Masteroppgave

NTNU
Norges teknisk-naturvitenskapelige universitet
Fakultet for økonomi
NTNU Handelshøyskolen

Sara Strandkleiv Øverland
Olav Haugen

Lønnsomhetsdrivere for SMB i den norske bygg- og anleggsbransjen

Masteroppgave i Business Analytics og Økonomistyring
Veileder: Tor-Eirik Olsen, Ole Jakob Sønstebø og Ranik Raaen Wahlstrøm

Mai 2024

NTNU
Kunnskap for en bedre verden

Sara Strandkleiv Øverland
Olav Haugen

Lønnsomhetsdrivere for SMB i den norske bygg- og anleggsbransjen

Masteroppgave i Business Analytics og Økonomistyring
Veileder: Tor-Eirik Olsen, Ole Jakob Sønstebø og Ranik Raaen Wahlstrøm
Mai 2024

Norges teknisk-naturvitenskapelige universitet
Fakultet for økonomi
NTNU Handelshøyskolen

NTNU
Kunnskap for en bedre verden

Forord

Denne masteroppgaven er det avsluttende arbeidet for mastergraden vår innen økonomi og administrasjon ved NTNU Handelshøyskolen. Masteroppgaven er skrevet innen hovedprofilene Business Analytics og Økonomistyring. Dette mener vi har vært til fordel for oppgaven, fordi det komplementerer hverandres kunnskap. Oppgaven utgjør 30 studiepoeng, og er skrevet våren 2024. Prosessen med å skrive masteroppgaven har vært både lang og utfordrende, men samtidig også svært lærerik. Det har vært spennende å utforme sin egen oppgave, og det har gitt oss god erfaring med å gjennomføre en vitenskapelig studie.

Vi ønsker å rette en stor takk til veilederne våre Tor-Eirik Olsen, Ole Jakob Sønstebø og Ranik Raaen Wahlstrøm for bistand og verdifulle tilbakemeldinger. Vi ønsker å gi en ekstra takk til Ranik for tilgang til datasettet som har utgjort grunnlaget for oppgaven. Videre vil vi også rette en stor takk til våre medstudenter som har gitt gode råd og motivasjon gjennom hele prosessen. Avslutningsvis, ønsker vi å takke Martin Røstvoll og Karl Henrik Skulstad som har delt masterom med oss, og bidratt til å skape et godt arbeidsmiljø med god stemning og mye latter.

Innholdet i denne oppgaven står for forfatterenes regning.

Trondheim, 2024

Sara Strandkleiv Øverland

Olav Haugen

Denne siden er blank med hensikt

Sammendrag

Bygg- og anleggsbransjen står overfor utfordringer knyttet til lønnsomhet, og fra 2021 til 2022 var det et fall i driftsmargin fra 5,1 % til 4,5 %. Denne oppgaven bidrar til økt innsikt i hva som driver lønnsomheten i bransjen, med problemstillingen: *Hva er de viktigste lønnsomhetsdriverne for norske SMB i bygg- og anleggsbransjen?* For å svare på problemstillingen, ser vi på lønnsomheten fra tre forskjellige perspektiver, bedriftsspesifikke-, bransjespesifikke- og makroøkonomiske perspektiver.

Oppgaven er avgrenset til norske små- og mellomstore aksjeselskap, innen bygg og anlegg fra 2009 til 2022. Modellene trenes og testes på et datasett som omfatter alle ukonsoliderte årsregnskap gjennom en rolling window tilnærming. Vi benytter først denne dataen til å utvikle forklaringsmodeller i inneværende år ved ordinary least squares (OLS) og fixed effects. Lønnsomheten blir målt gjennom nøkkeltallet return on assets (ROA). I forkant benyttes Least Absolute Shrinkage and Selection Operator (LASSO), sammen med teori og tidligere forskning, til variabelseleksjon. Videre utvikles tre prediktive modeller som predikerer kontinuerlig ROA ett år frem i tid ved Extreme Gradient Boosting (XGBoost) og PyCaret, hvor sistnevnte er et automaskinlæringsbibliotek. XGBoost benyttes også til å predikere om bedrifter klassifiseres som «lønnsom» eller «ikke lønnsom». For å identifisere de viktigste variablene, og effekten av dem, benyttes Shapley Additive exPlanations (SHAP), Partial Dependence Plots (PDP) og Individual conditional expectation (ICE) kurver.

Funnene viser at de bedriftsspesifikke faktorene er mest egnet til å forklare lønnsomhet. Kapasitetsutnyttelse, målt ved nøkkeltallene produktivitet og lønnskostnad i % av driftsinntekt, er to av de viktigste lønnsomhetsdriverne, og høyere verdier er generelt assosiert med bedre lønnsomhet. Selskapsstørrelse, målt gjennom antall årsverk, omsetning og markedsandel, er mindre viktige for lønnsomhet, og viser at en økning i selskapsstørrelse ikke nødvendigvis fører til økt lønnsomhet. Erfaring i form av bedriftens alder viser at yngre bedrifter er mer volatile og kan være assosiert med lavere lønnsomhet, men for eldre bedrifter er ikke alder utslagsgivende for lønnsomheten. Lokalisering kan påvirke lønnsomheten, og beliggenhet i Oslo er assosiert med bedre lønnsomhet, mens beliggenhet i Hordaland og Møre om Romsdal har motsatt effekt. Kapitalens omløpshastighet blir ansett som en viktig lønnsomhetsdriver, og understreker viktigheten av effektiv kapitalutnyttelse. Når det gjelder likviditet, er likviditetsgrad 1 og 2, samt arbeidskapital, viktige forklaringsvariabler, og verdier over gjennomsnittet assosieres med lavere ROA. For soliditet er rentedekningsgrad, egenkapitalandel og gjeldsgrad viktige forklaringsvariabler, og forholdet mellom egenkapitalandel og gjeldsgrad tyder på at strategisk bruk av gjeld kan øke lønnsomhet. Funnene tyder på at bransjespesifikke- og makroøkonomiske variabler ikke er viktige lønnsomhetsdrivere.

Abstract

The construction industry is facing challenges with profitability, as seen by a decline in operating margin from 5.1 % in 2021 to 4.5 % in 2022. This thesis aims to contribute with insight in profitability drivers, hence the research question: *What are the key drivers of profitability for Norwegian SMEs in the construction industry?* To answer the research question, we explore firm-specific, industry-specific and macroeconomic factors.

The scope of the study is limited to Norwegian small and medium-sized enterprises (SMEs) in the construction industry from 2009 to 2022. The dataset consists of all unconsolidated annual financial statements with a rolling window approach. We use this data to construct explanatory models within the current year using ordinary least squares (OLS) regression and fixed effects. Profitability is assessed using return on assets (ROA). Prior to the regression models, Least Absolute Shrinkage and Selection Operator (LASSO), along with theory and previous literature, are applied for variable selection. In the next stage, we predict ROA a year ahead of time using Extreme Gradient Boosting (XGBoost) and PyCaret, an automated machine learning library. XGBoost is also used for classification, predicting whether a firm will be "profitable" or "non-profitable". To identify the key drivers of profitability and their impact on prediction, SHapley Additive exPlanations (SHAP), Partial Dependence Plots (PDP), and Individual Conditional Expectation (ICE) curves are utilized.

The results suggest that firm-specific factors are most suitable for explaining profitability. Capacity utilization, measured by productivity and wage costs as a % of operating revenues, are two key determinants of ROA, where higher values are associated with improved profitability. Scale, measured by the number of employees, annual turnover, and market share have less impact on profitability, suggesting that larger companies are not necessarily associated with higher ROA. Experience, reflected by the age of the firm, indicate that younger companies tend to achieve lower profitability, whereas age does not significantly influence profitability for established companies. Location also influences ROA, where companies situated in Oslo are associated with higher ROA, whereas Hordaland and Møre og Romsdal indicate the opposite effect. Asset turnover significantly impacts ROA, emphasizing the importance of efficient capital utilization. For measuring liquidity, current ratio, quick ratio and working capital are considered key profitability drivers, and values above average are associated with lower profitability. In terms of solvency, the interest coverage ratio, debt ratio and equity ratio are important variables, and the relationship between debt ratio and equity ratio suggests that increased profitability can be achieved by the strategic use of debt. Industry-specific and macroeconomic variables are not considered to be key drivers of profitability.

Innhold

Figurer

Tabeller

Forkortelser

AUC	Area Under the Curve
BLUE	Best Linear Unbiased Estimator
ICE	Individual Conditional Expectation
LASSO	Least Absolute Shrinkage and Selection Operator
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MdAPE	Median Absolute Percentage Error
MSE	Mean Squared Error
PDP	Partial Dependence Plot
RMSE	Root Mean Squared Error
ROC	The Receiver Operator Characteristics Curve
SHAP	SHapley Additive exPlanations
SMB	Små og Mellomstore Bedrifter
XAI	Explainable Artificial Intelligence
XGBoost	Extreme Gradient Boosting

1 Introduksjon

1.1 Bakgrunn

Bygg- og anleggsbransjen er en av de største bransjene i Norge, og sysselsetter omtrent 265 000 arbeidstakere (NHO, 2024). Bransjen har en årlig omsetning på rundt 680 milliarder kroner, og i 2021 var det i underkant av 60 000 foretak (SSB, 2023). Bransjen utgjør en vital del av den norske økonomien, men den står overfor en rekke utfordringer, særlig med hensyn til lønnsomhet. I 2023 opplevde bygg- og anleggsbransjen økonomisk nedgang som hovedsakelig skyldtes økte kapitalkostnader, drevet av høye rentesatser. Situasjonen ble videre forverret av logistikkproblemer hos leverandørene, og stigende prisnivå på materialer. Dette, i tillegg til usikkerhet i økonomien, medførte en bråstopp i oppstart av nye prosjekter for bransjen (Dalsegg & Lidsheim, 2023). I perioden fra 2021 til 2022, viser rapporten til Dalsegg og Lidsheim (2023) at det var et fall i driftsmargin fra 5,1 % til 4,5 %, og at bedrifter innen bygg og anlegg stod for 29 % av alle konkurser i Norge i 2022. Det gjør bygg og anlegg til den bransjen med flest konkurser i Norge. Dalsegg og Lidsheim (2023) skriver også at antall konkurser har økt i 2023, og at de forventer at dette tallet vil øke i fremtiden. Dette understreker viktigheten av å forske på hva som er de viktigste lønnsomhetsdriverne, for å kunne bidra til økt innsikt og forståelse for hva som driver lønnsomheten. Dette vil kunne bidra til at bedriftene vil klare å holde seg i drift i fremtiden, noe som er viktig for den norske økonomien. Bransjen skaper arbeidsplasser og driver økonomisk vekst, i tillegg til å stå for utvikling av infrastruktur, slik som veier, offentlige bygninger og broer, som samfunnet er helt avhengig av.

1.2 Problemstilling

Tidligere studier på lønnsomhetsdrivere i bygg- og anleggsbransjen har vært en kombinasjon av kvalitative og kvantitative analyseteknikker, med fokus på kun de største entreprenørene. For denne studien, som i hovedsak er kvantitativ, har vi i stedet valgt å fokusere på små- og mellomstore bedrifter ettersom disse utgjør 99 % av alle norske foretak, og står for 84 % av verdiskapingen i bygg- og anleggsbransjen (NHO, 2024). Fokuset i oppgaven er faktorene som påvirker lønnsomhet, og vi skal i oppgaven gjennomføre kvantitative undersøkelser i inneværende år, i tillegg til prediksjon av bedriftenes lønnsomhet ett år frem i tid. Denne typen studie, har så vidt vi vet, ikke blitt gjennomført i Norge før, og vi håper den vil gi verdifull innsikt i hva som driver lønnsomheten for bedrifter i bransjen. Vi har med det landet på følgende problemstilling: *Hva er de viktigste lønnsomhetsdriverne for norske SMB i bygg-*

og anleggsbransjen? Vi anser en lønnsomhetsdriver som en viktig forklaringsvariabel som påvirker lønnsomheten, enten i positiv eller negativ retning.

1.3 Formål og målsetninger

Formålet med denne masteroppgaven er å utforske temaet lønnsomhet med fokus på små- og mellomstore bedrifter i bygg- og anleggsbransjen. Vi ønsker at dette kan bidra med kunnskap som kan føre til bedre økonomisk forståelse og færre konkurser i bransjen.

Målsetningene er følgende:

- Gjennomgå eksisterende teori og forskning innen lønnsomhet og bygg- og anleggsbransjen for å identifisere uavhengige variabler som potensielt kan bidra til å forklare lønnsomhet.
- Utvikle forklaringsmodeller for lønnsomhet for SMB i bygg- og anleggsbransjen.
- Utvikle prediktive modeller som kan forutsi lønnsomhet ett år frem i tid for SMB i bygg- og anleggsbransjen.
- Identifisere de viktigste forklaringsvariablene til både forklaringsmodellene og de prediktive modellene.
- Diskutere betydningen og den praktiske innvirkningen til de viktigste forklaringsvariablene.

1.4 Oppgavens struktur

Masteroppgaven er delt inn i 7 hovedkapitler. I kapittel 2 presenteres det teoretiske rammeverket, som er delt inn i bredriftsspesifikt, bransjespesifikt og makroøkonomisk nivå, samt tidligere litteratur om bygg- og anleggsbransjen i Norge og internasjonalt. I kapittel 3 beskrives datasettet som er brukt i oppgaven, samt valg og begrunnelser for avgrensninger. Videre presenteres også variabelsettet med de uavhengige variablene og avhengig variabel, samt deskriptiv statistikk. I kapittel 4 presenteres den metodiske fremgangsmåten, og inkluderer beskrivelser av de ulike modellene og fortolkningsteknikkene som benyttes i oppgaven. Resultatene blir presentert i kapittel 5, og diskutert i kapittel 6. I det siste kapittelet, kapittel 7, presenteres konklusjonen, samt begrensninger ved oppgaven og forslag til videre forskning.

2 Teoretisk rammeverk

I dette kapittelet presenteres det teoretiske rammeverket som blir brukt i studien. Oppgaven har en tredelt tilnærming, hvor vi undersøker lønnsomheten fra ulike nivåer for å danne et helhetlig bilde. Først skal vi se på bedriftsspesifikke forhold, med sentrale regnskapsmessige nøkkeltall og teorien til Porter (1985) om kostnadsdrivere. Deretter skal vi se på bransjespesifikke forhold, for å analysere hvordan bransjekarakteristikker påvirker lønnsomheten for bedrifter i bransjen. For dette skal vi bruke Porters femfaktormodell som gir god innsikt i lønnsomhetspotensialet og konkurranseforholdene i bransjen. Til slutt skal vi, ved hjelp av PESTEL-rammeverket, utforske de makroøkonomiske forholdene.

2.1 Bedriftsspesifikke forhold

Ifølge McGahan og Porter (2002) er det de bedriftsspesifikke forholdene som påvirker lønnsomheten i størst grad. Dette bekreftes også av Føyen og Danielsen (2020), samt Lorentzen og Bergander (2022) sine funn, som begge bekrefter at de bedriftsspesifikke forholdene er av størst betydning for bedrifter i bygg- og anleggsbransjen. For å analysere de bedriftsspesifikke forholdene, tar vi utgangspunkt i regnskapsbaserte nøkkeltall og rammeverket til Porter (1985) om kostnadsdrivere.

2.1.1 Regnskapsbaserte nøkkeltall

Lønnsomhetsmål

I denne studien skal vi måle lønnsomheten gjennom nøkkeltallet *Return on Assets* (ROA), som måler bedrifters lønnsomhet i forhold til de totale eiendelene. Vi bruker gjennomsnittlige totale eiendeler fordi det vil gjøre det enklere å sammenligne bedrifter innen bygg og anlegg ettersom bedrifter starter og ferdigstiller prosjekter på ulike tidspunkt i løpet av et år. Totale eiendeler er summen av egenkapitalen og bedriftens totale forpliktelser. ROA gir innsikt i hvor effektivt en bedrift er i stand til å konvertere eiendeler til nettoinntekt, og nøkkeltallet er en indikator på operasjonell effektivitet. En økning i ROA indikerer at bedriften er i stand til å generere større profitt for hver krone de har i totale eiendeler. På den annen side, kan en fallende ROA tyde på at bedriften bruker for mye penger eller har gjort dårlige investeringer (Hargrave, 2024). En av fordelene med nøkkeltallet er at det er gunstig for å sammenligne bedrifter av ulik størrelse. Dette fordi det er et forholdstall som tar utgangspunkt i totale eiendeler, og fokuserer på hvor effektivt bedrifter utnytter eiendelene, og ikke på mengden av for eksempel egenkapital, som vil kunne variere

i stor grad mellom bedrifter. Nøkkeltallet viser altså hvordan bedrifter, uavhengig av størrelse, presterer i forhold til sine totale eiendeler. Det er vanskelig å si hva som regnes som en god ROA, men generelt sett vil alt over 5 % regnes som bra. Hva nøkkeltallet burde ligge på vil avhenge av flere faktorer og bedrifter i kapitalintensive bransjer, slik som bygg og anlegg, vil generelt sett ha lavere ROA enn mindre kapitalintensive bransjer, som teknologibransjen (Hargrave, 2024). Formelen for ROA ligger under:

$$\text{ROA} = \frac{\text{Driftsresultat} + \text{finansinntekter}}{\text{Gj.sn totale eiendeler}} \tag{2.1}$$

Likviditet

Med likviditet menes betalingsevne, at en er i stand til å innfri økonomiske forpliktelser (Berg, 2021, s. 123). Det er viktig å se på lønnsomheten til bedrifter, men det kan være enda viktigere å se på likviditeten. Ulønnsomme bedrifter kan fortsette driften over lang tid dersom likviditeten er god nok, men lønnsomme bedrifter kan risikere å gå konkurs dersom de har dårlig likviditet (Knardal & Sending, 2022). Ved å se på nøkkeltall for likviditet, kan bedrifter danne seg et bilde av fremtidige inn- og utbetalinger, og om det er behov for tilførsel av kapital. Nedenfor skal vi vise til nøkkeltall som brukes for å se på likviditeten. Disse er arbeidskapital, likviditetsgrad 1, likviditetsgrad 2 og kapitalens omløpshastighet.

