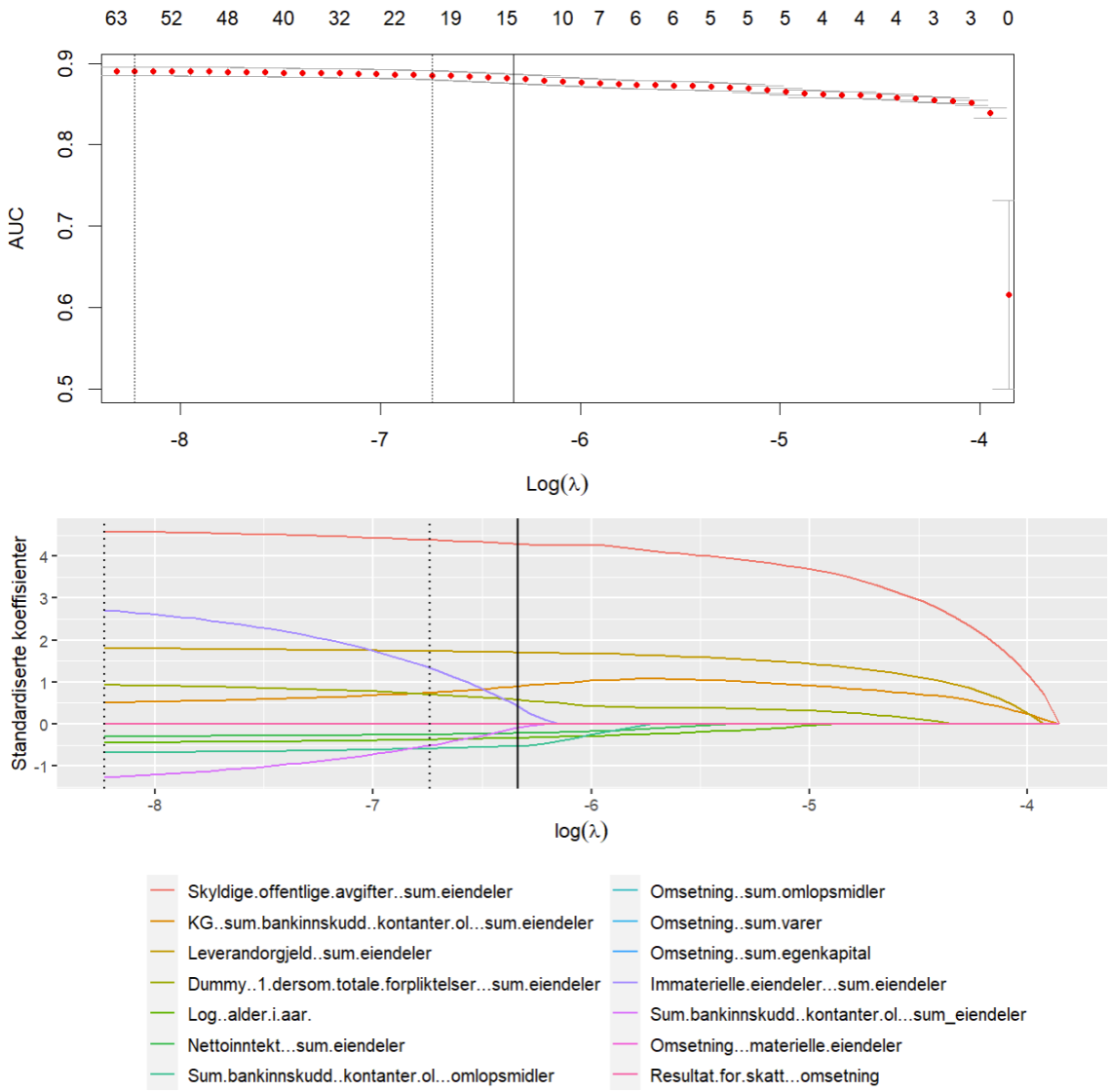
Periode 6. AUC og Standardiserte koeffisienter, over log(λ)-verdier. Den tykkeste linjen er lik log(λ) hvor vi tillater 1,5 standardfeil. Variablene til høyre for linjen velges av LASSO.



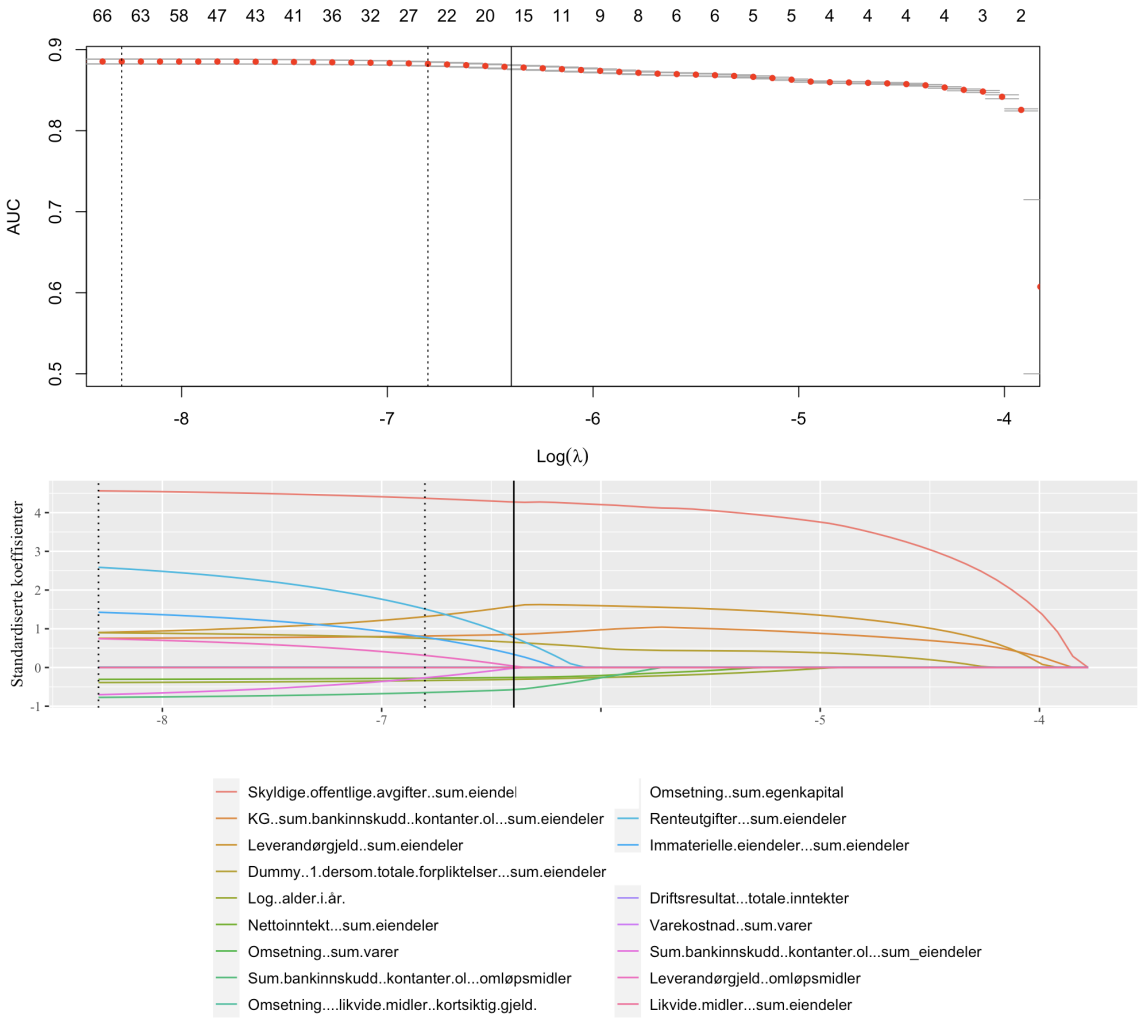
Periode 7. AUC og Standardiserte koeffisienter, over log(λ)-verdier. Den tykkeste linjen er lik log(λ) hvor vi tillater 1,5 standardfeil. Variablene til høyre for linjen velges av LASSO.