Arbeidskapital

Arbeidskapitalen er et nøkkeltall som indikerer bedrifters evne til å betale regninger på kort sikt. Når arbeidskapitalen er positiv, indikerer det at bedriften er i stand til å betale regningene sine. Omløpsmidlene består enten av kontanter eller eiendeler som innen kort tid kan omdannes til kontanter. Den kortsiktige gjelden vil være penger som forsvinner fra bankkontoen innen kort tid (Berg, 2021). Formelen for beregning av arbeidskapital ligger under:

$$\text{Arbeidskapital} = \text{Omløpsmidler} - \text{Kortsiktig gjeld} \tag{2.2}$$

Tallet for arbeidskapitalen er ikke alltid til å stole på, eksempelvis når man ikke vet sammensetningen av omløpsmidler og kortsiktig gjeld. Omløpsmidlene kan variere fra bedrift til bedrift, hvor noen bedrifter har omløpsmidler som nesten utelukkende består av bankinnskudd, mens andres hovedsakelig består av varelager. Det kan derfor være lurt å se på arbeidskapitalen i prosent av driftsinntektene. Når denne blir høy så kan dette bety at varelageret bygger seg opp eller at kundefordringene ikke betales i tide (Berg, 2021). Ifølge Knardal og Sending (2022) så er gjennomsnittet på dette rundt 10 % i Norge. Formelen for arbeidskapital i % av driftsinntekter ligger under:

$$\text{Arbeidskapital i \% av driftsinntekt} = \frac{\text{Arbeidskapital}}{\text{Driftsinntekter}} \tag{2.3}$$

Likviditetsgrad 1

Likviditetsgrad 1 og arbeidskapital måler i prinsippet det samme, men forskjellen ligger i at forholdstallet man kommer frem til i likviditetsgrad 1, er enklere å bruke til sammenligninger. I likhet med arbeidskapitalen, er det vanskelig å si hvilket nivå nøkkeltallet bør være på, ettersom man ikke vet hva den kortsiktige gjelden og omløpsmidlene består av. Berg (2021) sier at nøkkeltallet antageligvis burde være høyere enn 1, men at det i praksis er umulig å si hvilket nivå dette nøkkeltallet burde ligge på. Dette begrunner han med at det er flere bedrifter med likviditetsgrad 1 som ligger rundt 1, som aldri har hatt betalingsproblemer, men at flere bedrifter har gått konkurs med likviditetsgrad langt over 1. Knardal og Sending (2022) sier at verdien helst skal være rundt 2. Formelen for likviditetsgrad 1 ligger under:

$$\text{Likviditetsgrad 1} = \frac{\text{Omløpsmidler}}{\text{Kortsiktig gjeld}} \tag{2.4}$$

Likviditetsgrad 2

For beregningen av likviditetsgrad 2, trekker man varelageret fra omløpsmidlene. På denne måten får man et tydeligere bilde av sammensetningen av omløpsmidlene, og hvor stor del av omløpsmidlene som er likvide (Berg, 2021). Knardal og Sending (2022) hevder dette nøkkeltallet burde være på minst 1, men at det ofte kan forekomme bransjeforskjeller i forhold til varelageret. Formelen for likviditetsgrad 2 ligger under:

$$\text{Likviditetsgrad 2} = \frac{\text{Omløpsmidler - Varelager}}{\text{Kortsiktig gjeld}} \tag{2.5}$$

Kapitalens omløpshastighet

Kapitalens omløpshastighet måler hvor effektivt bedrifter utnytter kapitalen til å generere inntekter (Sending, 2009). Selv om kapitalens omløpshastighet ikke er et direkte mål på likviditet, kan en høy omløpshastighet bidra til å forbedre likviditeten ved at bedriften er i stand til å raskt konvertere kapitalen til inntekter. Formelen for kapitalens omløpshastighet ligger under:

$$\text{Kapitalens omløpshastighet} = \frac{\text{Driftsinntekter}}{\text{Gj.sn totale eiendeler}} \tag{2.6}$$

Soliditet

Soliditet betyr robusthet, og henviser til bedrifters evne til å tåle tider med lav lønnsomhet (Berg, 2021, s. 123). Høy egenkapitalandel signaliserer god soliditet (Knardal & Sending, 2022). Det viktigste nøkkeltallet for soliditet er egenkapitalandel, men vi skal også se på finansieringsstrukturen, som kan betegnes som en forlengelse av soliditet (Berg, 2021). For dette skal vi se på nøkkeltallene finansieringsgrad 1, rentedekningsgrad og gjeldsgrad.

Egenkapitalandel

Det viktigste nøkkeltallet for å måle bedrifters soliditet, er egenkapitalandelen. Jo høyere dette nøkkeltallet er, desto bedre rustet vil bedriften være for tøffe tider. En bedrift vil i praksis være konkurs dersom de ikke har egenkapital (Berg, 2021). Nøkkeltallet forteller oss andelen av bedriftens eiendeler som er finansiert med egenkapital (Knardal & Sending, 2022). Formelen for beregning av egenkapitalandel ligger under:

$$\text{Egenkapitalandel} = \frac{\text{Egenkapital}}{\text{Totale eiendeler}} \tag{2.7}$$

Finansieringsgrad 1

Finansieringsgrad 1 måler i hvilken grad anleggsmidlene er langsiktig finansiert. Nøkkeltallet bør helst være under 1, som indikerer at anleggsmidlene er langsiktig finansiert, i tillegg til deler av omløpsmidlene. At de er langsiktig finansiert, vil si at de er finansiert gjennom egenkapital og langsiktig gjeld, i kontrast til kortsiktig gjeld (Berg, 2021). Formelen for beregning av finansieringsgrad 1 ligger under:

$$\text{Finansieringsgrad 1} = \frac{\text{Anleggsmidler}}{\text{Langsiktig kapital}} \tag{2.8}$$

Rentedekningsgrad

Rentedekningsgraden sier noe om i hvilken stand man er i til å betjene et lån. Dersom dette nøkkeltallet er 1, betyr det at bedriften vil ha null i resultat, fordi alt bedriften genererer går til å dekke rentekostnadene. Er nøkkeltallet 15 betyr det at resultatet før rentekostnader er 15 ganger større enn rentekostnadene (Berg, 2021). Formelen for beregning av rentedekningsgrad ligger under (Visma, udatert):

$$\text{Rentedekningsgrad} = \frac{\text{Resultat før skatt} + \text{finanskostnader}}{\text{finanskostnader}} \tag{2.9}$$

Gjeldsgrad

Gjeldsgraden måler størrelsen på gjelden i forhold til egenkapitalen i en bedrift. Nøkkeltallet gir innsikt i hvor stor del av bedriftens eiendeler som er finansiert gjennom gjeld, sammenlignet med egenkapitalen. Gjeldsgraden er en indikator på bedriftens soliditet, hvor bedrifter med lavere gjeldsgrad indikerer lavere risiko for konkurs (Langli, 2010). Formelen for beregning av gjeldsgrad ligger under:

$$\text{Gjeldsgrad} = \frac{\text{Sum gjeld}}{\text{Egenkapital}} \tag{2.10}$$

2.1.2 Porters kostnadsdrivere

Det er viktig å se på kostnadssiden og hvordan kostnadene kan forklare lønnsomhetsvariasjoner. Tradisjonelt sett har det vært vanlig å kun se på produksjonskostnader, og utelatt hvilken innvirkning forskjellige aktiviteter som markedsføring kan ha på kostnadene. For å få et mer nyansert kostnadsbilde, skal vi bruke rammeverket til Porter (1985) som fokuserer på kostnadsdrivere fremfor produksjonskostnader.

I rammeverket til Porter (1985) presenterer han ti kostnadsdrivere han mener viser bedrifters kostnadsposisjon. Disse kostnadsdriverne er: *stordriftsfordeler, læring, mønsteret i kapasitetsutnyttelsen, bindeledd, samhørighet, integrasjon, tidspunkt, skjønnsmessige retningslinjer, lokalisering* og *institusjonelle faktorer* (Porter, 1985, s. 91). Viktigheten av disse kostnadsdriverne kan variere fra bedrift til bedrift, uavhengig av om de er i samme bransje eller ikke. I denne studien utelates kostnadsdriverne *bindeledd, samarbeid, integrasjon, tidspunkt*, *skjønnsmessige retningslinjer* og *institusjonelle faktorer* fordi de er vanskelige å måle basert på tilgjengelig kvantitativ data.

Stordriftsfordeler

Porter (1985) mener stordriftsfordeler oppnås i tilfeller der bedrifter klarer å drive aktiviteter unikt og mer rasjonelt ved større volum, eller når bedrifter kan fordele kostnader over høyere omsetning. Han mener stordriftsfordeler kan ha både positive og negative virkninger på lønnsomheten. De negative virkningene, kalt stordriftsulemper, oppstår når kostnader er overproporsjonale med økt volum. Eksempler på stordriftsulemper knyttet til økt volum, kan være problemer knyttet til økt kompleksitet eller ved innkjøp av varer dersom det er uelastisk varetilgang ved store behov. En økning i volum kan også føre til økte lønnskostnader og redusert motivasjon blant de ansatte (Porter, 1985).

Læring

Ifølge Porter (1985) vil læring kunne føre til at kostnader ved aktiviteter reduseres over tid ved at de gjennomføres mer effektivt. Kostnadene kan eksempelvis reduseres ved bedre utnyttelse av arbeidskraft og effektiv planlegging av aktiviteter. Læring som holdes innad i foretaket vil kunne gi et kostnadsfortrinn i bransjen, men kun dersom det ikke er læringslekkasjer. Med læringslekkasjer så menes det at de andre aktørene i bransjen tar lærdom av hverandre, ofte gjennom tidligere ansatte, leverandører og konsulenter. Konkurransefortrinnet som et foretak oppnår gjennom læring, er avhengig av graden av læringslekkasje. Dersom læringslekkasjen er stor, vil det føre til at kostnadene for hele bransjen reduseres (Porter, 1985).

Kapasitetsutnyttelse

Kapasitetsutnyttelsen vil kunne ha stor påvirkning på bedriftens kostnadsstruktur, og muligheter til å oppnå konkurransefortrinn. Forskjellen mellom kapasitetsutnyt-

telse og stordriftsfordeler er at kapasitetsutnyttelse ser på hvor effektivt bedrifter utnytter sin tilgjengelige kapasitet, men for stordriftsfordeler ønsker man å øke produksjonsvolumet for å redusere kostnaden per enhet. Forholdet mellom bedriftens faste og variable kostnader, indikerer hvor følsom bedriften er for kapasitetsutnyttelsen. I tilfeller der bedriften har store faste kostnader, vil det være dyrt å ikke utnytte kapasiteten.

Lokalisering

Porter (1985) ser på lokalisering som en egen kostnadsdriver, og hevder at lokaliseringen til en bedrift påvirker kostnadene. Dette kan være kostnader knyttet til lønn, energi og råmaterialer. Bedriftens lokalisering i forhold til leverandører og kunder kan være en viktig faktor for transport- og logistikkostnader, i tillegg til aktivitetsnivå og hvor tilgjengelige ressurser er. Lokal infrastruktur, kulturelle normer og klima er også faktorer som kan påvirke kostnadene til bedrifter (Porter, 1985). En bedrift vil altså kunne oppnå et konkurransefortrinn ved gunstig lokalisering.

2.2 Bransjespesifikke forhold

I denne delen presenterer vi det teoretiske rammeverket som anvendes for å analysere hvordan bransjespesifikke faktorer vil kunne påvirke lønnsomheten til bedrifter innen bygg og anlegg. For dette skal vi benytte femfaktormodellen, som Michael Porter presenterte for første gang i 1979 i artikkelen om hvordan konkurransekrefter former strategi (Porter, 1979). Vi tar utgangspunkt i den oppdaterte versjonen av denne artikkelen, som ble publisert i 2008.

2.2.1 Porters femfaktormodell

Porters femfaktormodell gir innsikt i hvordan nøkkelkrefter i markedet kan påvirke lønnsomhetsforskjeller og konkurransekrefter i en bransje (Barney, 2006). Modellen tar for seg fem konkurransekrefter som sammen kan forklare bransjens konkurransekraft og lønnsomhetspotensiale. I denne oppgaven skal vi kun fokusere på én sentral kraft: intern rivalisering, fordi det særlig er relevant for å forstå lønnsomhetsutfordringene i en bransje.

Intern rivalisering

Intern rivalisering innebærer at bedriftene i samme bransje konkurrerer om de samme kundene, og tilbyr like tjenester og produkter. Ifølge Porter (2008), vil høy grad av rivalisering i bransjen føre til redusert lønnsomhet på grunn av konstant konkurranse. Det er konkurranseintensiteten, og grunnlaget de konkurrerer på, som bestemmer i hvilken grad lønnsomheten i bransjen reduseres. Når det er et stort antall konkurrenter som har lik størrelse og makt, vil graden av rivalisering være høy. Når det er lav vekst i bransjen, vil også rivaliseringen være høy ettersom bedriftene vil ha fokus på

å tilegne seg markedsandeler. I tillegg vil høye utgangsbarrierer, ofte ved høye faste kostnader, føre til at det er vanskelig å komme ut av markedet, og dermed øke rivaliseringen. Ofte vil mangelen på differensierte produkter føre til at bedrifter ender opp med å konkurrere på pris. Dette mener Porter (2008) er skadelig for lønnsomheten ved at profitten overføres fra bransjen og over til kunden.

2.3 Makroøkonomiske forhold

De makroøkonomiske forholdene gjør opp det ytterste laget i bedriftenes omgivelser, og de er i svært liten grad kontrollerbare for bedriftene. Bedrifter som klarer å tilpasse seg omgivelsene bedre enn andre, vil kunne oppnå konkurransefortrinn (Whittington mfl., 2020). For å få forståelse for hvordan makroøkonomiske faktorer kan påvirke lønnsomheten til bedrifter i bygg- og anleggsbransjen, skal vi benytte PESTEL-rammeverket.

2.3.1 PESTEL-rammeverket

PESTEL-rammeverket er et verktøy for å analyse de makroøkonomiske omgivelsene, og hvordan de påvirker en spesifikk bransje. Rammeverket består av seks aspekter som er *politiske, økonomiske, sosiale, teknologiske, miljømessige* og *juridiske faktorer* (Whittington mfl., 2020). Vi har valgt å begrense fokuset til de økonomiske og sosiale faktorene på grunn av relevans og tilgjengeligheten på kvantitativ data for disse faktorene.

Økonomiske faktorer

De økonomiske faktorene handler om hvordan makromiljøet blir påvirket av faktorer som rentenivå, valutakurser, og globale endringer i økonomisk vekst. Selv om disse faktorene ikke nødvendigvis gir god innsikt i lønnsomhetsvariasjoner innad i en bransje, kan de være viktige for å forstå hvorfor visse bransjer oppnår høyere lønnsomhet enn andre (Whittington mfl., 2020). En undersøkelse av økonomiske faktorer vil kunne kaste lys over hvordan bransjer blir berørt av både nasjonale og internasjonale trender. Eksempelvis kan utenlandske valutakursendringer eller endring i rentenivå påvirke bedrifter ved internasjonal handel. Det er viktig å ha innsikt i hvordan den økonomiske situasjonen og trender påvirker markedene som bransjer opererer i. Whittington mfl. (2020) trekker spesielt frem økonomisk utvikling som viktig å følge med på da det kan påvirke andre viktige økonomiske faktorer.

Sosiale faktorer

De sosiale faktorene tar for seg hvordan faktorer som demografi, kultur og trender påvirker makromiljøet. Whittington mfl. (2020) trekker blant annet frem den aldrende befolkningen i vestlige land, som vil føre til mindre tilgjengelig arbeidskraft. Kulturelle endringer kan føre til strategiske utfordringer. Et eksempel på dette er

det økte fokuset på miljøet fremfor profittmaksimering. Dette kan føre til endringer for tilbud og etterspørsel i en bransje. Videre kan geografisk plassering også være en viktig faktor for lønnsomhet, der noen bedrifter opplever høyere lønnsomhet på grunn av plassering (Whittington mfl., 2020).

2.4 Tidligere litteratur

For å forstå lønnsomhetsdrivere innen bygg- og anleggsbransjen, har vi sett på tidligere forskning, hvor særlig to masteroppgaver er relevante for studien. Disse oppgavene har søkt å finne de viktigste lønnsomhetsdriverne for bedrifter i bransjen, og har også utforsket lønnsomhetsfaktorer på et bedriftsspesifikt, bransjespesifikt og makroøkonomisk nivå. Videre er bygg- og anleggsanalysene til BDO sentrale for å danne et bilde over faktorene som påvirker lønnsomheten i bransjen.

Bedriftsspesifikke faktorer

Masteroppgavene av Lorentzen og Bergander (2022) samt Føyen og Danielsen (2020) kommer begge frem til at det er de bedriftsspesifikke faktorene som er de viktigste lønnsomhetsdriverne. Funnene deres viser en signifikant, men varierende effekt av selskapsstørrelse på lønnsomheten. Mens Føyen og Danielsen (2020) observerer at en økning i selskapsstørrelse, målt i antall årsverk, har en negativ effekt på lønnsomheten, indikerer Lorentzen og Bergander (2022) at det finnes en terskelverdi for størrelsen på bedriften og lønnsomheten. Dette vil si at en ytterligere vekst av bedriften, over terskelverdien, reduserer lønnsomheten, til tross for at de trekker frem at det foreligger stordriftsfordeler med hensyn til kapitalens omløpshastighet. En annen undersøkelse fra 1991, av 80 britiske entreprenørselskaper, konkluderte med at selskapsstørrelse, målt i omsetning, var signifikant og positivt korrelert med lønnsomhet (Akintoye & Skitmore, 1991). Dalsegg og Lidsheim (2023) peker likevel på at det er en generell trend i bransjen med økt omsetning, men med avtagende marginer.

Lorentzen og Bergander (2022) trekker frem lokalisering som en nøkkelfaktor for lønnsomhet, som påvirker tilgangen på ressurser og på aktivitetsnivået. Ifølge Dalsegg og Lidsheim (2023), henger endringen i aktivitetsnivået sammen med at de store byene har høyere befolkningstetthet og flere prosjekter. På den annen side, kan dette slå negativt ut på marginene, ettersom det vil være større konkurranseintensitet og derfor også økt prispress. Dalsegg og Lidsheim (2023) nevner også at bransjen står overfor utfordrende tider med lavere aktivitet, men at aktivitetsnivået på Vestlandet fremdeles er høy i de områdene som er tilknyttet olje- og gassindustrien på grunn av høye energipriser.

Føyen og Danielsen (2020) viser til at kapasitetsutnyttelse, målt gjennom produktivitet, har en signifikant og positiv sammenheng med lønnsomhet. Produktivitet er et mål på verdiskapingen bedriften oppnår per årsverk, og de hevder at sammenhengen er positiv ettersom bygg- og anleggsbransjen regnes for å være en arbeidsintensiv bransje. Det er også funnet at effektiv arbeidskapitalstyring bidrar til økt lønnsom-

het (Lorentzen & Bergander, 2022). Dette blir videre bekreftet av Deloof (2003), som hevder at effektiv arbeidskapitalstyring kan bidra til høyere lønnsomhet for bedrifter, gjennom å redusere antall dager på kundefordringer og lagertiden. Videre er likviditetsgrad 1 funnet å ha positiv effekt på lønnsomheten, som understreker viktigheten av god likviditetsstyring (Lorentzen & Bergander, 2022). Jolly Cyril og Singla (2020) sin undersøkelse av 67 bedrifter innen bygg og anlegg i India, bekrefter viktigheten av god likviditetsstyring, og hevder det er avgjørende for at bedrifter innen bygg og anlegg skal være lønnsomme. De trekker også frem at høy gjeldsgrad kan være lønnsomt ved at opptak av gjeld kan øke lønnsomheten når investeringsavkastningen overstiger gjeldskostnaden. De nevner også at høy gjeldsgrad fører til lavere skatt på grunn av rentekostnadene, noe som reduserer kostnaden av gjeld, sammenlignet med egenkapital. Dette står i kontrast til funnene til Ab. Halim mfl. (2014), som undersøkte bedrifter innen bygg og anlegg i Malaysia, og konkluderte med at høyere andeler av gjeld reduserte lønnsomheten.