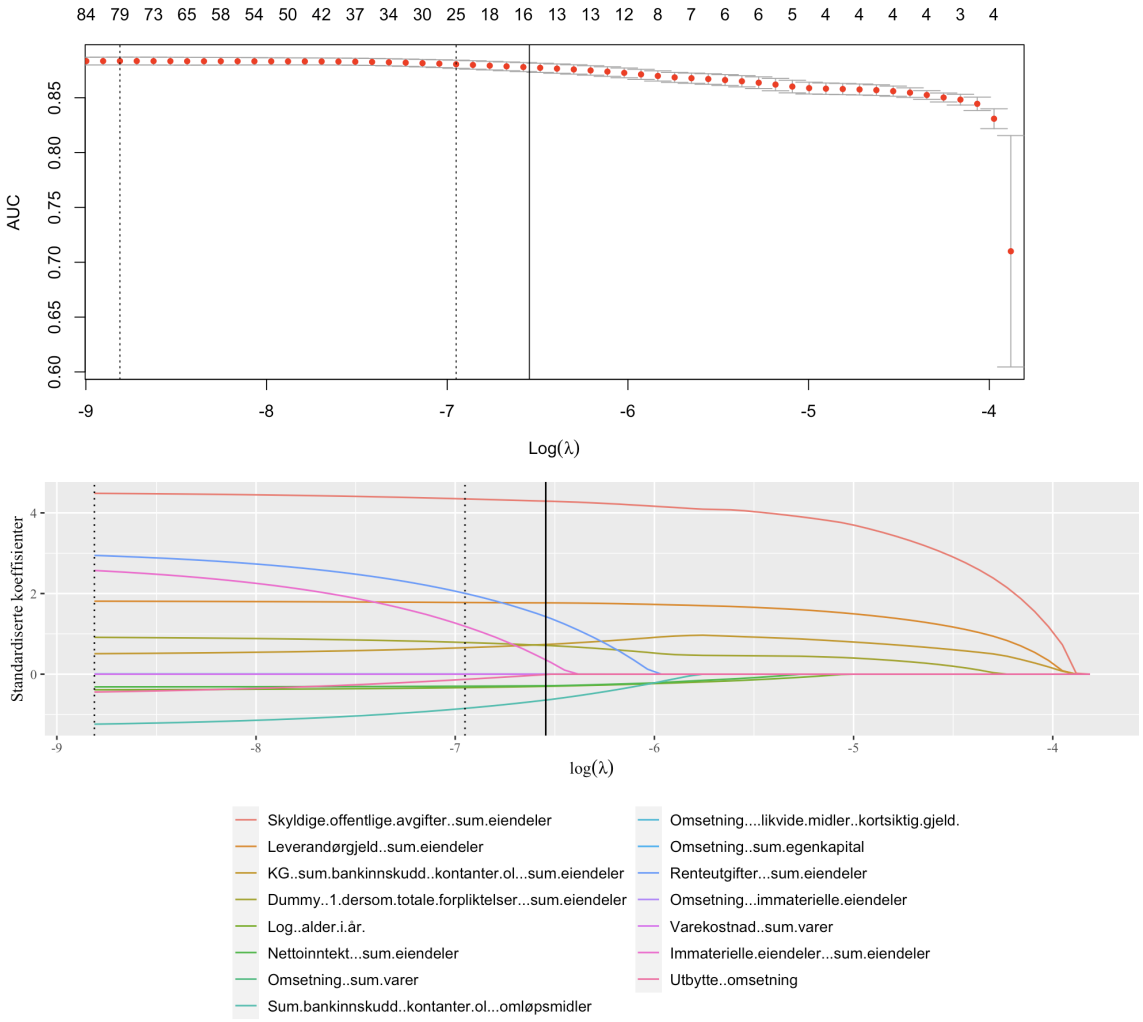
Periode 8. AUC og Standardiserte koeffisienter, over log(λ)-verdier. Den tykkeste linjen er lik log(λ) hvor vi tillater 1,5 standardfeil. Variablene til høyre for linjen velges av LASSO.



Periode 9. AUC og Standardiserte koeffisienter, over log(λ)-verdier. Den tykkeste linjen er lik log(λ) hvor vi tillater 1,5 standardfeil. Variablene til høyre for linjen velges av LASSO.



Periode 10. AUC og Standardiserte koeffisienter, over log(λ)-verdier. Den tykkeste linjen er lik log(λ) hvor vi tillater 1,5 standardfeil. Variablene til høyre for linjen velges av LASSO.



Periode 11. AUC og Standardiserte koeffisienter, over log(λ)-verdier. Den tykkeste linjen er lik log(λ) hvor vi tillater 1,5 standardfeil. Variablene til høyre for linjen velges av LASSO.

II Resultater fra variabelsettene

Resultater periode 1. Oversikt over AUC og Brier Score til de logistiske regresjonsmodellene med variablene fra Altman og Sabato (2007) og variablene i SEBRA-modellen.

	Altman og Sabato (2007)	**SEBRA**
AUC treningssett	0,809	0,859
Brier treningssett	0,019	0,018
AUC testsett	0,794	0,848
Brier testsett	0,017	0,016

Resultater periode 2. Oversikt over AUC og Brier Score til de logistiske regresjonsmodellene med variablene fra Altman og Sabato (2007) og variablene i SEBRA-modellen.

	Altman og Sabato (2007)	**SEBRA**
AUC treningssett	0,803	0,859
Brier treningssett	0,019	0,018
AUC testsett	0,793	0,852
Brier testsett	0,017	0,016

Resultater periode 3. Oversikt over AUC og Brier Score til de logistiske regresjonsmodellene med variablene fra Altman og Sabato (2007) og variablene i SEBRA-modellen.

	Altman og Sabato (2007)	**SEBRA**
AUC treningssett	0,800	0,859
Brier treningssett	0,018	0,017
AUC testsett	0,774	0,854
Brier testsett	0,016	0,016

Resultater periode 4. Oversikt over AUC og Brier Score til de logistiske regresjonsmodellene med variablene fra Altman og Sabato (2007) og variablene i SEBRA-modellen.

	Altman og Sabato (2007)	**SEBRA**
AUC treningssett	0,793	0,856
Brier treningssett	0,017	0,016
AUC testsett	0,784	0,863
Brier testsett	0,016	0,015

Resultater periode 5. Oversikt over AUC og Brier Score til de logistiske regresjonsmodellene med variablene fra Altman og Sabato (2007) og variablene i SEBRA-modellen.