Føyen og Danielsen (2020) tester om alder, som en indikator på erfaring, påvirket lønnsomheten. Funnene deres indikerer imidlertid ikke at alder er en signifikant variabel, som kan bety at langvarig erfaring ikke nødvendigvis henger sammen med økt lønnsomhet. Dette bekreftes også av Jolly Cyril og Singla (2020), som ikke finner noen sammenheng mellom bedriftens alder, og lønnsomhet.

Bransjespesifikke og makroøkonomiske faktorer

Capon mfl. (1990) hevder at bedrifter vil være mer lønnsomme jo større markedsandeler de har. Ifølge Dalsegg og Lidsheim (2021) er prispress og konkurranseintensitet ansvarlig for å drive ned marginene for bedrifter i bransjen. Økningen av utenlandske aktører som kommer inn i det norske markedet, i tillegg til veksten i antall norske bedrifter, er med på å øke konkurranseintensiteten. Bransjens konkurranseintensitet blir også påvirket av om produktene som tilbys er differensierte. Siden aktørene i bygg- og anleggsbransjen tilbyr lite differensierte tjenester, kan det argumenteres for at det vil være vanskelig å oppnå konkurransefortrinn på andre måter enn å tilby lavest pris. Eventuelt kan konkurransefortrinn oppnås gjennom gode kunderelasjoner og et sterkt omdømme (Dalsegg & Lidsheim, 2021). Produksjonsindeksen reflekterer bransjens aktivitetsnivå, og kan brukes for å se på makroøkonomiske svingningene i bransjen. Lorentzen og Bergander (2022) kom frem til at produksjonsindeksen, som måler utviklingen i aktivitet basert på timeverkstall for hele bransjen, var en signifikant forklaringsvariabel, men kun for en av modellene. Det vil si at en økning i produksjonsindeksen, som hadde økt betydelig i perioden, vil føre til en økning i lønnsomheten, målt gjennom ROA. Føyen og Danielsen (2020) hevder økningen i bransjens aktivitetsnivå skyldes store offentlige investeringer i utbygging. Videre skriver Dalsegg og Lidsheim (2023) at økningen av styringsrenten fra 2020 til 2023 kan føre til mindre privat konsum, noe som videre vil redusere etterspørselen etter kapitalvarer. Dette kan etter hvert resultere i redusert aktivitetsnivå i bygg- og anleggsbransjen. Det er likevel, som Dalsegg og Lidsheim (2023) påpeker, en treghet knyttet til forventninger i markedet.

Denne siden er blank med hensikt

3 Data

I dette kapittelet starter vi med å presentere datasettet som er grunnlaget for studien, og videre presenteres og begrunnes avgrensningene. Deretter presenterer vi oppgavens variabler, både den avhengige variabelen ROA som måler lønnsomhet, og de uavhengige variablene vi mener kan forklare lønnsomhet blant bedriftene i utvalget. Vi forsøker å ha gjennomgående kritisk refleksjon til de valgene som blir gjort i oppgaven i tilknytning til validitet og reliabilitet. Til videre i oppgaven vil vi informere om at preprosessering, modeller og figurer er generert i Python, som er programvare på engelsk. For å få gjennomgående tegnsetting i tekst, tabeller og figurer har vi valgt å benytte punktum som desimalskilletegn, selv om dette ikke er standard tegnsetting på norsk.

3.1 Beskrivelse av datasett

I den kvantitative analysen, benyttes et datasett som omfatter alle ukonsoliderte årsregnskap for norske bedrifter som er innrapportert til de norske myndighetene i perioden 2006-2022 (Wahlstrøm, 2023). Dataen er levert av Brønnøysundregistrene. Det er ingen manglende informasjon i datasettet, og dersom en regnskapspost ikke har en oppføring betyr det at verdien er null. I oppgaven vil hvert årsregnskap bli beskrevet som én observasjon.

Datasettet kan betegnes som paneldata, som betyr at det er en blanding av tverrsnitt- og tidsseriedata. Hver rad i datasettet er årsregnskapet for én spesifikk bedrift, for ett spesifikt år. Hvert årsregnskap er én observasjon, men vi har også tilgang på årsregnskap for den samme bedriften over flere år. I løpet av dataperioden har noen bedrifter startet opp og noen gått konkurs. Det betyr at vi ikke har årsregnskap for alle bedrifter i alle tidsperioder, og paneldataen er ubalansert.

I tillegg til de ukonsoliderte årsregnskapene, benytter vi datasett for årlig BNP per innbygger (SSB, udatert-e), gjennomsnittlig årlig drivstoffpris for diesel (SSB, udatert-a), gjennomsnittlig årlig styringsrente og endring i styringsrente gjennom året (Norges Bank, 2024a), årlig produksjonsindeks for bygg- og anleggsvirksomhet (SSB, udatert-c), og gjennomsnittlig årlig valutakurs mellom nok og euro (Norges Bank, 2024b). Vi ser kun på valutakurs mellom nok og euro ettersom Europa, særlig EU-land, er de viktigste handelspartnerene til Norge. Fra 2010 til 2024, har andelen av varer vi importerer fra Europa ligget mellom 60 og 70 prosent av den totale importen, hvorav 88 % av dette kom fra EU-land i 2023 (SSB, udatert-f). Vi har også datasett for lokasjon, hvor vi henter ut bedriftenes fylker basert på postkode (Hellesnes, udatert). Vi gjennomfører *left join*-operasjoner for å sammenkoble data-

settet for de ukonsoliderte årsregnskapene med denne dataen. En *left join*-operasjon tillegger hver observasjon i årsregnskapsdatasettet med nye variabler dersom de har en felles nøkkelvariabel. Makrovariablene ble sammenkoblet basert på regnskapsår, og lokasjonsdatasettet ble sammenkoblet basert på postnummer. Ettersom de fleste av observasjonene har tidsperiode fra før de ulike fylkessammenslåingene og fylkesoppdelingene, benyttes fylkene fra før 2018 med en inndeling i 19 ulike fylker (SSB, udatert-d). Vi fjerner bedrifter som tilhører Svalbard på grunn av få observasjoner, og at det ikke tilhører fastlandet.

Vi har valgt å benytte *winsorizing* for å redusere effekten av uteliggere uten å slette observasjoner i likhet med Paraschiv mfl. (2023). Vi *winsorizer* med 1% og 99% persentil. Det betyr at de 1% laveste verdiene for hver variabel hvert år blir endret til verdien i 1% persentilen, og de 1% høyeste verdiene for hver variabel hvert år blir endret til 99% persentilen. På grunn av divisjon ved utarbeidelse av de regnskapsbaserte nøkkeltallene, har noen observasjoner fått verdier som uendelig eller minus uendelig. Ettersom datasettet består av innrapporterte årsregnskap vil de regnskapspostene som mangler oppføring få verdi null. Når vi regner ut regnskapsbaserte nøkkeltall, vil en observasjon som har null i nevneren få verdi uendelig eller minus uendelig for det aktuelle nøkkeltallet. Dette gjelder særlig finansieringsgrad 1, fordi omtrent 9400 observasjoner i datasettet har ingen langsiktig gjeld. Det gjelder også for rentedekningsgrad der omtrent 2500 observasjoner er uten verdi for sum finanskostnader. Dette gjelder i tillegg arbeidskapital i % av driftsinntekt og lønnskostnad i % av driftsinntekt, fordi 15 observasjoner har innrapportert null i sum driftsinntekter. Avslutningsvis gjelder det også for likviditetsgrad 1 og 2, fordi 16 observasjoner har ingen kortsiktig gjeld. For observasjoner hvor en eller flere av disse regnskapsbaserte nøkkeltallene blir uendelig eller minus uendelig på grunn av dette, blir verdiene satt til maksimum og minimumsverdi for de variablene det gjelder for det samme året, før *winsorizingen*. På den måten reduserer winsorizingen effekten av uteliggere, samtidig som den fjerner denne svakheten ved utregning av regnskapsbaserte nøkkeltall.

3.2 Avgrensning

For å hente ut relevant informasjon, har vi avgrenset oppgaven. Kort oppsummert, ser vi på norske små- og mellomstore aksjeselskap, innen bygg og anlegg fra 2009-2022. I det følgende skal vi utdype de ulike avgrensningene. En oversikt over steg i avgrensningen og antall observasjoner i hvert steg finnes i tabell 3.1.

De fleste bedriftene i Norge er enten AS eller ENK (Brønnøysundregistrene, 2024). Alle AS har de samme formelle kravene for rapportering. Ved å kun fokusere på AS vil derfor den innrapporterte dataen vi har tilgang på være konsistent. I tillegg ønsker vi at fokuset skal være på norske bedrifter, og vi har derfor valgt å ekskludere bedrifter som ikke har hovedkontor i Norge. Dette vil være bedrifter som ikke har landskode «NO» i datasettet. Oppgavens problemstilling er knyttet til bedrifter i bygg- og anleggsbransjen, og vi inkluderer derfor kun bedrifter med næringskode «F» bygge- og anleggsvirksomhet. Innenfor denne næringskoden, er 2009 bedrifter

kategorisert som næringskode 41 - oppføring av bygninger, 316 bedrifter er kategorisert som næringskode 42 - anleggsvirksomhet, og 3741 bedrifter er kategorisert som næringskode 43 - spesialisert bygge- og anleggsvirksomhet. Til sammen utgjør dette 6066 unike bedrifter.

Norske bedrifter består nesten utelukkende av små- og mellomstore bedrifter (SMB), og de står for nesten halvparten av den årlige verdiskapingen i Norge. Innen bygg- og anleggsvirksomhet, står SMB for 84% av verdiskapingen i bransjen (NHO, 2024). Tidligere forskning har i stor grad fokusert på de største bygg- og anleggsbedriftene, og vi ønsker derfor å bidra med forskning på de mindre bedriftene. Vi antar også at flere av de mindre bedriftene i Norge har få eller ingen administrativt ansatte, og at denne oppgaven muligens kan bidra til at disse bedriftene får økt lønnsomhetsforståelse. Siden vi ikke har tilgang på data for antall årsverk for observasjonene har vi valgt å ta utgangspunkt i EU sin definisjon for SMB som er basert på omsetning og eiendeler (Directorate-General for Internal Market, Industry, Entrepreneurship and SMEs (European Commission), 2015). Denne definisjonen består av mikrobedrifter, småbedrifter og mellomstore bedrifter. Vi har valgt å avgrense oppgaven til å kun inkludere mellomstore og små bedrifter. Dette fordi vi mistenker at mikrobedrifter vil være mer preget av enkeltindivider, og at det vil være krevende å generalisere de mest sentrale lønnsomhetsdriverne for disse bedriftene. EU definerer SMB som bedrifter med mindre enn 43 millioner euro i eiendeler eller bedrifter med mindre enn 50 millioner euro i omsetning. Fordi vi ønsker å ekskludere mikrobedrifter, tar vi også i bruk denne definisjonen for å sette en nedre grense, og vi avgrenser til bedrifter med 10 eller flere estimerte antall årsverk.

For å kunne sette den nedre grensen må vi estimere antall årsverk. Gjennomsnittslønnen til en ansatt i bygg- og anleggsbransjen i 2022 var kr 599 100 (SSB, udatert-b). Vi har tilgang på informasjon om lønnskostnad i bedriftene, og benytter derfor dette til å lage et estimat for antall årsverk. Altinn (2024) skriver at en tommelfingerregel for å finne ut av hvor mye en ansatt koster, er å legge på 20-30 % på bruttolønnen. Vi har derfor valgt å estimere at en ansatt har lønnskostnad på 599 100 * 1,3 = 778 830. Dette gjør at vi kan estimere antall årsverk i en bedrift basert på lønnskostnader.

Finansdepartementet (2011) beskriver i sin rapport at utviklingen etter finanskrisen stabiliserte seg mot årsskiftet 2008/2009, og at det var bedring i realøkonomien utover 2009. For å sikre robuste analyser uten påvirkning fra finanskrisen, har vi valgt å benytte data fra 2009 og utover i forklaringsmodellene. Selve hoveddelen av analysen er i perioden 2010-2022, men vi er avhengig av data fra 2009 for å kunne beregne ROA for år 2010, da ROA beregnes med gjennomsnittlige eiendeler. I tillegg trenger vi data fra 2009 i de prediktive modellene for å lagge variablene. En avgrensning vi gjør i de prediktive modellene er å fjerne oppstartsåret til bedrifter som starter opp i løpet av utvalgsperioden 2010-2022. Dette fordi vi ikke har tilgang på årsregnskapet fra året før, slik at prediksjon blir umulig.

Tabell 3.1: Avgrensning av data

Steg i avgrensning	Antall observasjoner
All data: 2009-2022	4,379,351
Kun AS	3,865,839
Kun næringskode F og landskode NO	511,219
Omsetning < 50 mill euro, eiendeler < 43 mill euro	510,534
Antall årsverk > 10	38,027
Fjerner regnskapsår 2009	36,421
Fjerner Svalbard	36,380
For de prediktive modellene:	
Fjerner oppstartsåret dersom oppstart er mellom 2010-2022	31,817

3.3 Avhengig variabel

I denne delen skal vi presentere den avhengige variabelen, ROA. Den avhengige variabelen benyttes både som kontinuerlig variabel og som dummyvariabel i ulike deler av oppgaven. Tabell 3.3 inneholder deskriptiv statistikk for ROA. I oppgaven vil ROA bli beskrevet både som prosent og desimaltall.

De kvantitative analysene er delt inn i tre deler, hvor den første er deskriptive analyser ved anvendelse av ulike regresjonsmodeller. I denne første delen er ROA kontinuerlig. Den andre delen, predikerer kontinuerlig ROA, og den tredje delen klassifiserer ROA etter gitte kategorier. Generelt sett blir en ROA på mer enn 5 % regnet som bra, spesielt i kapitalintensive bransjer, og alt over 10 % regnes som veldig bra. Under 2 % blir regnet som dårlig, og mellom 2 % og 5 % blir regnet som moderat lønnsomt (Hargrave, 2024). Vi gjennomførte noen tester med fire kategorier som fulgte Hargrave (2024) sine avgrensninger for dårlig, moderat, bra og veldig bra ROA. Det ga antydninger til at modellen omtrent klassifiserte alle observasjonene som enten dårlig eller veldig bra. Vi valgte derfor heller tilnærmingen med to kategorier fordi det forenkler tolkning av modellen, og vi skiller ROA på 2%. Oversikt over klassifiseringsinndelingen finnes i 3.2. Bedrifter som har ROA på under 2 % blir ansett som «ikke lønnsom», og bedrifter med ROA over 2 % blir ansett som «lønnsom». Etter denne definisjonen har datasettet totalt 6622 ikke lønnsomme observasjoner og 25195 lønnsomme observasjoner. Dette trenes og testes på med en *rolling window* tilnærming som vi kommer til i kapittel 4.

Tabell 3.2: Klassifisering av ROA

Kategori	ROA	Antall
Ikke lønnsom	$< 2\%$	6622
Lønnsom	$> 2\%$	25195

Tabell 3.3: Deskriptiv statistikk for ROA

Statistikk	ROA
Antall	36380
Gjennomsnitt	0.10
Standardavvik	0.16
Minimum	-0.49
25%	0.00
50%	0.09
75%	0.20
Maksimum	0.55

3.4 Uavhengige variabler

I denne delen presenteres de uavhengige variablene vi mener kan forklare lønnsomhet for små- og mellomstore bedrifter innen bygg og anlegg. Alle variablene er enten forankret i teori eller tidligere forskning, og en komplett oversikt over variablene og kategori finnes i tabell 3.5. Variablene som ikke ble presentert i kapittel 2 med formler, vil bli gjennomgått her. I beregningene måtte vi foreta valg i forhold til utregning, og grunnlaget for avgjørelsene blir forklart for de variablene det gjelder i denne delen. Tabell 3.4 inneholder deskriptiv statistikk for de uavhengige variablene.

Markedsandel

Vi har valgt å ta utgangspunkt i bedrifters omsetning for å beregne markedsandeler. Vi ser kun på markedsandeler innenfor utvalget og ikke for hele bransjen. Vi beregner markedsandeler ved å dele hver enkelt bedrift sin omsetning, på utvalgets totale omsetning i det inneværende året. Grunnen til at vi har valgt å bruke omsetning som mål på markedsandeler, er fordi det er en robust indikator på markedets preferanse, og reflekterer verdien av varer og tjenester solgt for hver bedrift. Funnene til Capon mfl. (1990) indikerer at høyere markedsandeler, målt gjennom omsetning, kan føre til økt lønnsomhet.

Erfaring

Flere studier, blant annet av Føyen og Danielsen (2020) og Jolly Cyril og Singla (2020), bruker alder for å undersøke om erfaring fører til økt lønnsomhet for bedrifter innen bygg og anlegg. Porter (1985) mener bedrifter kan oppnå konkurransefortrinn dersom de lærer over tid. Vi bruker derfor alder for å undersøke om erfaring har noe å si for lønnsomheten. I datasettet, er alderen for hver bedrift oppgitt i antall dager, men for å forenkle tolkningen av denne variabelen har vi endret til alder i antall år ved å dele antallet dager på 365. Videre har vi gjennomført en logaritmisk transformasjon av variabelen, for å redusere skjevhet og spredning i fordelingen. Vi

antar at ett års økt erfaring vil ha ulik effekt på lønnsomheten for en yngre enn en eldre bedrift, altså at det er en ikke-lineær sammenheng som avtar med økende alder. Vi anser derfor ln(alder i år) som en mer passende variabel.

ln(antall årsverk)

Antall årsverk er en av variablene vi bruker for å se om selskapsstørrelse har en effekt på lønnsomheten. Siden datasettet ikke har data for antall årsverk, har vi laget et estimat selv. Vi har, som nevnt i punkt 3.2, kommet frem til et estimat for antall årsverk ved å dele bedriftenes lønnskostnader på gjennomsnittlig lønn, multiplisert med 1.3. Både Føyen og Danielsen (2020) og Lorentzen og Bergander (2022) hadde antall årsverk som forklaringsmodeller i sine undersøkelser, men de hadde tilgang på nøyaktige data for antall årsverk. Videre har vi valgt å gjennomføre en logaritmisk transformasjon av denne variabelen av samme grunn som for ln(alder i år).

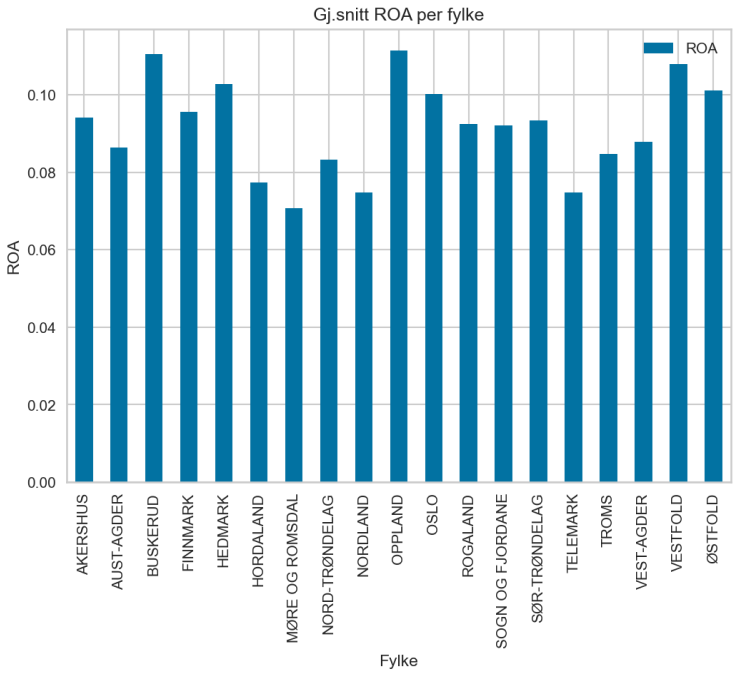
ln(omsetning)

Vi vil også se om omsetning har effekt på lønnsomheten, som et estimat for selskapsstørrelse. Siden det er så store forskjeller mellom omsetningen blant bedriftene i utvalget, ref. tabell 3.4, har vi valgt å logaritmisk transformere denne variabelen. Dette demper også effekten av mulige uteliggere i variabelen.

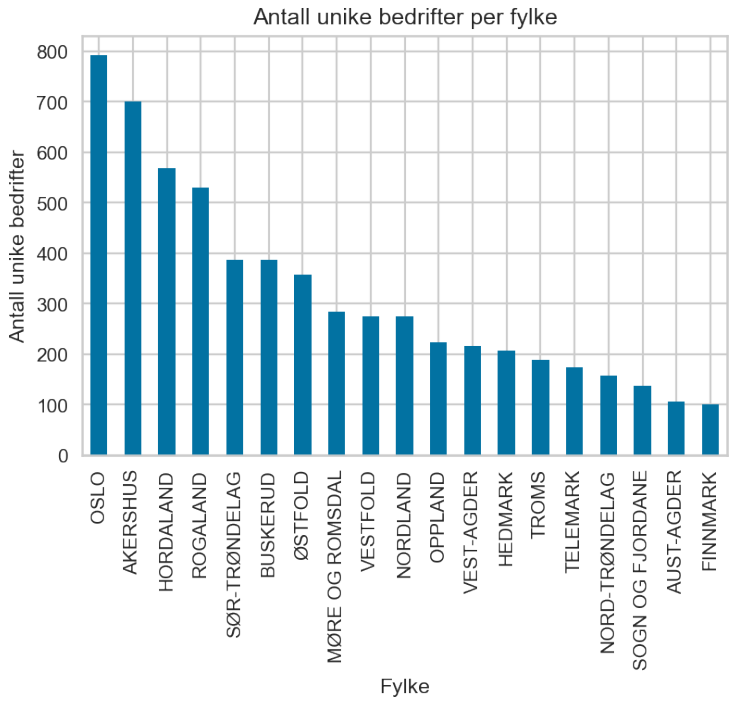
Fylke

Porter (1985) mener lokalisering kan ha effekt på lønnsomheten. For å undersøke om bedrifters lokalisering har effekt på lønnsomheten, har vi inkludert dummyvariabler for fylker. Variabelen består av 19 fylker og fylkesinndelingen er fra 2018. Følgende fylker er inkludert: *Østfold, Akershus, Oslo, Hedmark, Oppland, Buskerud, Vestfold, Telemark, Aust-Agder, Vest-Agder, Rogaland, Hordaland, Sogn og Fjordane, Møre og Romsdal, Trøndelag, Nordland, Troms, Finnmark* og *Svalbard*. Vi har valgt å ekskludere Svalbard på grunn av få observasjoner, og at det ikke tilhører fastlands-Norge. Etter sammenkobling av datasettet som inneholdt postnummer og fylker, manglet 1263 av observasjonene fylke. For å sikre at dummyvariablene for fylke var mest mulig korrekte valgte vi å innhente informasjon om fylke for disse observasjonene manuelt. Vi har valgt Akershus som referansefylke for regresjonsmodellene. Det er vanlig praksis å fjerne den første dummyvariabelen, som i dette tilfeller er alfabetisk, for å unngå problemer med multikollinearitet.

Figur 3.1 viser en oversikt over gjennomsnittlig ROA per fylke, og vi observerer variasjoner i gjennomsnittlig ROA for de ulike fylkene. Buskerud, Oppland, Oslo og Vestfold har de høyeste verdiene av ROA, mens Møre og Romsdal, Nordland og Telemark har lavere gjennomsnittlig ROA. Dette kan ses i sammenheng med figur 3.2 som viser antall bedrifter per fylke.



Figur 3.1: Gjennomsnittlig ROA per fylke



Figur 3.2: Antall unike bedrifter per fylke

Produktivitet

Kapasitetsutnyttelse, som diskutert i kapittel 2, er en viktig kostnadsdriver for å kunne oppnå et konkurransefortrinn. Det er spesielt viktig for bedriftenes evne til å konkurrere på pris. Faktoren vi bruker for å måle kapasitetsutnyttelsen er produktivitet, som er verdiskaping per årsverk (Bygballe mfl., 2019). Vi mener dette vil være en god faktor for å måle kapasitetsutnyttelsen for bedrifter innen bygg og anlegg, spesielt siden bransjen er arbeidsintensiv. Føyen og Danielsen (2020) brukte også produktivitet som forklaringsvariabel, med lik utregning. Formelen for produktivitet ligger under:

$$\text{Produktivitet} = \frac{\text{Verdiskaping}}{\text{Antall årsverk}} \tag{3.1}$$

For verdiskaping benytter vi definisjonen til Bygballe mfl. (2019, s. 1):

$$\begin{aligned}\text{Verdiskaping} = {} & \text{Lønn} + \text{Rentekostnader} + \text{Skatt} + \text{Eieravkastning} \\ & + \text{Avskrivninger} + \text{Nedskrivninger}\end{aligned} \tag{3.2}$$

Lønnskostnad i % av driftsinntekt

Bygg- og anleggsbransjen er som nevnt en arbeidsintensiv bransje, noe vi ser i tabell 3.4, hvor lønnskostnadene i gjennomsnitt utgjør 36 % av driftsinntektene blant bedriftene i utvalget. Ved å se på lønnskostnader som en prosentandel av driftsinntektene, vil vi kunne få innsikt i hvor effektive bedriftene er til å generere inntekter i forhold til lønnskostnader. Dette vil være en viktig variabel for måling av skala da bygg- og anleggsbransjen er svært arbeidsintensiv. Høyere andel kan indikere at bedriften er mindre effektiv i bruk av arbeidskraft. Vi finner lønnskostnader i prosent av driftsinntekter ved å dele lønnskostnadene på driftsinntektene. Variabelen vil beskrives både i prosent og som desimaltall.

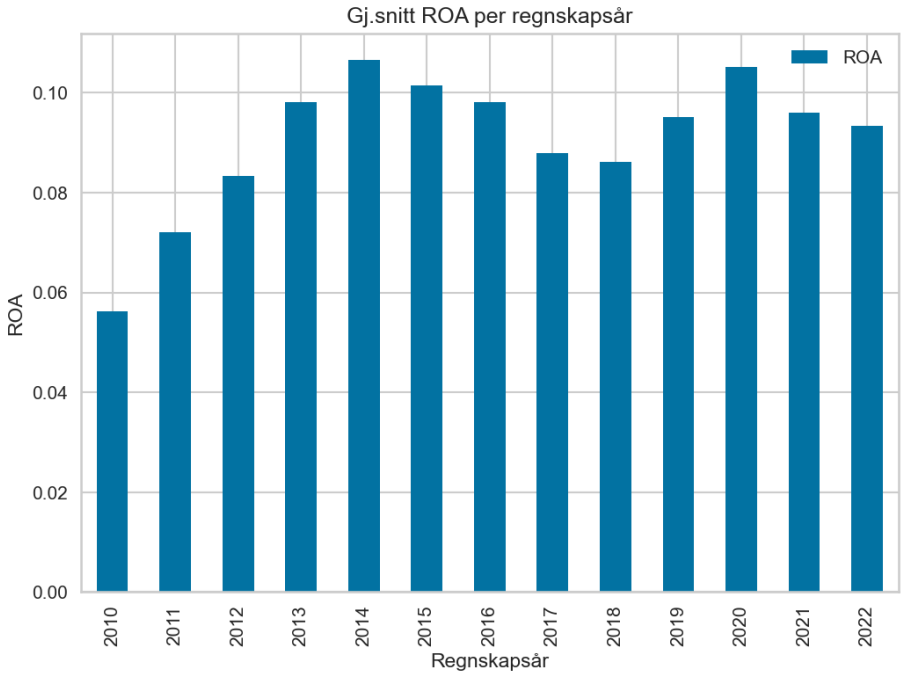
Kapitalens omløpshastighet

Kapitalens omløpshastighet blir beregnet ved å dele driftsinntekter på gjennomsnittlig totale eiendeler. I prediksjon og klassifisering av ROA bruker vi laggede variabler for å predikere og klassifisere ROA ett år fram i tid. Dataen er fra 2009 og utover, som vil si at når vi skal predikere og klassifisere ROA for 2010, benyttes de uavhengige variablene i lagget form fra årsregnskapet i 2009. Hvis vi skulle ha regnet ut kapitalens omløpshastighet med gjennomsnittlige totale eiendeler, måtte dette blitt et gjennomsnitt av totale eiendeler fra 2008 og 2009, men vi har ekskludert data fra 2008 for å unngå år som er særlig påvirket av finanskrisen. Ettersom vi har valgt å ekskludere disse årene i modellene, har vi valgt å være konsistent i dette valget ved å også unngå bruk av disse årene i beregning av de regnskapsbaserte nøkkeltallene. Dette problemet oppstår ikke i forklaringsmodellene som ikke lagger de uavhengige

variablene ett år bak i tid, men vi velger å benytte samme utregning på de ulike modellene i oppgaven for at de enklere kan sammenlignes. Kapitalens omløpshastighet beregnes derfor som driftsinntekter delt på totale eiendeler.

Selv om kapitalens omløpshastighet ikke er et direkte mål på likviditet gir variabelen innsikt i hvor effektivt bedrifter utnytter kapitalen, og er en indikator på hvor raskt en bedrift omgjør sine aktiva til inntekter. Høye verdier kan bidra til å forbedre likviditeten ved at bedriften er raske til å konvertere kapital til inntekter, og gjør at bedriften i stedet kan bruke kapitalen til mer lønnsomme aktiviteter.

Regnskapsår



Figur 3.3: Gjennomsnittlig ROA per regnskapsår

Vi har valgt regnskapsår 2010 som referanseår for de øvrige regnskapsårene. Dette er gjort fordi det er vanlig praksis å ekskludere den første verdien når det lages dummyvariabler til regresjonsanalyser, for å unngå potensielle problemer med multikollinearitet.

Figur 3.3 viser gjennomsnittlig ROA per regnskapsår. Her ser vi at ROA generelt er lavere for de første årene i datasettet. ROA har gjennomsnittlig lavere verdier i 2010, 2011 og 2012, før den øker igjen i 2014. Det er igjen en dypp i gjennomsnittlig ROA for årene 2017 og 2018, mens de resterende årene holder seg ganske jevnt rundt ROA på 0.09 og 0.10.

Tabell 3.4: Deskriptiv statistikk for uavhengige variabler

Variabler	Gj. snitt	Std. avvik	Median	Min	Maks
AK	8,106,770	15,676,592	4,136,305	-49,768,422	106,368,949
AK i %	0.11	0.21	0.09	-0.74	2.24
Likviditetsgrad 1	1.50	0.62	1.37	0.37	4.86
Likviditetsgrad 2	1.37	0.57	1.27	0.27	4.19
Kap. omløpshastighet	2.50	1.05	2.43	0.02	6.58
Egenkapitalandel	0.30	0.21	0.29	-0.63	0.81
Finansieringsgrad 1	231,477	902,627	2.78	0.00	4,730,371
Rentedek. grad	71,617	497,638	19.59	-12,528	6,845,474
Gjeldsgrad	4.04	9.12	2.22	-26.17	86.29
Årsverk	27.69	27.84	17.62	10.07	197.34
ln(årsverk)	3.05	0.65	2.87	2.31	5.28
Omsetning	77,326,363	95,827,019	43,343,817	9,944,178	677,101,988
ln(omsetning)	17.74	0.83	17.58	16.11	20.33
Alder i år	17.60	11.39	15.97	0.73	54.35
ln(alder i år)	2.59	0.84	2.77	-0.31	4
Produktivitet	995,519	373,836	924,970	253,149	4,118,677
Lønnskostn. i %	0.36	0.16	0.34	0.07	1.31
Markedsandel	0.000348	0.000440	0.000196	0.000028	0.003850
Prod. indeks	95.02	6.62	97.70	80	104.40
Rente snitt	0.98	0.59	1.05	0.08	2.14
Renteendring	0.19	0.92	0.00	-1.50	2.25
EUR valutakurs	9.26	0.98	9.33	7.47	10.72
Drivstoff	14.48.00	2.77	13.44	11.66	21.72
BNP	684,654	141,137	629,737	532,873	1,045,431

Tabell 3.5: Komplett oversikt over variablene

Variabel	**Kategori**
ROA	Lønnsomhetsmål
Arbeidskapital	Likviditet
Arbeidskapital i % av salgsinntekt	Likviditet
Likviditetsgrad 1	Likviditet
Likviditetsgrad 2	Likviditet
Kapitalens omløpshastighet	Likviditet
Egenkapitalandel	Soliditet
Finansieringsgrad 1	Soliditet
Rentedekningsgrad	Soliditet
Gjeldsgrad	Soliditet
ln(antall årsverk)	Selskapsstørrelse
ln(total omsetning)	Selskapsstørrelse
Markedsandel	Selskapsstørrelse og konkurranseintensitet
ln(alder)	Erfaring
Produktivitet	Kapasitetsutnyttelse
Lønnskostnader i % av driftsinntekter	Kapasitetsutnyttelse
Fylke	Lokalisering
Produksjonsindeks	Bransjens aktivitetsnivå
Gjennomsnittlig årlig styringsrente	Makrovariabel
Endring i styringsrente gjennom året	Makrovariabel
Valutakurs	Makrovariabel
Drivstoffpriser	Makrovariabel
BNP	Makrovariabel

Denne siden er blank med hensikt

4 Metode

I denne oppgaven anvendes teori og tidligere forskning for å utarbeide et variabelsett med forklaringsvariabler vi mener kan forklare lønnsomhetsforskjeller mellom bedriftene i utvalget. Sammen med dette benyttes LASSO som metode for å velge ut de viktigste forklaringsvariablene som danner grunnlaget for forklaringsmodellene OLS og fixed effects. Forklaringsmodellene analyserer de uavhengige variablenes påvirkning på lønnsomhet i det samme årsregnskapet, altså i det inneværende året. Det er derfor også interessant i de videre analysene å predikere hvilke og hvordan de uavhengige variablene påvirker lønnsomheten ett år frem i tid. Dette gjøres ved XGBoost og PyCaret. Ettersom det er krevende å predikere nøyaktig lønnsomhet skal vi også gjennomføre en klassifisering, som deler ROA inn i to kategorier «lønnsom» og «ikke lønnsom». I tillegg skal vi ta i bruk SHAP, PDP og ICE for å få bedre innsikt i hvordan de viktigste forklaringsvariablene påvirker ROA. Metodikken blir implementert ved bruk av Python, og man kan finne en fullstendig oversikt over programvarer og bibliotek som anvendes i oppgaven i vedlegg A. I dette kapittelet skal vi presentere disse metodene som i stor grad er basert på maskinlæring.

4.1 Maskinlæring

Maskinlæring er en type kunstig intelligens som benytter data og algoritmer for å etterligne måten mennesker lærer på, for å forbedre nøyaktighet (IBM, udatert). Forskning ved bruk av maskinlæringsteknikker har blitt stadig mer populært for å få innsikt, analysere og ta fornuftige beslutninger basert på store mengder data. Tidligere økonomisk forskning har blant annet brukt maskinlæring til å predikere konkurs (Paraschiv mfl., 2023), til å forbedre regnskapsmessige estimater (Ding mfl., 2020), og til å forutsi kundefrafall (Patil mfl., 2017).

Denne oppgaven er kvantitativ, og vi har både et regresjonsproblem hvor vi skal predikere en kontinuerlig verdi for ROA, og et klassifiseringsproblem hvor vi skal predikere kategori for ROA, som beskrevet i kapittel 3. Vi har tilgang på historisk data, og anvender derfor teknikker innen supervised maskinlæring. For hver observasjon av forklaringsvariabler xi, $i = 1, .., n$ finnes en assosiert avhengig variabel yi. Supervised maskinlæring trener en modell som knytter forklaringsvariablene til den avhengige variabelen ved å minimere forventningsverdien av en spesifisert objektiv funksjon. Supervised maskinlæring kan anvendes både for å predikere fremtidige observasjoner, samt for å fremme innsikt i forholdet mellom den avhengige variabelen og de tilhørende uavhengige variablene (James mfl., 2023).