	Altman og Sabato (2007)	**SEBRA**
AUC treningssett	0,785	0,854
Brier treningssett	0,016	0,016
AUC testsett	0,767	0,879
Brier testsett	0,015	0,014

Resultater periode 6. Oversikt over AUC og Brier Score til de logistiske regresjonsmodellene med variablene fra Altman og Sabato (2007) og variablene i SEBRA-modellen.

	Altman og Sabato (2007)	**SEBRA**
AUC treningssett	0,779	0,862
Brier treningssett	0,016	0,015
AUC testsett	0,752	0,866
Brier testsett	0,015	0,014

Resultater periode 7. Oversikt over AUC og Brier Score til de logistiske regresjonsmodellene med variablene fra Altman og Sabato (2007) og variablene i SEBRA-modellen.

	Altman og Sabato (2007)	**SEBRA**
AUC treningssett	0,773	0,866
Brier treningssett	0,015	0,015
AUC testsett	0,763	0,857
Brier testsett	0,015	0,015

Resultater periode 8. Oversikt over AUC og Brier Score til de logistiske regresjonsmodellene med variablene fra Altman og Sabato (2007) og variablene i SEBRA-modellen.

	Altman og Sabato (2007)	**SEBRA**
AUC treningssett	0,770	0,868
Brier treningssett	0,015	0,015
AUC testsett	0,772	0,864
Brier testsett	0,016	0,015

Resultater periode 9. Oversikt over AUC og Brier Score til de logistiske regresjonsmodellene med variablene fra Altman og Sabato (2007) og variablene i SEBRA-modellen.

	Altman og Sabato (2007)	**SEBRA**
AUC treningssett	0,768	0,868
Brier treningssett	0,015	0,014
AUC testsett	0,778	0,845
Brier testsett	0,017	0,027

Resultater periode 10. Oversikt over AUC og Brier Score til de logistiske regresjonsmodellene med variablene fra Altman og Sabato (2007) og variablene i SEBRA-modellen.

	Altman og Sabato (2007)	**SEBRA**
AUC treningssett	0,769	0,861
Brier treningssett	0,016	0,015
AUC testsett	0,775	0,863
Brier testsett	0,012	0,011

Resultater periode 11. Oversikt over AUC og Brier Score til de logistiske regresjonsmodellene med variablene fra Altman og Sabato (2007) og variablene i SEBRA-modellen.

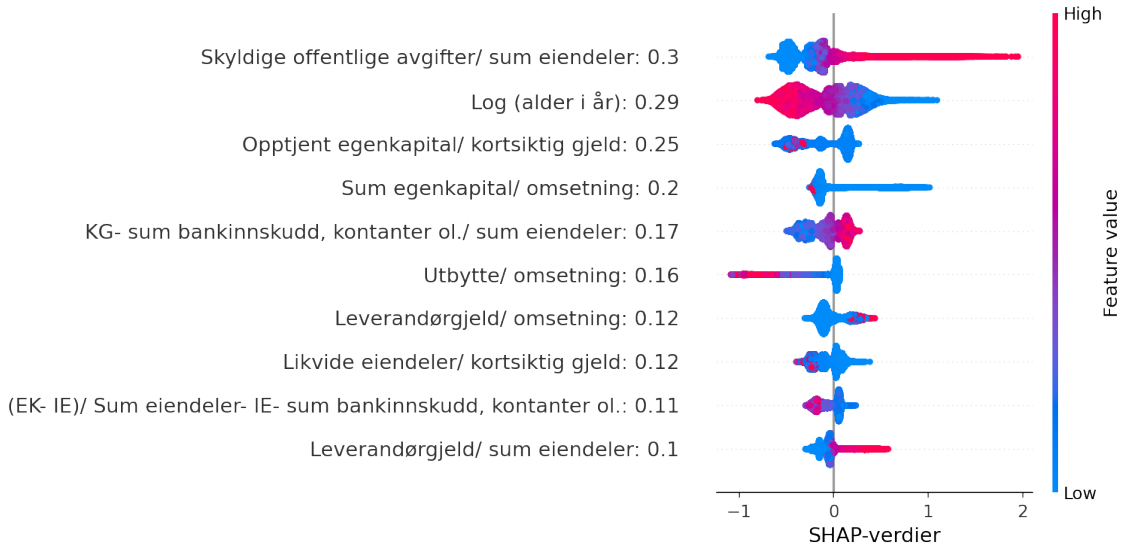
	Altman og Sabato (2007)	**SEBRA**
AUC treningssett	0,775	0,863
Brier treningssett	0,015	0,014
AUC testsett	0,808	0,905
Brier testsett	0,006	0,006

III Hyperparametere valgt ved optimalisering

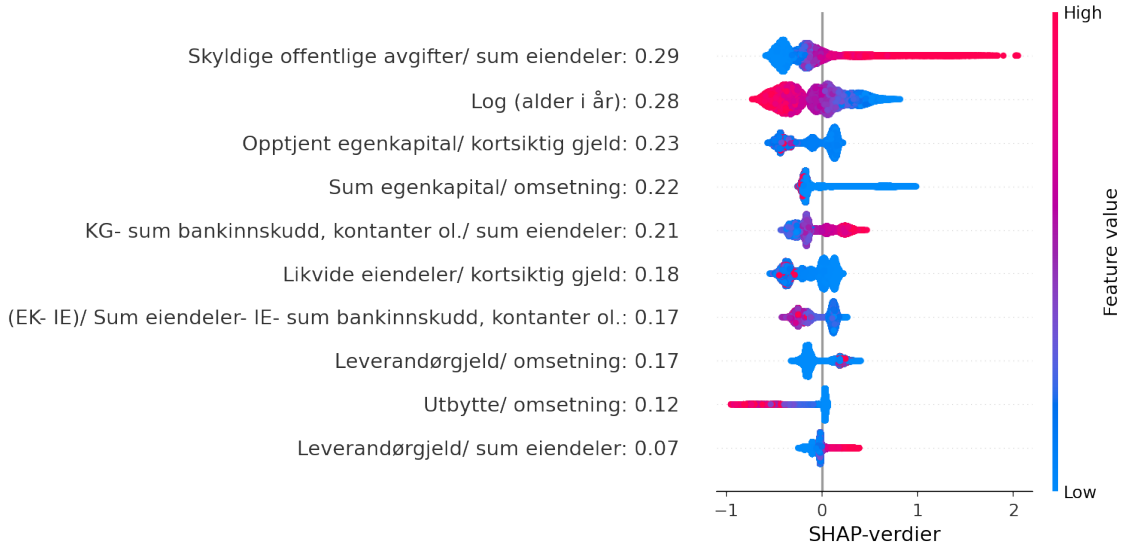
Hyperparametere valgt ved optimalisering for hver periode i XGBoost-modellene.