For å trene modellene, defineres en objektiv funksjon som måler hvor godt modellen passer til dataen. Denne objektive funksjonen består ofte av en tapsfunksjon L, og en regulariseringsterm Ω, som skal minimeres for å oppnå best tilpasning til treningsdataen.

$$\text{obj}(\theta) = L(\theta) + \Omega(\theta) \tag{4.1}$$

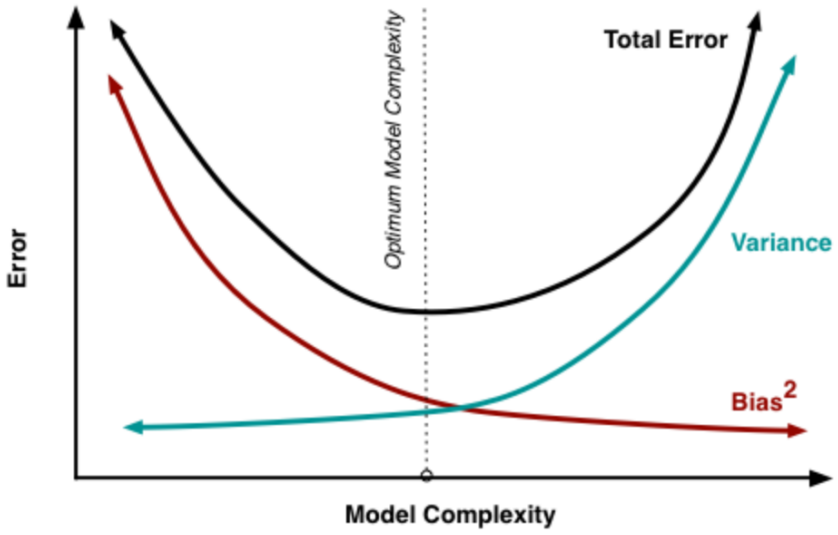
4.1.1 The Bias-Variance Trade-Off

En betydelig utfordring ved maskinlæring er å bygge en modell som også predikerer godt på nye observasjoner. Dette kan løses ved å dele dataen inn i et trenings- og testsett. Da bygges en modell på treningsdataen, for deretter å teste hvor godt den gjør det på data den ikke har sett før, som er testsettet. Inndeling av trenings- og testsett vil vi komme tilbake til i neste del. Forventet test MSE for en gitt verdi x_0 består av tre deler: variansen, den kvadrerte biasen og variansen av feilleddet. Dette kan vises ved følgende formel (James mfl., 2023):

$$E\left[\left(y_0 - \hat{f}(x_0)\right)^2\right] = \text{Var}(\hat{f}(x_0)) + [\text{Bias}(\hat{f}(x_0))]^2 + \text{Var}(\varepsilon) \tag{4.2}$$

Her kan vi se at for å minimere den forventede feilen i testsettet, må vi lage en modell som har både lav varians og lav bias. Varians beskriver hvor mye $\hat{f}$ ville endret seg om vi benyttet et annet treningssett, det vil altså si modellens sensitivitet til svingninger. En modell med høy varians, gjenkjennes ved at den er kompleks og fanger opp støy i treningsdataen. Den vil få gode resultater på treningssettet, men derimot ikke på testsettet. Da sier vi at modellen er overtilpasset. Bias beskriver derimot feilen ved å lage en forenklet modell til et virkelighetsproblem. Det kan føre til at den gjør sterke antakelser og derfor ikke klarer å fange opp underliggende trender i dataen. Bias kjennetegnes ved at den gjør de samme feilene på både treningssett og testsett, og vi sier at modellen er undertilpasset (Mudadla, 2023).

Hvis vi anvender fleksible metoder så vil typisk variansen øke og biasen synke. Et sentralt problem i maskinlæring er derfor å finne den avveiningen mellom bias og varians som gir minst testfeil. Dette kalles *the bias-variance trade-off*, og blir illustrert i figur 4.1. Det finnes ulike måter å håndtere denne avveiningen. Regulariseringstermen, som ble beskrevet som en del av den objektive funksjonen i formel 4.1, vil bidra til å redusere overtilpasning i modellene. Vi benytter regularisering både i LASSO og XGBoost-modellen. En annen måte er ved *feature engineering*, som handler om å være selektiv i valg av variabler som påvirker kompleksiteten til modellen (Mudadla, 2023). Dette gjør vi i regresjonsmodellene ved at vi foretar en gjennomgang av både økonomisk teori og tidligere forskning for valg av variabler.

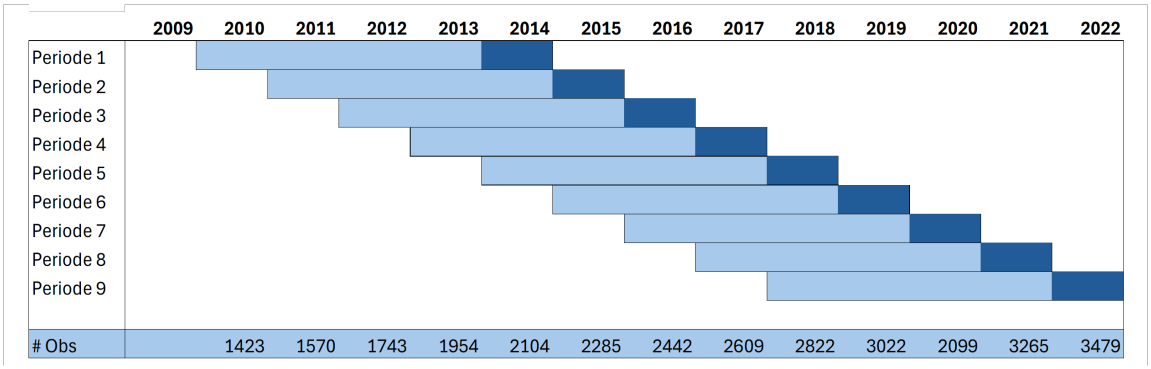


Figur 4.1: Illustrasjon av *The Bias-Variance Trade-Off* (Fortmann-Roe, 2012)

4.1.2 Trenings- og testsett

For metodene OLS, fixed effects og LASSO har vi forklaringsmodeller, og dataen fordeles ikke inn i trenings- og testsett. Modellene blir derfor trent og testet på hele datasettet. For XGBoost og PyCaret skal vi predikere og klassifisere ROA, og for å ta hensyn til *the bias-variance trade-off*, fordeles dataen i trenings- og testsett. Vi har valgt en *rolling window* tilnærming fordi det kan gjøre modellene mer relevante for testsettet. Treningssettet blir trent på observasjoner som er nærmere i tid til testsettet, i motsetning til om vi hadde satt en grense hvor de siste årene er testsett og de foregående årene treningssett. *En rolling window* tilnærming sikrer, i motsetning til en fast grense, at vi kan benytte flere år som både trenings- og testsett i ulike perioder, samtidig som modellene trenes uten å ha sett dataen i testsettet på forhånd. Dataen fordeles i ni ulike perioder som alle har et treningssett på fire år, og testsett det påfølgende året. På den måten blir modellen trent på data som er før testsettet i tid. Første periode trenes på data for 2010-2013 og testes på data for 2014. Videre ruller trenings- og testsettet ett steg fremover, som illustrert i figur 4.2.

Både prediksjonen og klassifikasjonen ser på ROA ett år frem i tid. De uavhengige variablene, med unntak av ln(alder i år) og lokasjon, *lagges* ett år bak i tid. Ln(alder i år) og lokasjon blir ikke *lagget*, fordi de er forutsigbare også ett år frem i tid. Bedriften har blitt ett år eldre, og vi antar at lokasjonen er den samme som året før. Dette betyr at for det første testsettet som er 2014, vil de resterende uavhengige variablene ha verdier fra 2013, og vi tester hvor godt disse verdiene predikerer ROA for 2014. Dette er tilsvarende for de resterende årene. I siste periode, år 2022, vil de uavhengige verdiene fra årsregnskapet og makroverdier være fra 2021 og benyttes for å predikere ROA i 2022. I figur 4.2 vil derfor året som står øverst representere det



Figur 4.2: Rolling window trenings- og testsett

året vi predikerer ROA for. Både trenings- og testsett er lagget på samme måte. Vi har også gjennomført noen tester hvor alle de uavhengige variablene var lagget for de to foregående årsregnskapene. Modellen fikk bedre resultater for treningssettene, men ikke testsettene som ga antydning til overtilpasning. Vi valgte derfor å utvikle modellene med kun *laggede* uavhengige variabler fra året før i denne oppgaven.

4.1.3 Evalueringsmål

Evalueringsmål for regresjonsmodeller

For å evaluere kvaliteten til modellene, benyttes forklaringsgraden og feil i prediksjonene. For regresjonsproblemer, er det vanlig å evaluere hvor nærme prediksjonene er den faktiske verdien. Prediksjonsfeil adresserer dette, og viser i gjennomsnitt hvor nærme prediksjonene er forventede verdier (Brownlee, 2021). For å evaluere modellene i oppgaven, brukes evalueringsmålene R^2, MSE, RMSE, MAPE, MdAPE og presisjon innenfor en toleransegrense. Vi anvender flere evalueringsmål for å underbygge kvaliteten til modellene. I prediksjonsmodellene beregnes verdier for både trenings- og testsett, men det er testsettet som blir vektlagt.

Forklaringsgraden til modellen betegnes ved R^2, og beskriver hvor mye de uavhengige variablene forklarer den avhengige variabelen. R^2 er en andel av forklart varians som vil ha en verdi mellom 0 og 1, hvor 1 betyr at de uavhengige variablene forklarer den avhengige variabelen perfekt. Det er derfor ønskelig å ha så høy verdi som mulig. Dersom antall observasjoner er mindre enn 10 ganger antall uavhengige variabler, bør man benytte korrigert forklaringsgrad (Studenmund & Johnson, 2017). Ettersom vi har et stort antall observasjoner, er det derimot ikke nødvendig og R^2 vil også være stabil. Formelen for R^2 er som følger :

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4.3}$$

For regresjonsproblemer, er det mest brukte evalueringsmålet Mean Squared Error (MSE) (Brownlee, 2021). MSE beregner gjennomsnittlig kvadrert forskjell mellom predikerte og forventede verdier i et datasett, og representerer derfor den uforklarte

variansen. MSE har lav verdi dersom prediksjonene er i nærheten av den faktiske verdien, og stor verdi dersom det er en betydelig forskjell. Siden MSE kvadrerer feilene, er den derfor avhengig av skalaen til dataen, og kan derfor være følsom for uteliggere. Store avvik fra den faktiske verdien kan derfor straffe modellen mer, men den vil likevel fungere godt til å sammenligne alternative modeller på samme datasett. MSE beregnes ut fra følgende formel:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 \tag{4.4}$$

RMSE er kvadratroten av MSE, og kan derfor betegnes som en utvidelse av formelen over (Brownlee, 2021). Vi tar også med RMSE fordi prediksjonsfeilen vil være i samme måleenhet som den avhengige variabelen, og kan derfor være mer intuitiv.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4.5}$$

Mean Absolute Error (MAE) beregner gjennomsnittet av de absolutte feilverdiene. I likhet med RMSE, vil MAE oppgi prediksjonsfeilen i samme måleenhet som den avhengige variabelen. Men i motsetning til MSE og RMSE, vil ikke MAE straffe store avvik mer enn små, men behandler alle avvik likt. En lavere MAE-verdi indikerer bedre modellnøyaktighet. MAE beregnes ved følgende formel:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{4.6}$$

Mean Absolute Percentage Error (MAPE) reflekterer prediksjonsfeil i regresjonsmodellene. Den måler den absolutte feilen mellom faktisk verdi og predikert verdi, og presenterer den som en prosentandel. Denne verdien beregnes for hver observasjon, for så å ta gjennomsnittet (Sagi, 2024). Formelen for MAPE er:

$$\text{MAPE} = \left(\frac{1}{n}\right)\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% \tag{4.7}$$

Median Absolute Percentage Error (MdAPE) er det samme som MAPE, men forskjellen ligger i at denne tar medianen av prediksjonsfeilen til de ulike observasjonene. Formelen er som følger:

$$\text{MDAPE} = \text{median}\left(\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\%\right) \tag{4.8}$$

I tillegg til å finne gjennomsnittlig og median prediksjonsfeil, skal vi finne hvor stor del av observasjonene i prosent som har prediksjonsfeil innen en satt toleransegrense. τ står for en viss toleransegrense som vi videre i oppgaven skal sette til 10%. Formelen blir som følger:

$$\text{Presisjon innen } \tau \text{ toleranse} = \left(\frac{1}{n}\right)\sum_{i=1}^{n}\mathbf{1}\left(\left|\frac{y_i - \hat{y}_i}{y_i}\right| \leq \tau\right) \tag{4.9}$$

Evalueringsmål for klassifiseringsmodeller

Vi skal bruke AUC og Brier Score til å evaluere kvaliteten til klassifiseringsmodellene for både trenings- og testsett. AUC står for Area Under the Curve, og er et mål på arealet under ROC-kurven. ROC-kurven, som står for The Receiver Operator Characteristics Curve, måler ytelsen til klassifiseringsmodeller, og kurven illustrerer forholdet mellom sann positiv rate og falsk positiv rate (James mfl., 2023). I denne oppgaven har vi en binær klassifiseringsmodell der bedrifter med ROA på over 2 % blir klassifisert som «lønnsom» og utgjør den positive klassen, mens bedrifter med ROA på under 2 % blir klassifisert som «ikke lønnsom», og utgjør den negative klassen. En oversikt over alle mulige utfall som modellen kan predikere er vist i tabell 4.1.

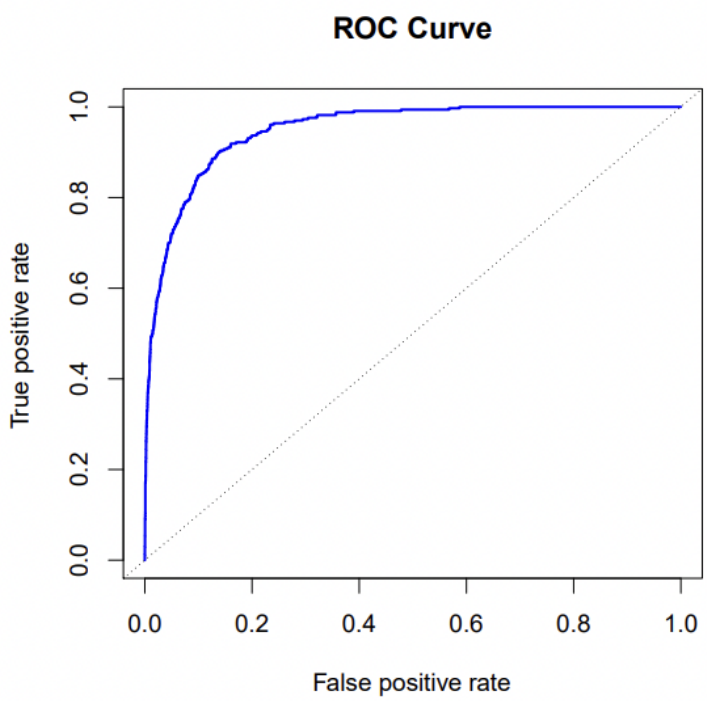
Tabell 4.1: Klassifiseringsutfall

	Faktisk lønnsom	**Faktisk ikke lønnsom**
Predikert lønnsom	Sann positiv	Falsk positiv
Predikert ikke lønnsom	Falsk negativ	Sann negativ

En hendelse vil være *sann positiv* dersom modellen predikerer den som positiv, og den faktisk er positiv. Det samme gjelder for *sann negativ* dersom modellen predikerer en negativ hendelse, og den faktisk er negativ. En *falsk positiv* hendelse tilsier at modellen har predikert en hendelse som positiv, men som faktisk er negativ. Dette er kjent som en type 1 feil. For falske negative hendelser, har modellen predikert en hendelse som negativ når den faktisk er positiv. Dette er kjent som en type 2 feil. ROC-kurven ser på forholdet mellom den sanne positive raten og den falske positiv raten. Formlene er gitt ved:

$$\text{Sann positiv rate} = \frac{\text{Antall korrekt klassifisert som lønnsomme}}{\text{Totalt antall lønnsomme bedrifter i utvalget}} \tag{4.10}$$

$$\text{Falsk positiv rate} = \frac{\text{Antall feilklassifisert som lønnsomme}}{\text{Antall } \textit{ikke lønnsomme} \text{ bedrifter i utvalget}} \tag{4.11}$$



Figur 4.3: ROC-kurve (James mfl., 2023, s. 155)

Figur 4.3 viser en ROC-kurve som går mot det øvre venstre hjørnet, som indikerer at modellen har høy sannsynlighet for å predikere sanne positive hendelser og lav sannsynlighet for falske positive. Det er ønskelig med en ROC-kurve så langt opp i venstre hjørne som mulig fordi det reflekterer en høy AUC-verdi. Verdien for AUC ønskes så høy som mulig, og verdiene ligger mellom 0 og 1 (James mfl., 2023). Ifølge Çorbacıoğlu og Aksel (2023), vil en AUC-verdi på over 0.80 indikere at modellen er nyttig til å predikere. En AUC-verdi på 0.50 tilsier at det er like sannsynlig at modellen predikerer riktig på halvparten av de positive hendelsene og halvparten av de negative hendelsene.

Brier Score er et vanlig evalueringsmål for ytelsen til klassifiseringsmodeller, og er spesielt nyttig for binære modeller. Evalueringsmålet ser på gjennomsnittlig kvadrert feil mellom de predikerte sannsynlighetene og de faktiske verdiene, og har derfor likheter med MSE (Dash, 2020). Formelen for Brier Score er gitt ved:

$$BS = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2 \tag{4.12}$$

I formelen er N antall observasjoner, f_t predikert sannsynlighet for observasjon t, og o_t er det faktiske utfallet. Verdiene for Brier Score ligger mellom 0 og 1, hvor verdi 0 indikerer en mer nøyaktig modell (Dash, 2020).

4.2 Modeller

4.2.1 OLS

I denne studien skal vi bruke OLS, som står for Ordinary Least Squares, for å modellere forholdet mellom den avhengige variabelen og de uavhengige variablene. Målet med OLS er å minimere summen av kvadratene av residualene, hvor residualene defineres som differansen mellom predikert og virkelig verdi (Studenmund & Johnson, 2017). En multippel regresjonsmodell med flere uavhengige variabler kan uttrykkes som:

$$\min_{\beta} \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{1i} + \ldots + \beta_n X_{ni}))^2 \tag{4.13}$$

Her er Y_i den avhengige variabelen, $X_1 i$,..., $X_n i$ er de uavhengige forklaringsvariablene, og n er antallet observasjoner. For å sikre at OLS-estimatene er Best Linear Unbiased Estimator (BLUE) som fremgår av Gauss-Markov-teoremet, er det en rekke forutsetninger som må være oppfylt. Forutsetningene fremgår på side 93 i Studenmund og Johnson (2017):

1. Regresjonsmodellen er lineær, korrekt spesifisert og har et additivt feilledd.
2. Forventningen til feilleddet skal være null.
3. Alle forklaringsvariablene er ukorrelert med feilleddet.
4. Ingen autokorrelasjon ved at feilleddene skal være uavhengig.
5. Ingen heteroskedastisitet ved at feilleddene har konstant varians
6. Ingen perfekt multikollinearitet ved at ingen forklaringsvariabel er en perfekt lineær funksjon av en annen forklaringsvariabel
7. Feilleddet er normalfordelt

Vi anvender OLS i studien for å danne et bilde av den overordnede sammenhengen mellom de uavhengige variablene og ROA. Fordelen med OLS er at den er enkel å tolke, også for de uten mye erfaring med kvantitativ analyse. En av begrensningene til OLS, er at det oppstår utfordringer dersom det er for mange forklaringsvariabler. Ifølge Tibshirani (1996) blir OLS mindre forståelig med et stort antall variabler. Variabelsettet i oppgaven er basert på teori og tidligere forskning, og inneholder 51 uavhengige variabler, og det er derfor en risiko for at modellen blir krevende å tolke. For å løse dette anvender vi LASSO, som blir forklart i neste underkapittel, i forkant av OLS til å velge ut de variablene som bidrar mest til å forklare ROA i inneværende år.

For å sikre at resultatene fra OLS er pålitelige, har vi gjennomført tester for autokorrelasjon, heteroskedastisitet og multikollinearitet for å sjekke om henholdsvis forutsetning 4, 5 og 6 i Gauss-Markov teoremet er oppfylt. Disse testene er dokumentert i vedlegg C. Vi har testet for mulige problemer med multikollinearitet ved å anvende analyse for Variance Inflation Factor (VIF), som ser på i hvor stor grad hver

uavhengig variabel kan forklares av de andre uavhengige variablene i modellen. Vi inkluderer ikke alle variablene som blir valgt av LASSO i OLS og fixed effects på grunn av mulige problemer med multikollinearitet. Som vi kan se i tabell C.1, indikerer VIF-verdiene at det ikke er vesentlige problemer med multikollinearitet, med unntak av produktivitet, som har en VIF-verdi på litt over 7. Det at produktivitet har noe høyere VIF-verdi henger trolig sammen med at modellen inkluderer lønnskostnad i % av driftsinntekt og kapitalens omløpshastighet. Vi betrakter produktivitet som en viktig forklaringsvariabel å inkludere i lys av den sterke teoretiske relevansen i forhold til kapasitetsutnyttelse, som Porter (1985) beskriver i teorien om kostnadsdrivere, i tillegg til tidligere forskningsresultater (Føyen & Danielsen, 2020). Selv om en VIF-verdi på over 7 kan indikere mulige problemer med multikollinearitet, anses likevel verdier under 10 som akseptabelt (O'brien, 2007). Vi har derfor valgt å beholde denne variabelen i oppgaven. For å ta hensyn til oppdaget heteroskedastisitet i Breusch Pagan testen, tabell C.2, implementeres HC0 robust regresjonsanalyse (Long & Ervin, 2000). Når det gjelder autokorrelasjon, indikerer Durbin-Watson-testen lite til ingen autokorrelasjon for residualene.

4.2.2 LASSO

LASSO står for Least Absolute Shrinkage and Selection Operator, og er en regulariseringsmetode som skal forbedre tilpasningen til en regresjonsmodell. Metoden ble introdusert og popularisert av Robert Tibshirani (1996), som beskriver i sin forskningsartikkel at det særlig er to svakheter ved OLS. Den første er at OLS estimatene ofte har lav bias og høy varians, noe som kan påvirke forklaringsgraden til regresjonen. Den andre er at OLS er mindre forståelig dersom man har et stort antall variabler. Tibshirani (1996) viser at ved å tillegge et straffeledd til 4.13, slik som vi kan se i formel 4.14, kan noen av modellens koeffisienter bli til null. LASSO benytter L1 straff, der straffeleddet er summen av absoluttverdiene til koeffisientene multiplisert med regulariseringsparameteren λ. LASSO vil derfor kunne redusere antallet variabler, noe som både kan forhindre overtilpasning og øke tolkbarheten.

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = RSS + \lambda\sum_{j=1}^{p}|\beta_j| \tag{4.14}$$

Hyperparameteren λ bestemmer regulariseringsgraden. Denne blir bestemt ved *GridSearchCV*, som fungerer ved at man velger et rutenett av λ-verdier, og utfører kryssvalidering for hver verdi av λ. Vi velger den parameteren som gir lavest kryssvalideringsfeil, og tilpasser modellen på nytt med den valgte verdien av λ. En tilstrekkelig stor verdi av λ, vil tvinge alle koeffisientene lik null. Ved å gradvis senke verdien til λ, kan vi analysere rekkefølgen for hvilke variabler som introduseres til modellen. På samme måte som Paraschiv mfl. (2023), tolker vi det slik at en variabel som inkluderes i modellen ved høyere verdier av λ, indikerer høyere prediktiv evne og derfor også høyere viktighet. På den måten blir LASSO en form for variabelseleksjon. Det vil si at vi i denne oppgaven tolker de variablene som inkluderes i LASSO

regresjonen ved høyere λ som en viktigere lønnsomhetsdriver. Dette illustreres i en figur med λ-verdi på x-aksen og koeffisientene på y-aksen.

LASSO har tidligere blitt anvendt i flere undersøkelser til utvelgelse av variabler, og Paraschiv mfl. (2023) hevder at LASSO var den beste metoden for variabelseleksjon i deres artikkel om konkursprediksjon. I tillegg kom Tian mfl. (2015) og Tian og Yu (2017), frem til at LASSO ga bedre prediksjonsevne enn andre populære modeller for konkursprediksjon. En stor fordel ved LASSO, i forhold til andre metoder innen variabelseleksjon, er at det er en metode som krever mindre datakapasitet (Paraschiv mfl., 2023).

4.2.3 Fixed Effects

En fixed effects-modell er en statistisk teknikk som blir brukt til å analysere paneldata. Selve grunnlaget for modellen er å hensynta individuell heterogenitet, når denne er konstant for hver enhet over tid, men varierer mellom enhetene. Modellen vil altså ta hensyn til de enhetsspesifikke egenskapene som ikke endres over tid, men som kan påvirke avhengig variabel. I en fixed effects-modell, antar man at alle enhetene har sin egen unike konstant som fanger opp de tidsuavhengige, uobserverte egenskapene. På denne måten vil forvridningene i estimatene bli redusert for de uavhengige forklaringsvariablene, som kan skyldes korrelasjon mellom de forklarende variablene og de uobserverte faktorene (Hill, 2012). Hausman-test brukes for å avgjøre om en fixed effects-modell foretrekkes fremfor en random effects-modell. Resultatene fra denne testen finnes i vedlegg C, og viser at fixed effects er å foretrekke.

I studien er hver bedrift en egen enhet som analyseres over flere år. Dette kalles en firm fixed effects-modell, og utfordringen er at hver enkelt bedrift opererer under særegne forhold, noe som kan påvirke lønnsomheten. Det er disse særegne og individuelle forskjellene som kan forvrenge forståelsen av hvordan forklaringsvariablene i modellen påvirker lønnsomheten. Ved å anta at det er en unik effekt for hver enkelt enhet, som fanger opp disse tidsuavhengige og uobserverte karakteristikkene, blir dette problemet løst. Fixed effects er derfor en supplerende modell til OLS som fjerner innflytelsen til disse individuelle forskjellene, og gjør at man kan analysere effekten forklaringsvariablene har på lønnsomheten (Hill, 2012). Fixed effects benytter de samme variablene som OLS.

4.2.4 XGBoost

Extreme Gradient Boosting, fra nå av XGBoost, er et optimert gradient-boosting maskinlæringsbibliotek som ble introdusert av Chen og Guestrin (2016). Det er en ensemble metode, bygget som et skalerbart maskinlæringssystem for boosting av beslutningstrær. I denne seksjonen skal vi først beskrive de svake lærerne, beslutningstrær, for deretter å gå inn på konseptet boosting, som sammen blir til den sterke læreren XGBoost.

Beslutningstrær består av en serie med delingsregler, som kan illustreres som et tre med beslutningspunkter og bladnoder. Måten dette gjøres på, kalles rekursiv

partisjonering, som betyr at man gjentatte ganger deler dataen i to for å oppnå maksimal homogenitet av resultatet innenfor hver del. For regresjonsproblemer vil prediksjonen være gjennomsnittet av observasjoner som faller innenfor hver bladnode. Et fullvokst tre vil være komplekst og tilpasse seg støy i dataen (Shmueli mfl., 2019). Selv om beslutningstrær er enkle og nyttige for forståelse, kan man få problemer med overtilpasning, og de gjør det heller ikke like bra som andre metoder når det gjelder prediktive evner (James mfl., 2023). I vedlegg D ser vi et eksempel på hvordan XGBoost bruker boosting til å trene beslutningstrær i oppgaven.

Gradient Boosting er en fremgangsmåte for å forbedre prediksjoner fra beslutningstrær som ble introdusert av Friedman (2001). Den er basert på å bygge et stort antall beslutningstrær sekvensielt, hvor hvert tre benytter informasjon fra tidligere bygde trær. Trærne blir tilpasset ved å bruke de gjenværende residualene, og vil på den måten gradvis redusere tapsfunksjonen. Boosting er en metode som lærer sakte, og James mfl. (2023) beskriver at modeller som lærer sakte, har en tendens til å prestere godt.

XGBoost algoritmen benytter gradient boosting ved å sekvensielt legge til det treet som bidrar mest til å minimere den objektive funksjonen (Chen & Guestrin, 2016):

$$
\begin{aligned}
L(\phi) &= \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \\
\text{hvor} \quad \Omega(f) &= \gamma T + \frac{1}{2}\lambda \|w\|^2
\end{aligned}
\tag{4.15}
$$

Den første delen av formelen er tapsfunksjonen som måler forskjellen mellom predikert og faktisk verdi. Hver f_k tilsvarer et uavhengig tre q med bladvekter w. Den andre delen av formelen er summen av regulariseringstermene for hvert tre i modellen. I regulariseringstermen får antallet blader T en straff gjennom parameteren γ. γ kontrollerer om en gitt node på et tre vil dele seg basert på forventet reduksjon i tapet. Høyere verdi av γ vil derfor føre til færre splittelser.

Modellen blir trent additivt som betyr at den iterativt velger det treet f_t som bidrar mest til å minimere formel 4.15. Ved at $\hat{y}_i^{(t)}$ er prediksjonen for instans i i iterasjon t, så legges det til den f_t som minimerer følgende formel (Chen & Guestrin, 2016):

$$
L^{(t)} = \sum_{n=1}^{N} l(y_n, \hat{y}_n^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{4.16}
$$

I en masteroppgave fra NTNU forsket Nielsen (2016) på hvorfor XGBoost har vunnet mangfoldige maskinlæringskonkurranser. Han begrunner dette ved at additive tremodeller har gode representasjonsevner, og metoden tar hensyn til avveiningen mellom bias og varians i nesten alle aspekter ved læringsprosessen. Læringsprosessen starter med lav varians og høy bias, og så reduseres denne biasen ved å redusere størrelsen på nabolag i modellen der det er nødvendig. Han understreker også betydningen av at modellen er adaptiv, og justerer seg etter dataen. Dette tyder på at XGBoost vil være en god metode til denne studien.

Tabell 4.2: Beskrivelse av hyperparametere i XGBoost modellen.

Hyperparameter	Beskrivelse
`learning_rate`	Hvor raskt modellen tilpasses residualfeil.
`n_estimators`	Antallet svake lærere/trær.
`max_depth`	Den maksimale dybden per tre.
`subsample`	Prosent av treningssettet som kan brukes per boostingrunde.
`colsample_bytree`	Prosent av variabler som kan brukes per boostingrunde.
`min_child_weight`	Minimumssummen av vekter som kreves for å lage en ny node.
`reg_alpha`	L1 regularisering/straff på bladvekter.
`gamma/min_split_loss`	Minimumsreduksjon i tapsfunksjonen som kreves for ny splitt.
`seed`	For å kunne reprodusere resultater.
`objective`	Tapsfunksjon med kvadratfeil.

XGBoost har mange justerbare parametere som må fastsettes før trening av en modell, disse omtales som hyperparametere. En oversikt over hyperparameterne som benyttes i modellene finnes i tabell 4.2. En naturlig regulariseringsparameter er antallet trær. Dette regulerer hvor mye det forventede tapet på treningsdataene kan minimeres. Friedman (2001) skriver at det ofte er funnet at regularisering ved å tillegge et straffeledd som krymper verdiene, er bedre enn å begrense antallet trær. For å gjøre dette i XGBoost-modellen, begrenses hver oppdatering av modellen ved læringsraten (*learning_rate*) i tillegg til antallet trær *(n_estimators)*. Å redusere verdien av læringsraten vil kreve et stort antall trær og motsatt. Etter anbefaling av Friedman (2001) skal vi finne de optimale verdiene for disse parameterne samtidig. Her må vi også ta hensyn til at et stort antall trær vil være beregningsmessig krevende.

For å optimere hyperparameterne i modellen, anvendes *RandomSearch* kryssvalidering. I en studie av Zahedi mfl. (2021), ble det funnet at bruk av både *GridSearch* og *RandomSearch* på maskinlæringsalgoritmer forbedret prediktiv evne. En begrensning ved *GridSearch* er derimot at den er beregningsmessig krevende, særlig for trebaserte modeller. Dette er også noe Zahedi mfl. (2021) diskuterer i sin forskningsartikkel. Siden vi anvender trebaserte modeller, har vi valgt å benytte *RandomSearch* med kryssvalidering for å gjøre tilpasning av hyperparametere mindre beregningsmessig utfordrende. *Randomsearch* velger tilfeldig ulike kombinasjoner av hyperparametere innenfor rutenettet, som finnes i tabell E.1 i vedlegg, og gjentar denne prosessen til den har nådd et spesifisert antall gjennomganger. I oppgaven har vi valgt å kjøre 100 tilfeldig genererte iterasjoner for hver modell, med 3-folds kryssvalidering. *RandomSearch* er på den måten mindre beregningsmessig utfordrende enn *GridSearch* som tester alle mulige kombinasjoner av hyperparametere. I denne analysen hadde *GridSearch* resultert i 2048 iterasjoner for hver modell, som er en for tidkrevende operasjon. Valgte hyperparametere for modellene finnes i vedlegg E.

4.2.5 PyCaret

PyCaret ble introdusert av Moez (2020), og er en *autoML*-metode som vil si at den automatiserer deler av maskinlæringsprosessen. PyCaret er et bibliotek i Python med åpen kildekode, og som krever få linjer med kode. De påstår selv at «PyCaret er et alternativt lavkodebibliotek som kan brukes til å erstatte hundrevis av kodelinjer med bare få ord» (Moez, 2020, avsn. 2). Den er bygd opp ved sekvensielle pipelines som bidrar blant annet til preprosessering, funksjonsvalg, og valg av verdi for hyperparametere. Den er basert på sklearn-rammeverket og benytter flere ulike metoder derfra, og kan med det betegnes som en Python-wrapper. Dette gjør oppbyggingen og analysering av maskinlæringsteknikker eksponentielt raskere, og mer effektivt (Moez, 2021). Vi benytter modulen til PyCaret for supervised regresjonsproblemer for å beregne kontinuerlig verdi av ROA, og den bygger prediksjonsmodeller basert på 19 forskjellige maskinlærings- og statistiske metoder, som er presisert i tabell 4.3 (Moez, udatert). Dette gjør at vi får testet dataen på forskjellige metoder, og sikrer at vi velger riktig metode til modellene i oppgaven. PyCaret er også bygd opp for å være transparent, og har flere innebygde funksjonaliteter for å få innsikt i analysens resultat. Blant annet SHAP-integreringen som vi kommer til i neste delkapittel.

Tabell 4.3: Metoder testet i PyCaret

Navn	Modul
Linear Regression	`sklearn.linear_model._base.LinearRegression`
Lasso Regression	`sklearn.linear_model._coordinate_descent.Lasso`
Ridge Regression	`sklearn.linear_model._ridge.Ridge`
Elastic Net	`sklearn.linear_model._coordinate_descent.ElasticNet`
Least Angle Regression	`sklearn.linear_model._least_angle.Lars`
Lasso Least Angle Regression	`sklearn.linear_model._least_angle.LassoLars`
Orthogonal Matching Pursuit	`sklearn.linear_model._omp.OrthogonalMatchingPursuit`
Bayesian Ridge	`sklearn.linear_model._bayes.BayesianRidge`
Passive Aggressive Regressor	`sklearn.linear_model.PassiveAggressiveRegressor`
Huber Regressor	`sklearn.linear_model._huber.HuberRegressor`
K Neighbors Regressor	`sklearn.neighbors._regression.KNeighborsRegressor`
Decision Tree Regressor	`sklearn.tree._classes.DecisionTreeRegressor`
Random Forest Regressor	`sklearn.ensemble._forest.RandomForestRegressor`
Extra Trees Regressor	`sklearn.ensemble._forest.ExtraTreesRegressor`
AdaBoost Regressor	`sklearn.ensemble._weight_boosting.AdaBoostRegression`
Gradient Boosting Regressor	`sklearn.ensemble._gb.GradientBoostingRegressor`
Extreme Gradient Boosting	`xgboost.sklearn.XGBRegressor`
Light Gradient Boosting Machine	`lightgbm.sklearn.LGBMRegressor`
CatBoost Regressor	`catboost.core.CatBoostRegressor`

4.3 Explainable Artificial Intelligence (XAI)

Flere maskinlæringsteknikker, blant annet XGBoost, er svart boks algoritmer. Det betyr at måten algoritmen gjennomfører prediksjoner på ikke er transparente og tolkbare. Explainable Artificial Intelligence (XAI) er den mest effektive praksisen for å endre på dette, og derav sikre at maskinlæringsteknikker er transparente, pålitelige og etiske (Bhattacharya, 2022). Foruten at XAI vil kunne bidra til å avdekke trender hos de ulike variablene i modellen og tilegne ny forståelse for prediksjon av lønnsomhet, så kan det være en mekanisme for å inkludere menneskelig erfaring i metodikken. Dette kan for eksempel benyttes ved feilaktige prediksjoner for å lettere identifisere årsaken til feil, og dermed kunne gjennomføre de nødvendige justeringene på data eller modell slik at den måler det den skal måle.

4.3.1 SHAP

Shapley Additive exPlanation, fra nå av SHAP, er en av de mest populære algoritmene innen XAI, og ble introdusert av Lundberg og Lee (2017). SHAP er modelluavhengig, og i denne oppgaven skal vi benytte dette rammeverket til å finne de viktigste lønnsomhetsdriverne til både XGBoost-modellene og PyCaret-modellen. At vi benytter det samme rammeverket for alle modellene i oppgaven gjør de også enklere å sammenligne. For å tolke SHAP-verdiene, benytter vi beeswarmplott som er integrert i SHAP. For å unngå for mange plott i oppgaven blir beeswarmplottene aggregert for alle periodene i *rolling window*. Bhattacharya (2022) beskriver at SHAP er en god metode for tolking av variabler fordi den i tillegg til å rangere hvor viktige de ulike variablene er, også visualiserer effekten som hver variabel har på prediksjonen.

SHAP er basert på Shapley-verdier, som er et konsept fra spillteorien som ble introdusert av Shapley (1951). SHAP teorien går ut på å sikre en rettferdig fordeling i et samarbeidsspill når man tar hensyn til de individuelle bidragene. Det er en additiv attribusjonsmetode, som vil si at summen av de individuelle bidragene, er det totale bidraget til gruppen. Denne spillteorien kan overføres til maskinlæringsteknikker, hvor vi kan benytte shapley-verdier til å beregne hvor mye den enkelte variabelen bidrar til prediksjonsverdien (Bhattacharya, 2022). En fordel med SHAP er at det er basert på spillteori med teoretisk evidens (Bhattacharya, 2022). Vi kan regne ut shapley-verdien til en variabel ved denne formelen:

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \left(v(S \cup \{i\}) - v(S)\right) \tag{4.17}$$

hvor $\phi_i(v)$ er shapley-verdien som viser det gjennomsnittlige bidraget til variabelen i over alle mulige kombinasjoner av variabelsammensetninger, N er antallet variabler, S er delmengden av variabler som ikke inneholder variabel i, og $v(S)$ er prediksjonsverdien for en gitt delmengde av S. Venstre side av formelen $(v(S \cup \{i\}) - v(S)$ er den marginale effekten av å tillegge variabelen i til variabelsettet S (Bhattacharya,

2022). Som vi kan se av formelen, er en ulempe ved SHAP at det vil være tidsmessig krevende å beregne shapley-verdier på høydimensjonell data. Dette har imidlertid ikke vært en utfordring i denne oppgaven, men noe vi har tatt hensyn til ved utvelgelse av variabler som benyttes i XGBoost-modellen. Bhattacharya (2022) beskriver også at en ulempe ved SHAP, er at den er komplisert å forstå for en person uten teknisk kompetanse.