	Gamma	Learning rate	Max depth	Reg lambda	Subsample
Periode 1	0,0	0,1	3	1	0,50
Periode 2	0,1	0,1	3	3	0,75
Periode 3	0,0	0,1	2	1	0,50
Periode 4	0,2	0,1	3	3	0,75
Periode 5	0,4	0,1	3	3	0,50
Periode 6	0,4	0,1	3	3	0,50
Periode 7	0,4	0,2	3	1	0,75
Periode 8	0,0	0,1	3	1	0,50
Periode 9	0,4	0,1	3	3	0,50
Periode 10	0,0	0,1	3	1	0,50
Periode 11	0,0	0,2	3	3	0,75

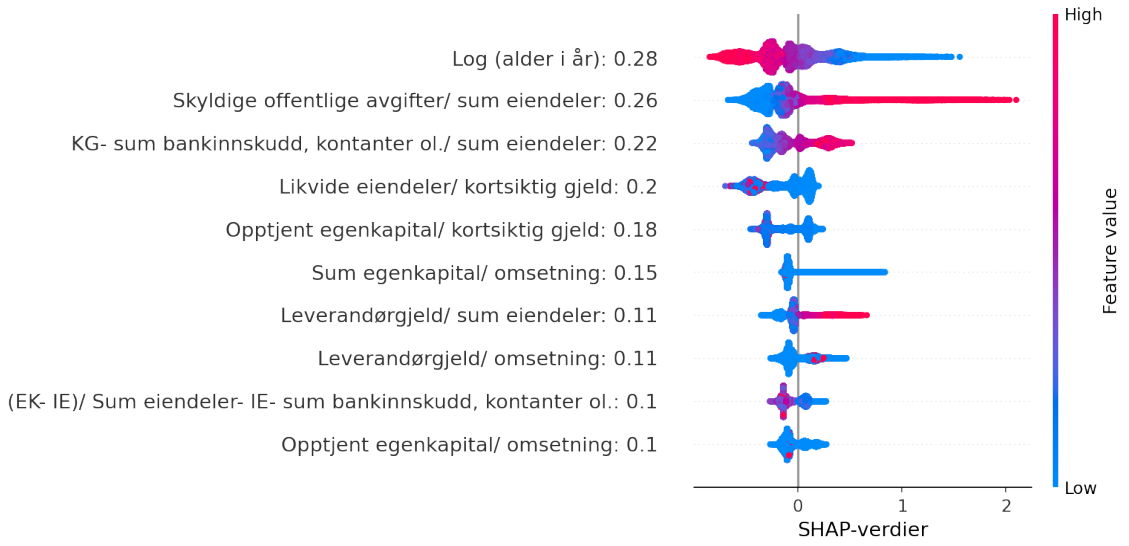
IV Bee swarm-plot



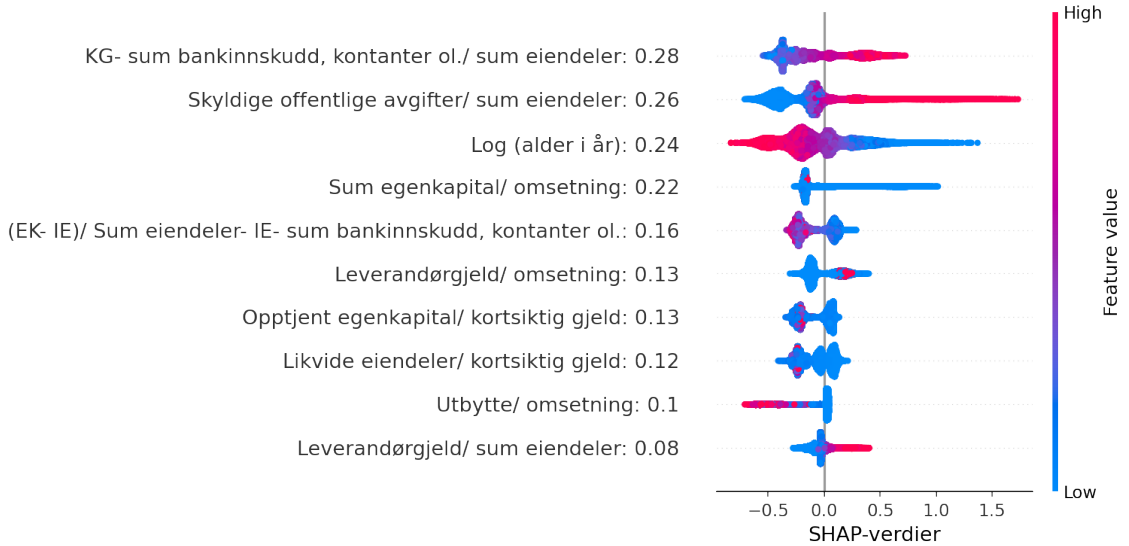
Periode 1. Bee swarm plot som viser SHAP-verdier for de 10 variablene med høyest SHAP-verdi i periode 1. Hver prikk i plottet representerer en observasjon, hvor fargen angir den opprinnelige verdien til variabelen.



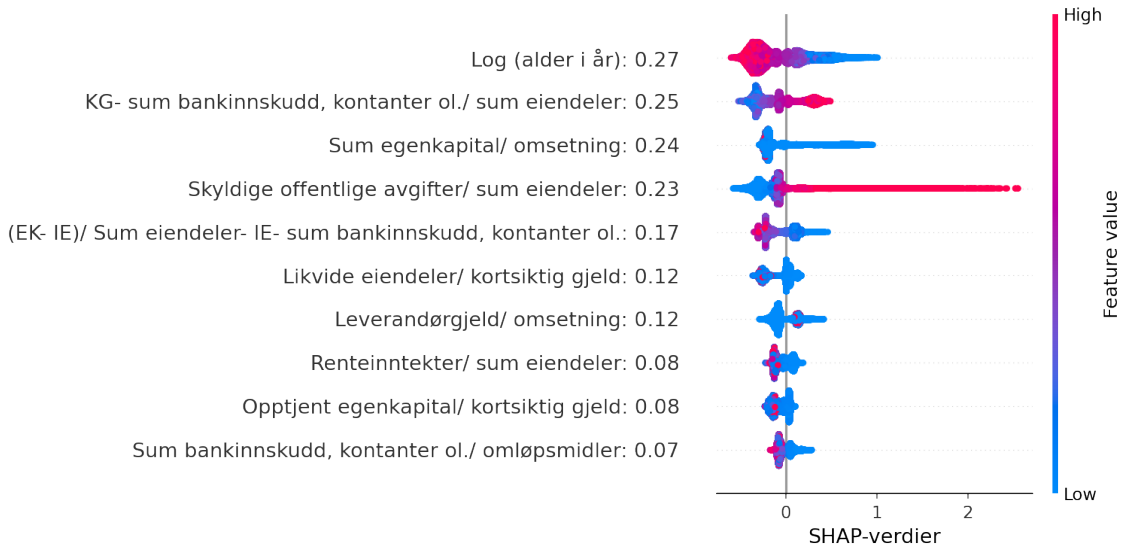
Periode 2. Bee swarm plot som viser SHAP-verdier for de 10 variablene med høyest SHAP-verdi i periode 2. Hver prikk i plottet representerer en observasjon, hvor fargen angir den opprinnelige verdien til variabelen.



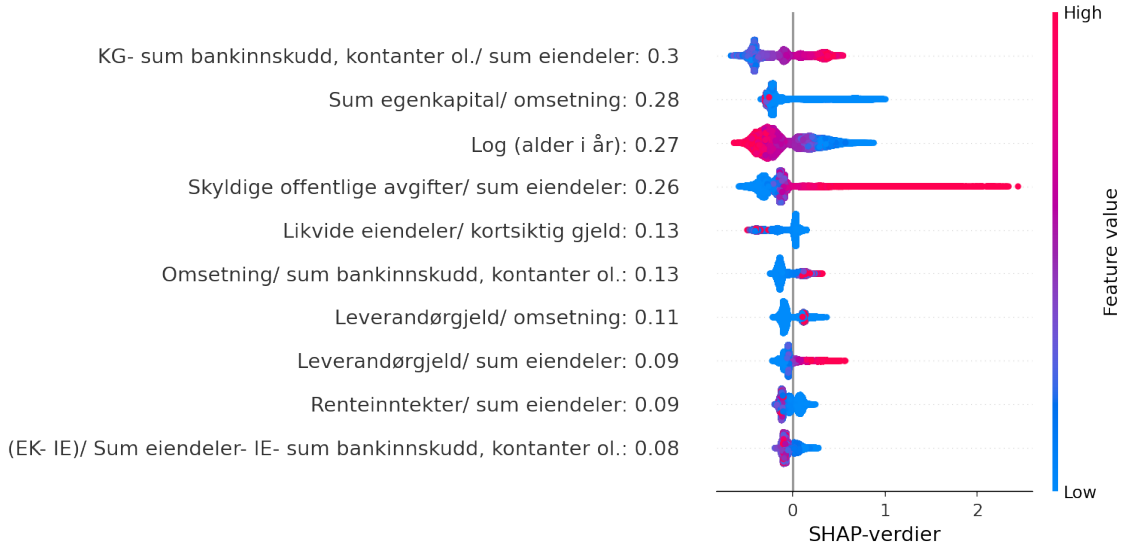
Periode 3. Bee swarm plot som viser SHAP-verdier for de 10 variablene med høyest SHAP-verdi i periode 3. Hver prikk i plottet representerer en observasjon, hvor fargen angir den opprinnelige verdien til variabelen.



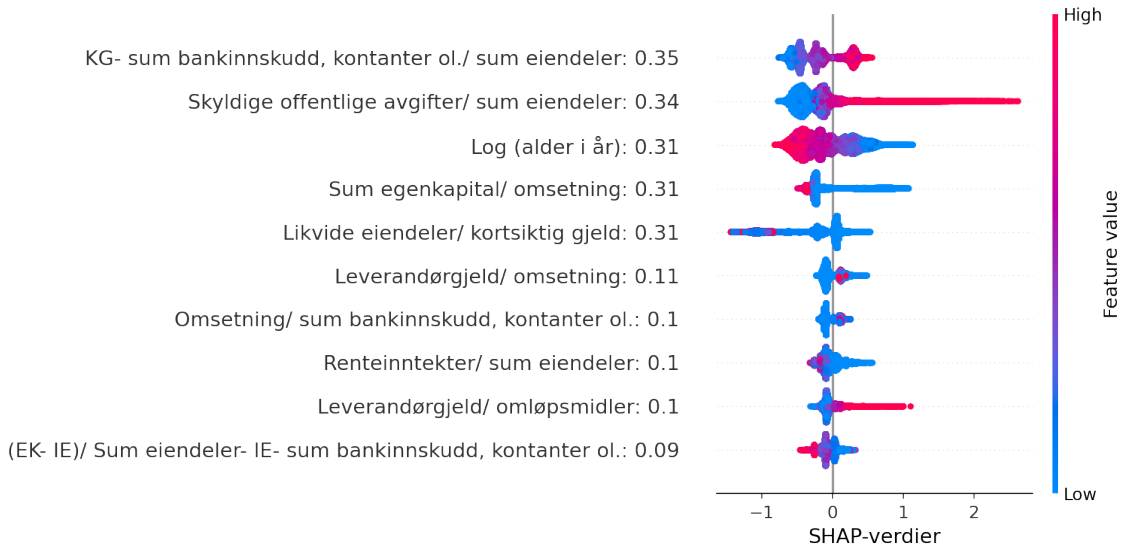
Periode 4. Bee swarm plot som viser SHAP-verdier for de 10 variablene med høyest SHAP-verdi i periode 4. Hver prikk i plottet representerer en observasjon, hvor fargen angir den opprinnelige verdien til variabelen.



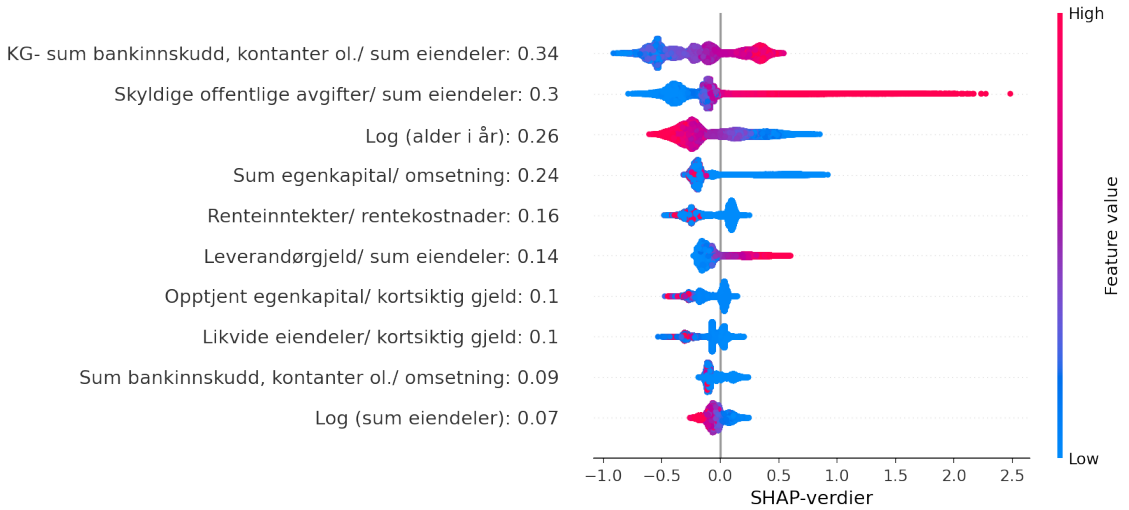
Periode 5. Bee swarm plot som viser SHAP-verdier for de 10 variablene med høyest SHAP-verdi i periode 5. Hver prikk i plottet representerer en observasjon, hvor fargen angir den opprinnelige verdien til variabelen.



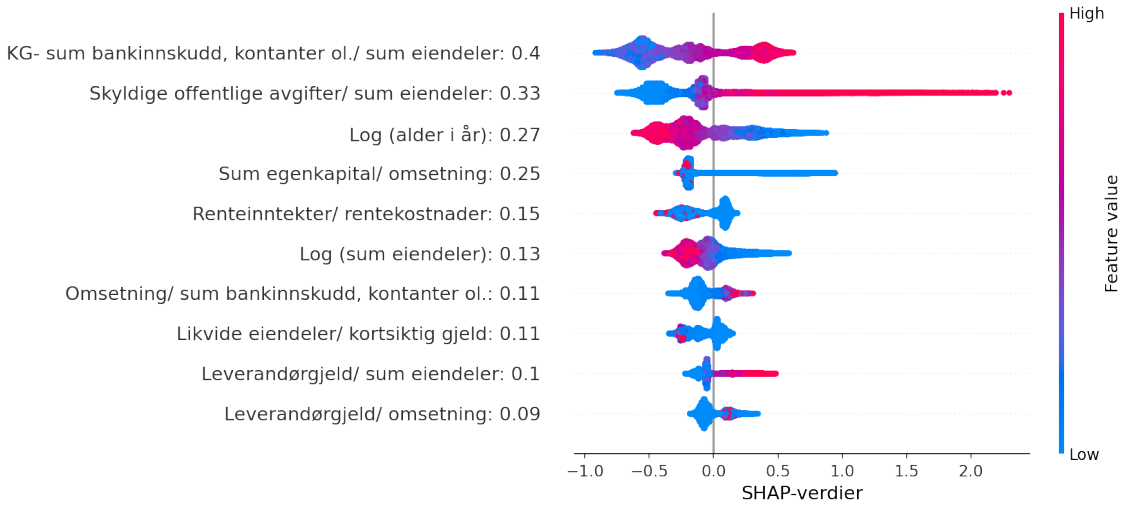
Periode 6. Bee swarm plot som viser SHAP-verdier for de 10 variablene med høyest SHAP-verdi i periode 6. Hver prikk i plottet representerer en observasjon, hvor fargen angir den opprinnelige verdien til variabelen.