4.3.2 Partial Dependence Plot (PDP)

SHAP er viktig for å forstå hvilke variabler som påvirker den avhengige variabelen mest, men den viser i mindre grad den funksjonelle sammenhengen mellom de uavhengige variablene og den avhengige variabelen. Partial dependence plots, eller delvis avhengighetsplot (PDP), ble introdusert av Friedman (2001), og kan identifisere lineære, monotone eller komplekse samspill mellom den avhengige variabelen, og de ulike forklaringsvariablene. PDP viser den gjennomsnittlige endringen i predikert verdi når spesifiserte variabler varierer over deres marginale fordeling (Goldstein mfl., 2015). Dette gjøres ved at PDP beregner den marginale effekten som én eller to variabler har på målvariabelen for lønnsomhet.

$$f_S = \mathbb{E}_{x_C}[f(x_S, x_C)] = \int f(x_S, x_C) dP(x_C). \tag{4.18}$$

hvor f_S er funksjonen for partiell avhengighet for en spesifikk variabel x_S, og hvor x_C er verdiene av de resterende variablene i modellen. Ettersom vi ikke vet den sanne f eller den sanne $dP(x_C)$, så estimerer vi den partielle avhengigheten ved følgende formel (Goldstein mfl., 2015):

$$\hat{f}_S = \frac{1}{N} \sum_{i=1}^{N} \hat{f}(x_S, x_{C_i}), \tag{4.19}$$

Fordelen med PDP er at illustrasjonene er enkle å forstå, også for de som ikke har statistisk eller teknisk bakgrunn. Friedman (2001) beskriver at når man har et større antall forklaringsvariabler, er det nyttig å anvende en metode for å måle relevansen til de ulike forklaringsvariablene, slik at man kan redusere antallet prediktorer å analysere. Vi skal derfor lage PDP til de variablene som SHAP utmerker som de viktigste i de prediktive modellene.

Bhattacharya (2022) beskriver også noen begrensninger ved bruk av PDP. Den første er at den antar at ingen av forklaringsvariablene er korrelerte, og at det ikke er noen interaksjoner mellom de ulike variablene. Dette er ikke en riktig antakelse i modellene, og vil generelt sjeldent stemme i reelle datasett. PDP viser kun gjennomsnittlige marginale effekter, og kan derfor ikke vise heterogene effekter. Et eksempel på dette er dersom en variabel har positiv effekt på den avhengige variabelen på halvparten av datasettet, men negativ på den andre halvparten. Effekten i PDP illustrasjonen vil da kunne utligne hverandre, og det kan gi inntrykk av at forklaringsvariabelen ikke har påvirkning på den avhengige variabelen. For å fange opp eventuelle svakheter ved heterogenitet i plottene, skal vi også illustrere ICE-kurver.

4.3.3 Individual conditional expectation (ICE)

Individual conditional expectation plots (ICE-kurver) ble introdusert av Goldstein mfl. (2015), og er en utvidelse av Friedmans PDPs. ICE-kurvene visualiserer også den funksjonelle relasjonen mellom forklaringsvariablene og den avhengige variabelen, men i stedet for å illustrere den gjennomsnittlige endringen i predikert verdi, viser ICE-kurvene endringen for individuelle observasjoner. Det vil si at plottene illustrerer N antall estimerte forventningskurver. ICE-kurvene vil fremheve variasjonen i de aktuelle variablene over hele spekteret av kovariater, som vil bidra til å fange opp eventuelle heterogeniteter (Goldstein mfl., 2015). ICE-kurvene har i likhet med PDP, en antakelse om at ingen av forklaringsvariablene er korrelerte.

Vi benytter Python-biblioteket Scikit-learn for å implementere ICE-kurver og PDP. Ettersom Scikit-learn kun støtter opp om å benytte en variabel av gangen for å kjøre ICE-algoritmen, benyttes enveis-plott både for PDP og ICE (Pedregosa mfl., 2011). Vi mener at å inkludere PDP og ICE vil gi en illustrasjon som er enklere å forstå ved hjelp av menneskelig persepsjon, i tillegg til at det vil gi en dypere forståelse av samspillet mellom de viktigste variablene og lønnsomheten til bedrifter i bygg- og anleggsbransjen.

5 Resultater

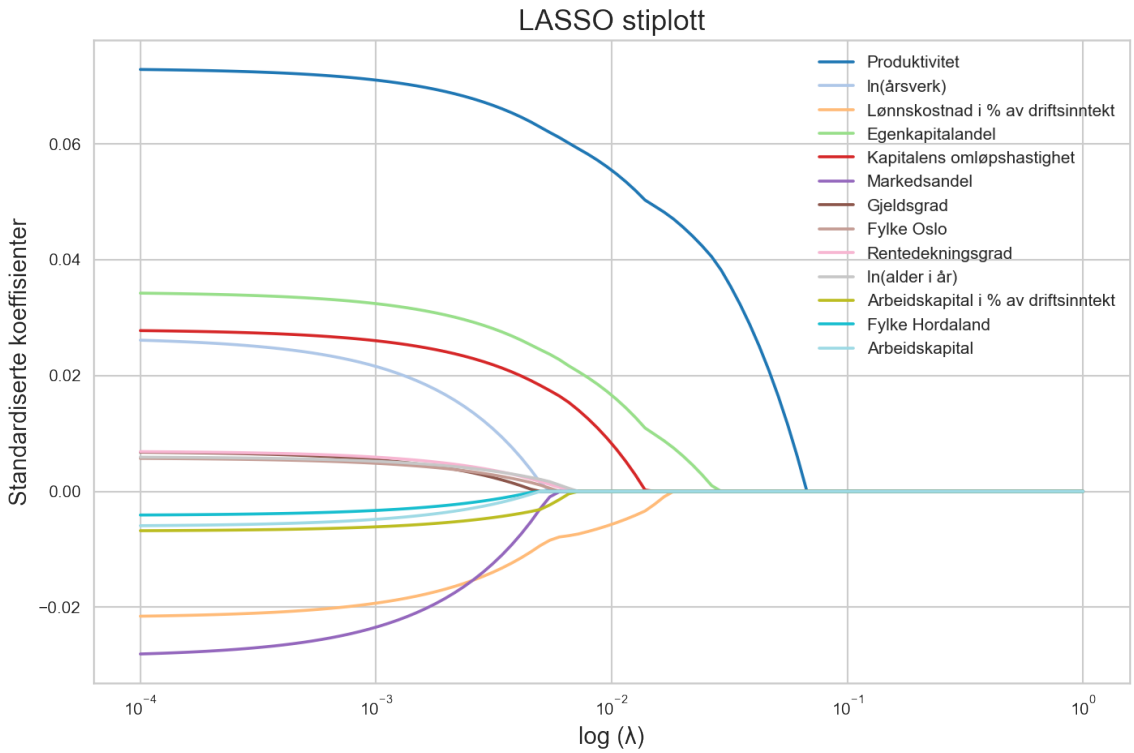
I dette kapittelet presenteres oppgavens resultater, og begynner med resultatene for forklaringsmodellene LASSO, OLS og fixed effects. Disse modellene ser på de uavhengige variablenes sammenheng med lønnsomheten i det samme regnskapsåret. Deretter presenteres resultatene for prediksjon med kontinuerlig ROA med XGBoost og PyCaret, i sammenheng med SHAP, for å finne de viktigste variablene for prediksjon. I denne delen presenteres også de syv viktigste forklaringsvariablene ytterligere ved PDP og ICE-kurver, som analyserer delvis avhengighet mellom disse forklaringsvariablene og ROA. Til slutt presenteres klassifisering av ROA med XGBoost, også i sammenheng med SHAP. For denne modellen skal vi også presentere PDP og ICE-kurver for de fire viktigste forklaringsvariablene. I denne delen presenteres funnene av analysene, før betydningen og den praktiske innvirkningen til funnene blir diskutert i kapittel 6.

5.1 Forklaringsmodeller

5.1.1 LASSO

I denne oppgaven anvendes LASSO, i sammenheng med teori og tidligere forskning, som en metode for variabelseleksjon for både OLS og fixed effects. Variablene som inkluderes i modellen ved høyere verdier av λ antas å ha høyere prediktiv evne, og derav også høyere viktighet. Modellen tar utgangspunkt i 51 variabler, og stiplottet for alle disse variablene finnes i vedlegg B. Dette plottet har 51 grafer, hvor hver graf illustrerer den standardiserte koeffisienten til en variabel for ulike verdier av logaritmen til λ. På grunn av antallet ulike grafer, er dette stiplottet visuelt krevende å tolke. For å redusere antallet variabler i plottet, ble ulike verdier av λ testet for å finne en modell hvor omtrent 10-15 variabler ble inkludert i stiplottet. Når man setter $\lambda = 0.004$ i modellen, er antallet variabler med koeffisienter over 0 redusert til 13 variabler. Dette er derfor de 13 viktigste variablene ifølge LASSO, og vi har valgt å vise et forenklet stiplott som kun inneholder disse 13 variablene.

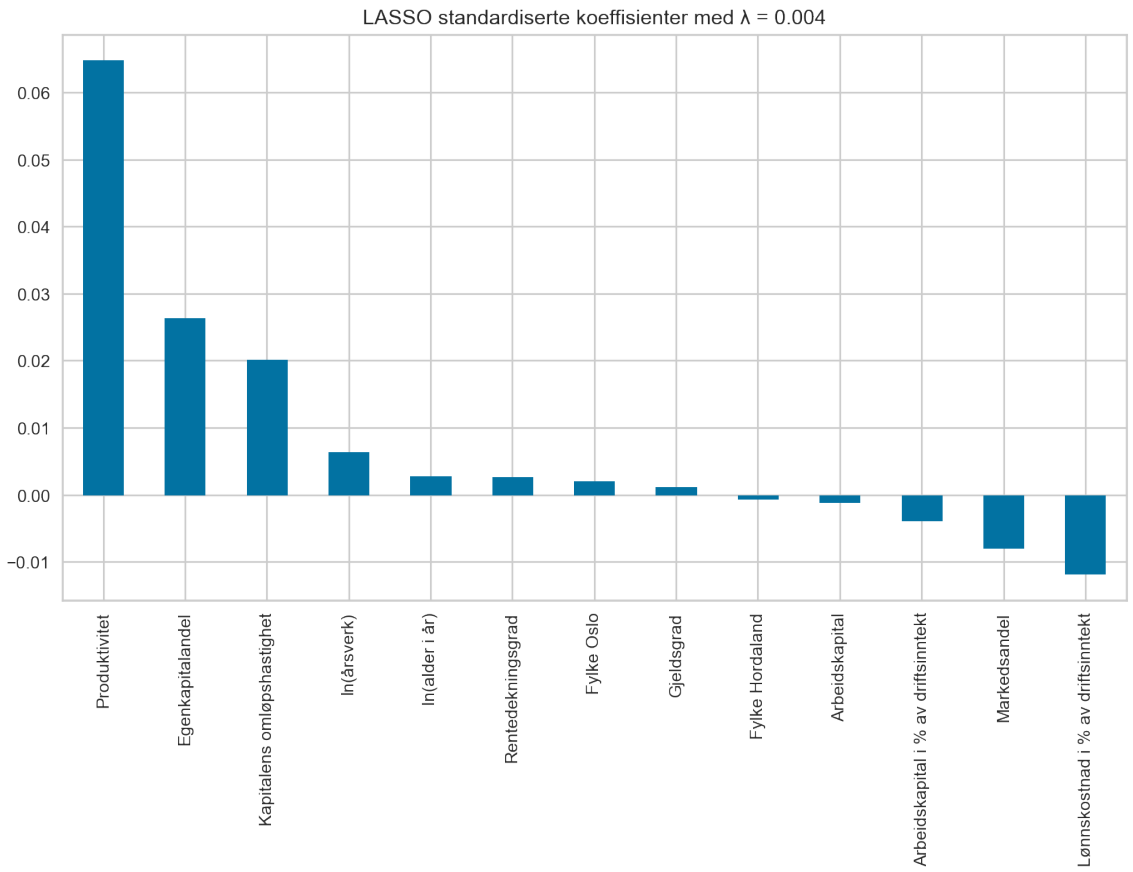
Stiplottet i figur 5.1 har de standardiserte koeffisientene på y-aksen, og logaritmen til straffeparameteren λ på x-aksen. I plottet ser man at produktivitet skiller seg ut som den variabelen hvor den standardiserte koeffisienten går mot 0 for høyest verdi av logaritmen til λ. LASSO viser derfor at produktivitet er den viktigste lønnsomhetsdriveren for norske SMB i bygg- og anleggsbransjen. Deretter finner vi egenkapitalandel, lønnskostnad i % av driftsinntekt og kapitalens omløpshastighet, som henholdsvis andre, tredje og fjerde viktigste lønnsomhetsdriver. De resterende



Figur 5.1: LASSO stiplott for de 13 viktigste variablene

variablene har standardiserte koeffisienter som går mot 0 for relativt lik verdi av logaritmen til λ. Rekkefølgen for de deretter viktigste variablene er henholdsvis ln(alder i år), arbeidskapital i % av driftsinntekt, rentedekningsgrad, markedsandel, Oslo, ln(årsverk), arbeidskapital, gjeldsgrad og Hordaland.

I figur 5.2 presenteres et stolpediagram med standardiserte koeffisienter for de 13 viktigste variablene til LASSO når $\lambda = 0.004$. Ettersom forklaringsvariablene er standardiserte, har de samme skala og kan enklere sammenlignes. I figur 5.2 kan man derfor si at de variablene som har de høyeste stolpene, både negativt og positivt, har størst innflytelse på den avhengige variabelen ROA. Produktivitet har den klart høyeste stolpen, og har en høy positiv standardisert koeffisient. Produktivitet er derfor også den variabelen med størst innflytelse på ROA. Deretter finner man egenkapitalandel og kapitalens omløpshastighet, som begge skiller seg ut med større innflytelse på ROA i positiv grad. Markedsandel og lønnskostnad i % av driftsinntekt har derimot noe større negative standardiserte koeffisienter, og har derfor også større innflytelse på ROA, men i negativ retning. De resterende variablene har ikke like stor påvirkning på ROA, men blir likevel ansett som viktige variabler av LASSO. Det er likevel interessant å bemerke seg retningen og koeffisientstørrelsen til de ulike forklaringsvariablene, ettersom flere av variablene ikke blir med i videre regresjonsanalyse på grunn av mulige problemer med multikollinearitet.



Figur 5.2: Standardiserte koeffisienter LASSO for de 13 viktigste variablene

5.1.2 OLS og fixed effects

Vi har brukt OLS og fixed effects for å se på sammenhengen mellom de uavhengige variablene og ROA. På grunn av mulige problemer med multikollinearitet, har vi ekskludert noen av de uavhengige variablene som LASSO har pekt ut til å være de viktigste. Variablene som er inkludert er produktivitet, lønnskostnad i % av driftsinntekt, egenkapitalandel, kapitalens omløpshastighet, gjeldsgrad og rentedekningsgrad. I tillegg har vi tatt med dummyvariabler for hvert regnskapsår, hvor vi har valgt regnskapsår 2010 som referansegruppe. Tolkningen av alle regnskapsårene blir derfor i forhold til regnskapsår 2010. Resultatene til OLS og fixed effects finnes i tabell 5.1.

Konstantleddene for OLS og fixed effects er negative, og er henholdsvis -0.197 og -0.163. Dette er predikert ROA dersom alle de uavhengige variablene settes lik null, men konstantleddet har ingen praktisk tolkning. Produktivitet har positivt fortegn i begge modellene, og er signifikant på 0.1 % nivå. Koeffisientene til produktivitet er henholdsvis 1.83×10^{-7} for OLS og 2.98×10^{-7} for fixed effects. Videre er lønnskostnad i % av driftsinntekt en viktig forklaringsvariabel som er signifikant på 0.1 % nivå. Koeffisientene er negative for begge modellene, men effekten er betydelig sterkere for

fixed effects. Egenkapitalandel har positive koeffisienter i begge modellene, men er sterkere i fixed effects enn i OLS. Variabelen er signifikant på 0.1 % nivå i begge modellene. Kapitalens omløpshastighet viser i OLS en positiv sammenheng med ROA, mens den har en negativ sammenheng i fixed effects. Variabelen er signifikant på 0.1 % nivå i begge modellene. For gjeldsgrad og rentedekningsgrad er koeffisientene små, men signifikante i begge modellene. Rentedekningsgrad er derimot kun signifikant på et 5 % nivå i fixed effects.

For begge modellene er alle regnskapsårene signifikante på 0.1 % nivå, med unntak av regnskapsår 2011 som kun er signifikant på 5 % nivå. For regnskapsårene er koeffisientene gjennomgående lavere for fixed effects enn for OLS.

Forklaringsgraden til OLS er 24.1 %, og er andelen varians i ROA som blir forklart av de uavhengige variablene i modellen. Forklaringsgraden for fixed effects er på 20.5 %, som er lavere enn for OLS. R^2 within viser til variasjonen i data som oppstår innenfor enheter, i dette tilfelle bedriftene i utvalget, over tid. Denne er på 36.8 % for fixed effects, noe som vil si at 36.8 % av variasjonen i ROA innenfor hver bedrift i utvalget over tid, forklares av de uavhengige variablene. R^2 between ser på hvor godt de uavhengige variablene forklarer forskjeller i ROA mellom enhetene. En R^2 between på 15.9 % vil si at modellen forklarer 15.9 % av variasjonen mellom enhetene, altså bedriftene i utvalget.