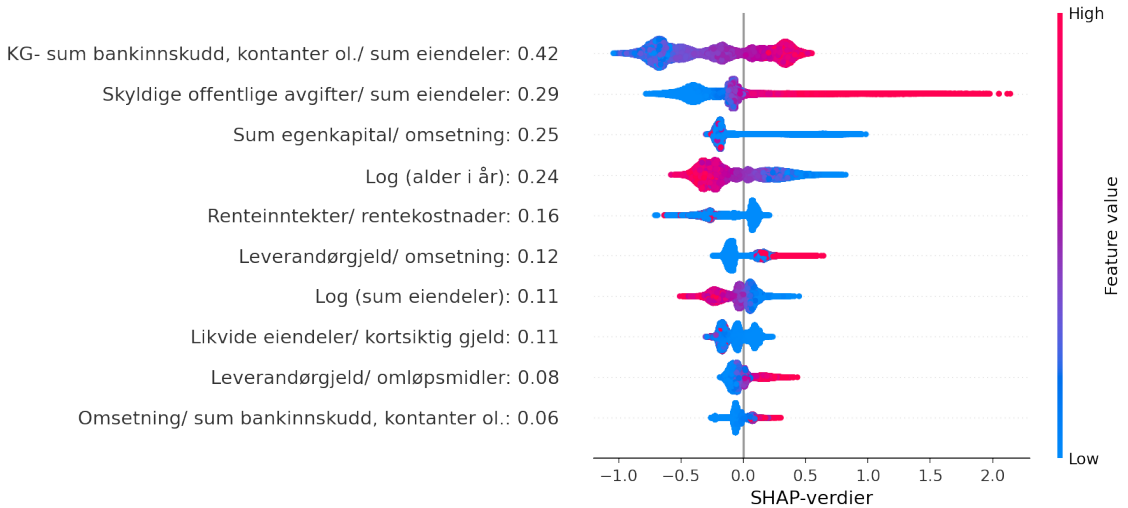
Periode 7. Bee swarm plot som viser SHAP-verdier for de 10 variablene med høyest SHAP-verdi i periode 7. Hver prikk i plottet representerer en observasjon, hvor fargen angir den opprinnelige verdien til variabelen.



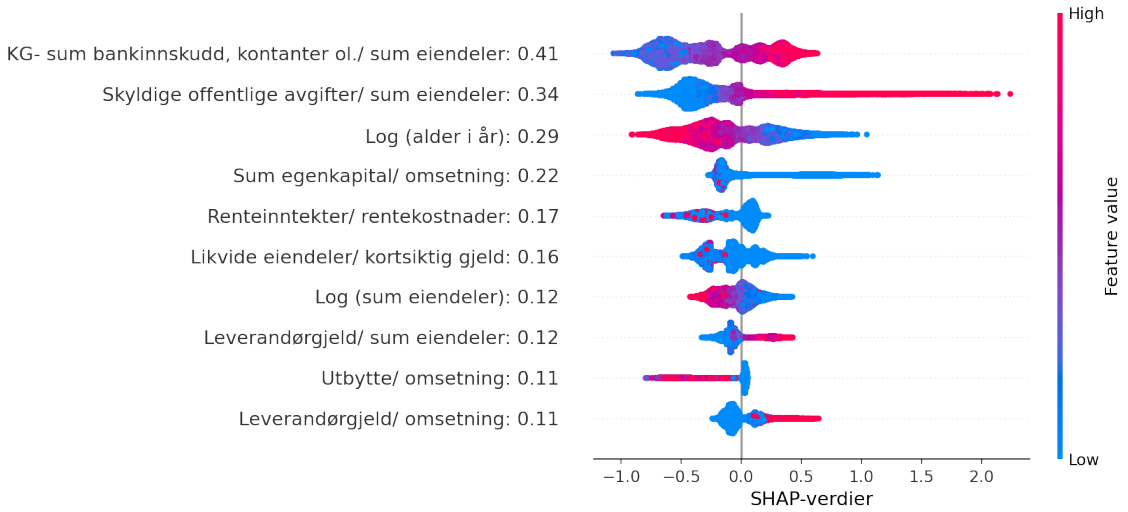
Periode 8. Bee swarm plot som viser SHAP-verdier for de 10 variablene med høyest SHAP-verdi i periode 8. Hver prikk i plottet representerer en observasjon, hvor fargen angir den opprinnelige verdien til variabelen.



Periode 9. Bee swarm plot som viser SHAP-verdier for de 10 variablene med høyest SHAP-verdi i periode 9. Hver prikk i plottet representerer en observasjon, hvor fargen angir den opprinnelige verdien til variabelen.

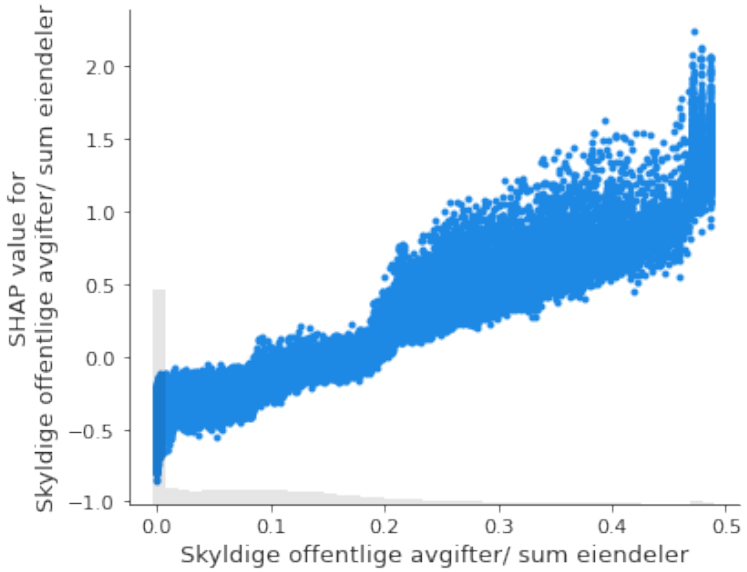


Periode 10. Bee swarm plot som viser SHAP-verdier for de 10 variablene med høyest SHAP-verdi i periode 10. Hver prikk i plottet representerer en observasjon, hvor fargen angir den opprinnelige verdien til variabelen.

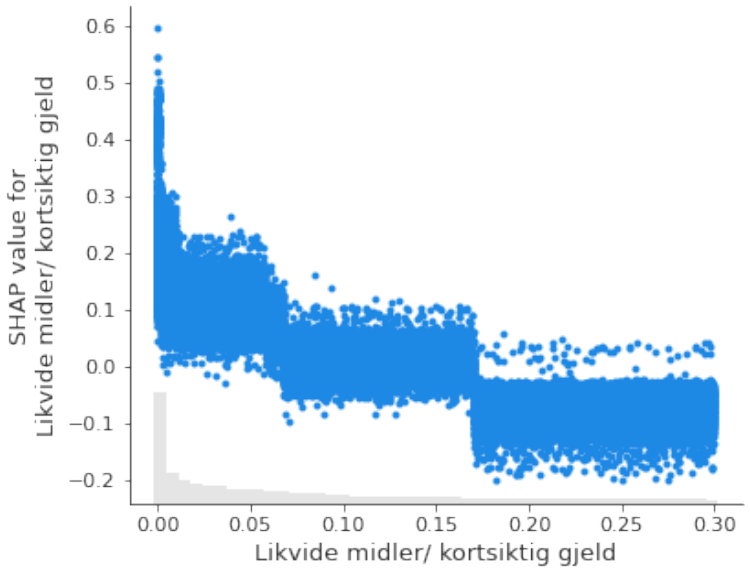


Periode 11. Bee swarm plot som viser SHAP-verdier for de 10 variablene med høyest SHAP-verdi i periode 11. Hver prikk i plottet representerer en observasjon, hvor fargen angir den opprinnelige verdien til variabelen.

V Dependence scatter plot



Dependence scatter plot for variabelen *skyldige offentlige avgifter/ sum eiendeler*. Y-aksen viser SHAP-verdier til variabelen for ulike verdier av variabelen. XGBoost-modellen er utviklet basert på data for bygg- og anleggsbransjen for periode 11.



Dependence scatter plot for variabelen *Likvide midler/ kortsiktig gjeld*. Y-aksen viser SHAP-verdier til variabelen for ulike verdier av variabelen. XGBoost-modellen er utviklet basert på data for bygg- og anleggsbransjen for periode 11.

VI Variabelsett 160 variabler

Figuren viser variabelsett med 160 variabler, som benyttes i oppgaven til variabelseleksjon. De 155 første variablene er hentet fra Paraschiv mfl. (2021). De fem siste er hentet fra Statistisk sentralbyrå (2022a, 2023a, 2023d)

Variabelnavn
Sum varer + kundefordringer/ sum egenkapital
(Langsiktig gjeld+ sum egenkapital)/ anleggsmidler
Kundefordringer/ omsetning
Likvide midler/ kortsiktig gjeld
(Likvide midler/ kortsiktig gjeld)* (Totalresultat/ rentekostnader)
Årsresultat/ sum egenkapital
Driftsresultat/ totale forpliktelser
Sum egenkapital/ sum forpliktelser
Sum bankinnskudd, kontanter ol./ sysselsatt kapital
Sum bankinnskudd, kontanter ol./ omsetning
Sum bankinnskudd, kontanter ol. / kortsiktig gjeld
Sum bankinnskudd, kontanter ol./ sum eiendeler
Omsetning/ omløpsmidler
Omløpsmidler/ sum egenkapital
Omløpsmidler/ omsetning
Omløpsmidler/ sum eiendeler
Kortsiktig gjeld/ omløpsmidler
Kortsiktig gjeld/ sum egenkapital
Kortsiktig gjeld/ totale forpliktelser
Kortsiktig gjeld/ omsetning
Totale forpliktelser/ sum eiendeler
Kundefordringer/ leverandørgjeld
Driftsresultat/ (driftsresultat- rentekostnader)
Totalresultat/ sum eiendeler
Driftsresultat/ rentekostnader
Skattekostnad/ ordinært resultat før skatt
Sum egenkapital/ sum eiendeler
Sum egenkapital/ totale forpliktelse
Omsetning/ sum egenkapital

Figuren viser variabelsett med 160 variabler, som benyttes i oppgaven til variabelseleksjon. De 155 første variablene er hentet fra Paraschiv mfl. (2021). De fem siste er hentet fra Statistisk sentralbyrå (2022a, 2023a, 2023d)

Variabelnavn
Resultat før skatt i prosent av sysselsatt kapital
Finansielle utgifter / omsetning
Totalresultat / omsetning
Omsetning / anleggsmidler
Anleggsmidler / sum eiendeler
Anleggsmidler / total egenkapital
Immaterielle eiendeler / sum eiendeler
Renteutgifter / totale inntekter
Rentebærende gjeld / total egenkapital
Sum varer / kortsiktig gjeld
Sum varer / arbeidskapital
Investeringsomsetning (omsetning / (total egenkapital + total gjeld))
Totale forpliktelser / total egenkapital
Langsiktig gjeld / omløpsmidler
Nettoinntekt / innbetalt kapital
Nettoinntekt / omsetning
(Totale inntekter - omsetning) / totale inntekter
Total egenkapital / anleggsmidler
Total egenkapital / omsetning
Intervall uten kreditt
Dummy: 1 dersom totale forpliktelser > sum eiendeler
Driftsutgifter / omsetning
Sum bankinnskudd, kontanter ol. / totale forpliktelser
Driftsresultat / totale inntekter
Driftsresultat / innskutt kapital
Driftsmiddel / sum eiendeler
Personalkostnader / merverdi
Resultat før skatt / innbetalt kapital
Nettoinntekt / totale inntekter
Overskudd / netto arbeidskapital
Likvide midler / omsetning

Figuren viser variabelsett med 160 variabler, som benyttes i oppgaven til variabelseleksjon. De 155 første variablene er hentet fra Paraschiv mfl. (2021). De fem siste er hentet fra Statistisk sentralbyrå (2022a, 2023a, 2023d)

Variabelnavn
Likvide midler / sum eiendeler
Resultat etter skatt og rentekostnader / netto sysselsatt kapital
Kortsiktig gjeld / resultat før skatt og rentekostnader
Opptjent egenkapital / omsetning
Opptjent egenkapital / sum eiendeler
Avkastning på gjeld (inntjening / total gjeld)
Nettoinntekt / sum eiendeler
Totale inntekter / anleggsmidler
Totale inntekter / sum eiendeler
Totale inntekter / netto arbeidskapital
Omsetning / sum eiendeler
Sum eiendeler / totale inntekter
Totale utgifter / eiendeler
Totale inntekter / totale utgifter
Arbeidskapital / kortsiktig gjeld
Arbeidskapital / omsetning
Arbeidskapital / sum eiendeler
Arbeidskapital / total egenkapital
Dummy: 1 dersom innbetalt egenkapital < total egenkapital
Arbeidskapital / totale inntekter
Leverandørgjeld / sum eiendeler
Offentlige betalbare skatter / totale eiendeler
Totalkapital / total gjeld
(Driftskostnader - lønn) / sum eiendeler
(Egenkapital + totale inntekter) / sum eiendeler
Omsetning / arbeidskapital
Sum bankinnskudd, kontanter ol. / omløpsmidler
Varekostnad / sum varer
Kostnad for solgte varer / omsetning
(Omløpsmidler - sum bankinnskudd, kontanter ol.) / sum eiendeler
Omløpsmidler / felles aksjonærkapital

Figuren viser variabelsett med 160 variabler, som benyttes i oppgaven til variabelseleksjon. De 155 første variablene er hentet fra Paraschiv mfl. (2021). De fem siste er hentet fra Statistisk sentralbyrå (2022a, 2023a, 2023d)

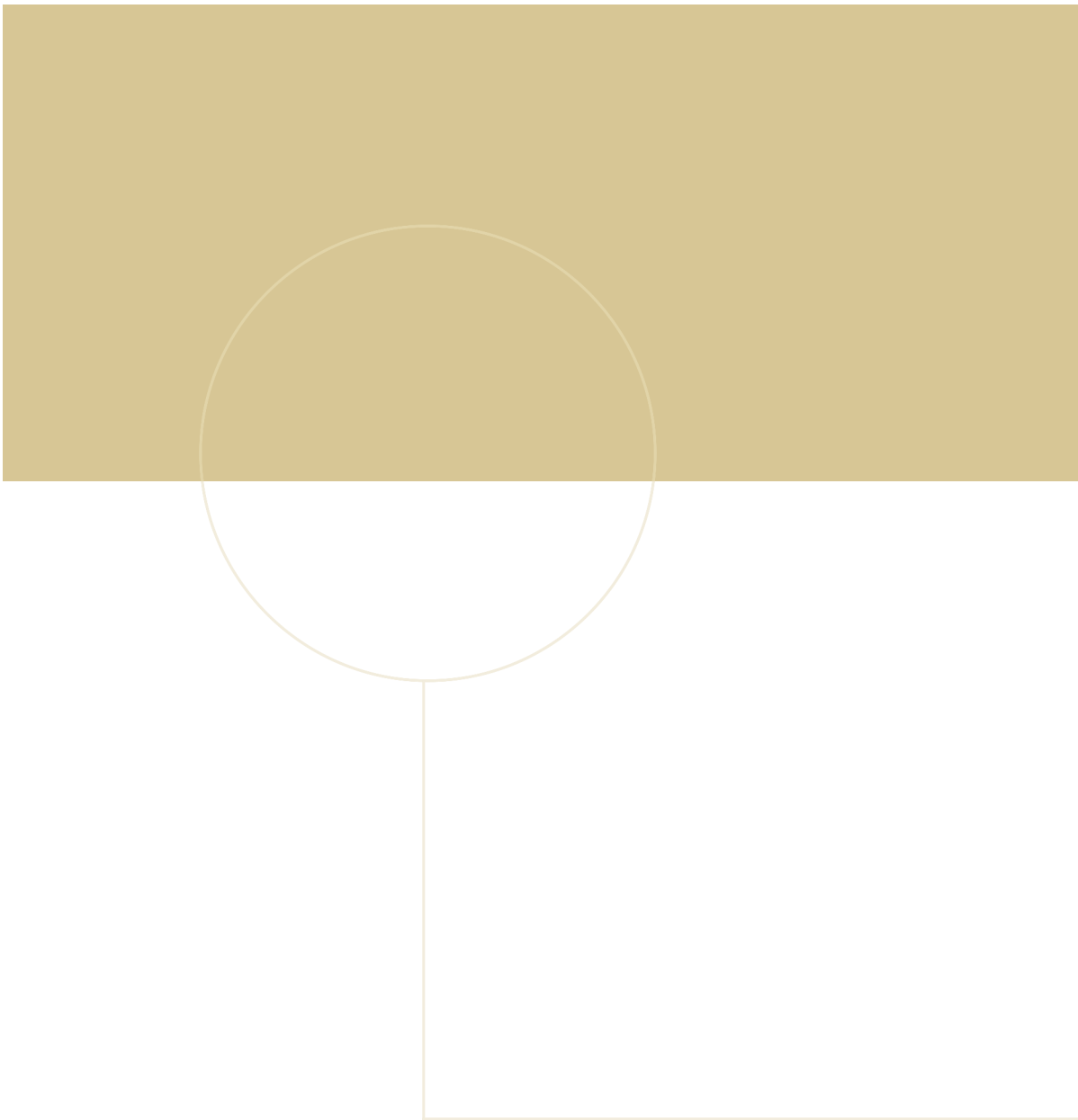
Variabelnavn
Kortsiktig gjeld / sum eiendeler
Utbytte / nettoinntekt
Arbeidskapital / langsiktig gjeld
Arbeidskapital / driftsutgifter
Totalresultat/ totale materielle eiendeler
Finansielle utgifter / salg
Anleggmidler / (innbetalt kapital + langsiktig gjeld)
(Omsetning - varekostnad) / omsetning
Inntektsgiring
Immaterielle eiendeler / omsetning
Renteutgifter / totale forpliktelser
Renteutgifter / totale utgifter
Renteinntekter / renteutgifter
Renteinntekter / sum eiendeler
Sum varer / varekostnader
Sum varer / omløpsmidler
Sum varer / omsetning
Langiktig gjeld / total egenkapital
Langsiktig gjeld / sum eiendeler
Omsetning / materielle eiendeler
Nettoinntekt / bruttofortjeneste
Nettoinntekt / (totall gjeld + innbetalt kapital)
Likvide midler / sum varer
Total egenkapital / (total egenkapital + langsiktig gjeld)
Ikke-renteutgifter / driftsresultat
Totale inntekter / omsetning
Resultat før skatt / total egenkapital
Resultat før skatt / ordinære utgifter
Resultat før skatt / omsetning
Resultat før skatt / sum eiendeler
Egenkapital / sum eiendeler

Figuren viser variabelsett med 160 variabler, som benyttes i oppgaven til variabelseleksjon. De 155 første variablene er hentet fra Paraschiv mfl. (2021). De fem siste er hentet fra Statistisk sentralbyrå (2022a, 2023a, 2023d)

Variabelnavn
Leverandørgjeld / kortsiktig gjeld
Leverandørgjeld / sum varer
Opptjent egenkapital/ sum varer
Opptjent egenkapital / materielle eiendeler
Avkastning på sysselsatt kapital
Avkastning på netto anleggsmidler
Lønn / sum eiendeler
Omsetning / sum bankinnskudd kontanter ol.
Omsetning / sum varer
Omsetning / kundefordringer
Omsetning / totale materialle eiendeler
Rentebærende gjeld / totale forpliktelser
Lønnskostnader/ (Omsetning - driftsresultat + finanskostnader)
(Kortsiktige eiendeler - totale forpliktelser) / sum eiendeler
Soliditetsgrad
Omsetning / egenkapital
(Totale inntekter + renteinntekter) / totale utgifter
Renteutgifter / sum eiendeler
Driftsutgifter / sum eiendeler
Omsetning / (likvide midler/ kortsiktig gjeld)
Driftsresultat / sum eiendeler
Driftsresultat / sum eiendeler
Driftsresultat / salg
(Kortsiktig gjeld - sum bankinnskudd, kontanter ol) / sum eiendeler
Leverandørgjeld / omsetning
Opptjent egenkapital / kortsiktig gjeld
(Total egenkapital - immaterielle eiendeler) / (totale eiendeler - immaterielle eiendeler - sum bankinnskudd, kontanter ol.)
Totalkapital / rentekostnader
Kundefordringer / totale forpliktelser
Resultat før skatt / kortsiktig gjeld

Figuren viser variabelsett med 160 variabler, som benyttes i oppgaven til variabelseleksjon. De 155 første variablene er hentet fra Paraschiv mfl. (2021). De fem siste er hentet fra Statistisk sentralbyrå (2022a, 2023a, 2023d)

Variabelnavn
Omløpsmidler / totale forpliktelser
Log (alder i år)
Log (sum eiendeler)
BNP
Styringsrente ved årsslutt
Gjennomsnittlig styringsrente
Styringsrente ved årsslutt - styringsrente ved starten av året
Gjennomsnittlig produksjonsindeks



NTNU

Kunnskap for en bedre verden