Tabell 5.1: Resultater fra OLS og Fixed effects

Variabler	**OLS**	**Fixed effects**
Konstantledd	-0.197***	-0.163***
	(0.007)	(0.007)
Produktivitet	$1.83 \times 10^{-7***}$	$2.98 \times 10^{-7***}$
	(4.53×10^{-9})	(2.93×10^{-9})
Lønnskostnad i % av driftsinntekter	-0.073***	-0.269***
	(0.006)	(0.009)
Egenkapitalandel	0.161***	0.211***
	(0.005)	(0.005)
Kapitalens omløpshastighet	0.027***	-0.007***
	(0.001)	(0.001)
Gjeldsgrad	0.001***	$3.17 \times 10^{-4***}$
	(0.000)	(7.94×10^{-5})
Rentedekningsgrad	$1.387 \times 10^{-8***}$	$2.84 \times 10^{-8*}$
	(1.32×10^{-9})	(1.37×10^{-9})
Regnskapsår 2011	0.009*	0.009*
	(0.004)	(0.004)
Regnskapsår 2012	0.022***	0.013***
	(0.004)	(0.004)
Regnskapsår 2013	0.029***	0.020***
	(0.004)	(0.036)
Regnskapsår 2014	0.035***	0.026***
	(0.004)	(0.004)
Regnskapsår 2015	0.025***	0.017***
	(0.004)	(0.004)
Regnskapsår 2016	0.029***	0.018***
	(0.004)	(0.003)
Regnskapsår 2017	0.024***	0.014***
	(0.004)	(0.003)
Regnskapsår 2018	0.024***	0.017***
	(0.004)	(0.003)
Regnskapsår 2019	0.023***	0.018***
	(0.004)	(0.003)
Regnskapsår 2020	0.027***	0.020***
	(0.004)	(0.003)
Regnskapsår 2021	0.022***	0.018***
	(0.004)	(0.003)
Regnskapsår 2022	0.025***	0.035***
	(0.004)	(0.003)
R^2	0.241	0.205
R^2 within		0.368
R^2 between		0.159

Standardfeil i parentes
$* \ p < 0.05$, $** \ p < 0.01$, $*** \ p < 0.001$

5.2 Prediksjon av ROA

I dette underkapittelet presenteres resultatene til XGBoost og PyCaret for prediksjon av kontinuerlig ROA ett år frem i tid. Først presenteres evalueringsmålene, og deretter går vi i dybden på variablene ved bruk av SHAP, PDP og ICE-kurver. Betydningen av funnene i denne delen vil videre diskuteres i kapittel 6.

5.2.1 XGBoost prediksjon av ROA

I tabell 5.2 ser vi evalueringsmålene for alle periodene i *rolling window* presentert med året som er testsettet for perioden. Vi skal i det følgende presentere alle evalueringsmålene med fokus på testsettet. For RMSE har treningssettene variasjon fra 0.125 til 0.137 mellom periodene, og MSE fra 0.016 til 0.019. Testsettene har RMSE fra 0.136 til 0.155, og MSE fra 0.018 til 0.024. Modellen har noe lavere RMSE og MSE for de tidligste periodene, i kontrast til de siste periodene. Det siste året med testsett for regnskapsår 2022 skiller seg noe ut med høyere verdi for RMSE og MSE, og modellen har derfor større kvadrerte feil i denne perioden.

MAE presenterer gjennomsnittet av de absolutte feilene mellom prediksjon og faktisk verdi. Treningssettene har MAE fra 0.092 til 0.099, og testsettene har MAE fra 0.098 til 0.113. Forklaringsgraden (R^2) viser hvor mye av variansen av ROA som forklares av de uavhengige variablene. Treningssettene har R^2 fra 0.307 til 0.374, og testsettene har R^2 fra 0.220 til 0.270. Det vil si at forklaringsvariablene i modellen forklarer mellom 22 % og 27 % av variansen i ROA ett år frem i tid.

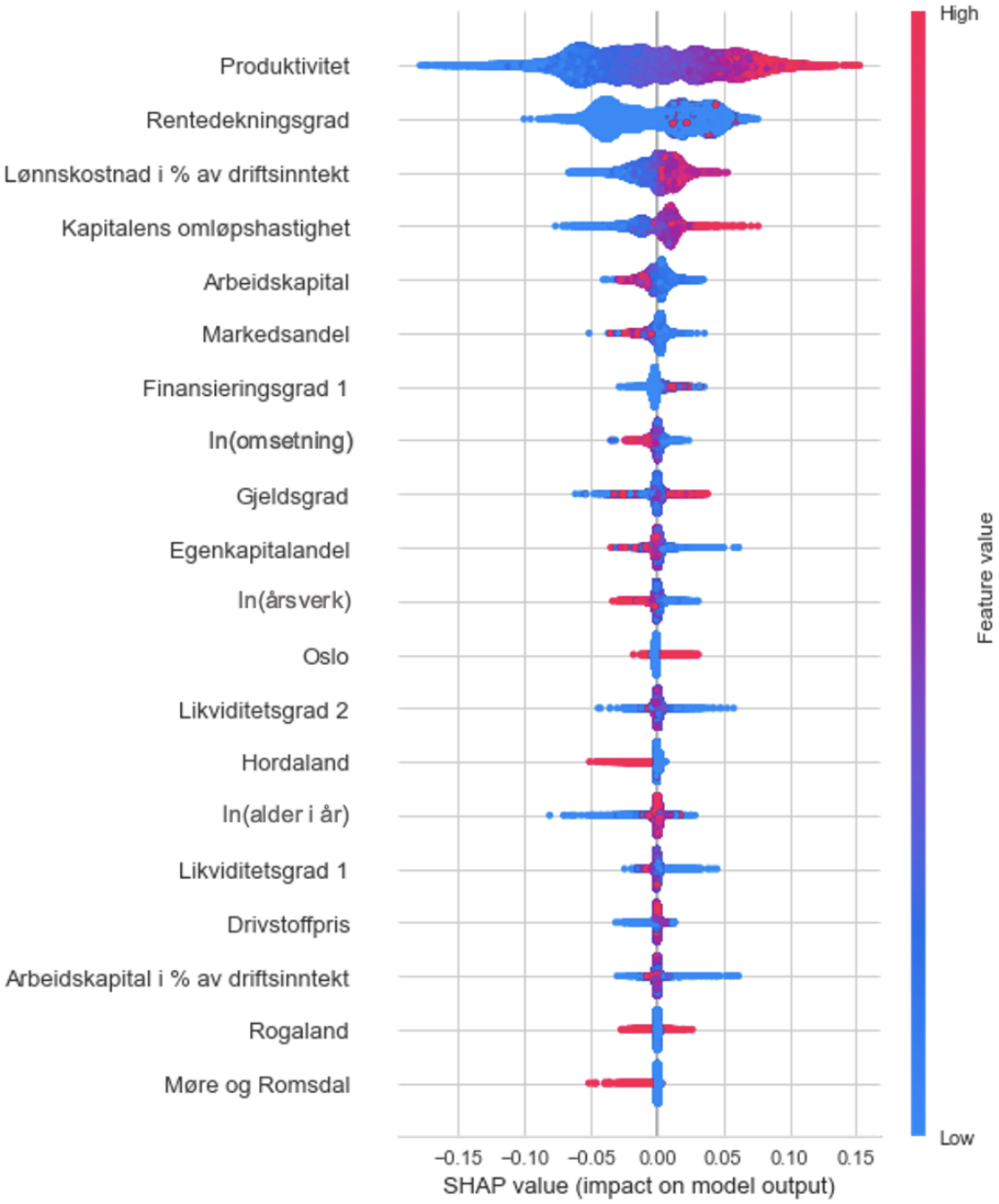
MAPE og MdAPE viser gjennomsnitt og median absoluttprediksjonsfeil i prosent. Treningssettene har MAPE fra 164 til 2366, og MdAPE fra 47 til 54. Testsettene har MAPE fra 147 til 7180, og MdAPE fra 51 til 58. For testsett 2014 viser både MAPE og MdAPE bedre verdier for testsettet enn treningssettet. For MAPE gjør testsettet i 2017, 2019, 2020, 2021 og 2022 det også bedre enn treningssettet. Dette er noe uventet.

For evalueringsmålet som viser 10 % toleransegrense, har treningssettene verdi på mellom 0.25 og 0.26, og testsettene mellom 0.23 og 0.27. Det vil si at for periodene så klarer XGBoost-modellen å predikere mellom 23 % og 27 % av observasjonene riktig med en 10 % toleransegrense. Her er andelen for testsettet høyere enn treningssettet for to perioder i *rolling window*.

Tabell 5.2: Evalueringsmål over alle periodene for prediksjon av ROA i XGBoost

	2014	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**	**2022**
RMSE treningssett	0.125	0.126	0.127	0.128	0.131	0.127	0.129	0.137	0.136
RMSE testsett	0.138	0.136	0.137	0.138	0.137	0.148	0.148	0.147	0.155
MSE treningssett	0.016	0.016	0.016	0.016	0.018	0.016	0.017	0.019	0.018
MSE testsett	0.020	0.018	0.019	0.019	0.019	0.022	0.022	0.022	0.024
MAE treningssett	0.092	0.092	0.093	0.093	0.095	0.093	0.094	0.099	0.098
MAE testsett	0.101	0.098	0.100	0.100	0.099	0.107	0.107	0.106	0.113
R^2 treningssett	0.363	0.374	0.358	0.352	0.320	0.353	0.364	0.307	0.333
R^2 testsett	0.269	0.238	0.244	0.242	0.270	0.232	0.236	0.228	0.220
MAPE treningssett	303	164	305	375	394	2490	2366	2269	1983
MAPE testsett	156	659	443	291	7180	214	147	190	450
MdAPE treningssett	53	50	47	48	51	51	52	54	53
MdAPE testsett	51	51	51	55	58	56	55	57	57
10 % toleranse treningssett	0.26	0.25	0.25	0.25	0.25	0.26	0.26	0.26	0.26
10 % toleranse testsett	0.23	0.25	0.25	0.27	0.26	0.25	0.24	0.25	0.26

SHAP Beeswarmplott



Figur 5.3: Aggregert SHAP beeswarmplott for XGBoost prediksjon av ROA for alle periodene i *rolling window*

For å identifisere og rangere forklaringsvariablene i modellen benyttes SHAP beeswarmplott som illustrert i figur 5.3. Dette plottet er aggregert for alle ni periodene i *rolling window*, og er sortert etter gjennomsnittlig absolutt SHAP-verdi på y-aksen. Dette tolkes som at viktigheten til variablene har synkende rekkefølge, og det viser at produktivitet, rentedekningsgrad, lønnskostnad i % av driftsinntekt og kapitalens omløpshastighet er de viktigste forklaringsvariablene i den rekkefølgen. Rangeringen har likheter med forklaringsmodellene tidligere i oppgaven. Hver av variablene har beeswarmplott som indikerer omfang, tetthet og hvilken retning variablene påvirker ROA. Prikker illustrerer individuelle datapunkter uten overlapping. Bredden på prikkene indikerer omfang av datapunkter langs aksen, og ved høy tetthet av observasjoner med samme SHAP-verdi, er prikkene stablet vertikalt. Med retning menes det hvorvidt forklaringsvariabelen påvirker lønnsomhet negativt eller positivt, og kan sammenlignes med om en koeffisient i regresjonsmodeller har pluss eller minus som fortegn. Videre i oppgaven refereres lave og høye verdier i samsvar med fargekodingen til høyre i beeswarmplottet. Observasjoner med gjennomsnittlig verdi for den aktuelle variabelen vil illustreres med lilla farge, under gjennomsnittet illustreres med blå, og over gjennomsnittet illustreres med rød. Dette vil derfor ha sammenheng med den deskriptive statistikken for de uavhengige variablene som finnes i datakapittelet, tabell 3.4.

Produktivitet er den viktigste variabelen i plottet, og er bred med SHAP-verdier som varierer fra omtrent -0.18 til 0.16. Det gjør produktivitet til den variabelen som har klart størst variasjon av SHAP-verdier. Tettheten viser derimot at flest observasjoner har SHAP-verdi mellom -0.075 og 0.075, noe som viser at for de fleste observasjonene i datasettet, vil produktivitet kunne påvirke ROA både negativt og positivt. Produktiviteten har lave (blå) verdier på venstresiden av plottet, før det gradvis går over til middels verdier (lilla) for SHAP-verdier fra 0 til omtrent 0.05, og deretter går over til høye verdier (rød). Denne gradvise overgangen med tydelige farger av både blå, lilla og rød, kan tyde på at produktiviteten har en lineær sammenheng med ROA. Høyere verdier av produktivitet indikerer høyere ROA, og lavere verdier av produktivitet indikerer lavere ROA.

Rentedekningsgrad er også en relativt bred variabel med stort omfang. Det er høy tetthet av observasjoner både til venstre i grafen rundt SHAP-verdi -0.05 til -0.025, og igjen fra SHAP-verdi rett over 0 til 0.05. Observasjonene i plottet har flest lave verdier for rentedekningsgrad, som man ser ved at nesten hele plottet er blått. Det er derimot en og annen rød prikk på høyresiden, som kan tyde på at høye verdier for rentedekningsgrad kan indikere høyere ROA, men at lave verdier verken tydelig indikerer høy eller lav ROA.

For lønnskostnad i % av driftsinntekt observeres det, i likhet med produktivitet, en noe gradvis overgang i farger med blå på venstresiden, til lilla, og deretter rød på høyresiden. Dette antyder også en lineær sammenheng med ROA. Tettheten er størst noe til høyre for nullpunktet i SHAP-verdi, som kan tyde på at for de fleste observasjonene, har lønnskostnad i % av driftsinntekt en moderat til positiv effekt på ROA. Lignende fargeovergang finnes også for kapitalens omløpshastighet, som tyder på linearitet. Imidlertid er det en ulikhet i fargedistribusjonen, hvor lønnskostnad

i % av driftsinntekt har en jevnere fordeling med noe mer rødt, mens kapitalens omløpshastighet har dominerende grad av lilla og røde prikker. Denne tettheten ligger noe til høyre i plottet, som vil si at de fleste observasjonene i datasettet har høye verdier av kapitalens omløpshastighet, og at det påvirker ROA i positiv retning.

Arbeidskapital og markedsandel har lignende plott med hensyn til omfang, tetthet, retning og viktighet. I plottene observeres lavere blå verdier på høyre side, som kan tyde på at lave verdier for arbeidskapital og markedsandel er assosiert med økt lønnsomhet, målt ved ROA. Til tross for dette, har begge variablene både røde og blå prikker på venstre side i plottene som antyder mer komplekse sammenhenger mellom disse variablene og ROA. Begge variablene har også en høyere tetthet i blå farge nær og til høyre for nullpunktet på x-aksen. Det indikerer at for majoriteten av observasjoner i datasettet, har arbeidskapital og markedsandel lave verdier som påvirker ROA i positiv retning. Det er også likheter med finansieringsgrad 1, men motsatt vei. Her er de lave blå verdiene til venstre i plottet, som kan tyde på at lave verdier av finansieringsgrad 1 er assosiert med lavere ROA. Til høyre i plottet finnes både røde og blå prikker som viser en kompleks sammenheng mellom finansieringsgrad 1 og ROA. Tettheten er konsentrert til venstre for nullpunktet på x-aksen som antyder at de fleste observasjonene har lave verdier, og at dette påvirker ROA i negativ retning.

For de resterende variablene som er presentert i figur 5.3, er tettheten konsentrert rundt nullpunktet på x-aksen, som kan antyde at for de fleste observasjonene har disse variablene liten til ingen påvirkning på prediksjonen av ROA ett år frem i tid. Videre analyse vil derfor fokusere på omfang og retning for disse variablene i den påfølgende delen.

I beeswarmplottet viser ln(omsetning), gjeldsgrad og egenkapitalandel ingen lineære trekk. Til høyre for ln(omsetning), er det kun blå prikker, som kan bety at lave verdier indikerer høyere ROA. De røde prikkene er lokalisert til venstre, som forsterker denne indikasjonen, selv om det også er noen blå prikker helt til venstre. Gjeldsgrad har et noe bredere plott enn flere av de resterende variablene, og har derfor større omfang. Det er mest rødt til høyre som tyder på høy ROA for høye verdier av gjeldsgrad, selv om det også her er både røde og blå prikker til venstre. Egenkapitalandel har til høyre kun blå prikker, som kan indikere at lave verdier av egenkapitalandel er assosiert med gunstig ROA. På venstresiden er det en god blanding av rødt og blått. Beeswarmplottet viser derfor at det er komplekse sammenhenger mellom ROA og henholdsvis ln(omsetning), gjeldsgrad og egenkapitalandel.

Ln(årsverk) synes å vise en noe lineær sammenheng med ROA. Det er en fargeovergang med røde høye verdier av ln(årsverk) til venstre, lilla i midten, og blå til høyre. Det indikerer negativ linearitet, hvor et større antall årsverk er assosiert med lavere ROA, og færre årsverk er assosiert med høyere ROA.

Når det gjelder lokaliseringsvariablene for fylker, er Oslo, Hordaland, Rogaland og Møre og Romsdal blant de 20 viktigste variablene for å predikere ROA. Siden det er dummyvariabler, vil en lav verdi (blå) bety at bedriften ikke er lokalisert i det gitte fylket, og en høy verdi (rød) bety at bedriften er lokalisert i fylket. Plottet viser at bedrifter med beliggenhet i Oslo fylke ofte vil ha høyere ROA, mens beliggenhet

i Hordaland og Møre og Romsdal kan tyde på lavere ROA. Rogaland har derimot større variasjon i beeswarmplottet, og man kan ikke tolke en bestemt retning.

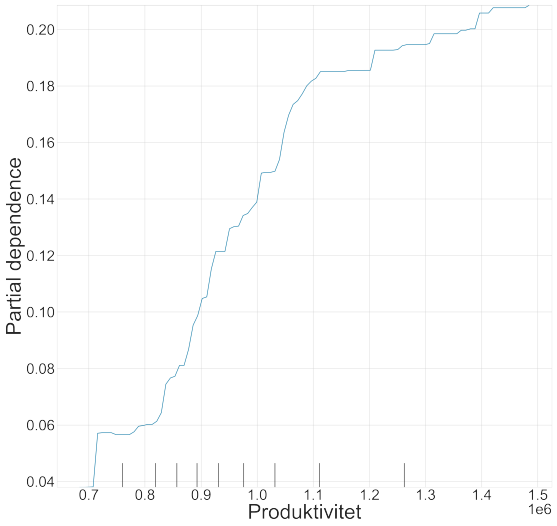
For variablene likviditetsgrad 2, ln(alder i år), likviditetsgrad 1, drivstoffpris og arbeidskapital i % av driftsinntekt, er det nesten utelukkende blå farge i bredden, og rødt i midten der SHAP-verdien er omtrent på nullpunktet. Det kan derfor indikere at lave verdier har større påvirkning på lønnsomheten enn høye verdier. Likviditetsgrad 2 er noe bred, og har verdier både til høyre og venstre. Det tyder på at lave verdier av likviditetsgrad 2 kan påvirke lønnsomheten både positivt og negativt, men at høye verdier påvirker svært lite. Ln(alder i år) er også noe bred, men mesteparten av plottet ligger til venstre. Det kan tyde på at dersom bedriften har eksistert i mange år påvirker det ikke lønnsomheten særlig, men om bedriften er ung vil det kunne påvirker lønnsomheten negativt. For likviditetsgrad 1 er bredden for det meste til høyre som kan tyde på at lave verdier kan predikere høyere lønnsomhet, og høye verdier påvirker i mindre grad eller potensielt negativt fordi det er noen røde prikker rett til venstre for nullpunktet. Drivstoffpris er den eneste makrovariabelen som befinner seg blant de 20 viktigste variablene. Lave verdier er stort sett til venstre som kan tyde på at lav drivstoffpris predikerer lavere ROA, men høy drivstoffpris verken påvirker ROA positivt eller negativt. Arbeidskapital i % av driftsinntekt har for det meste blå prikker til høyre som indikerer at lav prosent arbeidskapital i forhold til driftsinntekt, vil indikere høyere ROA. Her er det også blå verdier til venstre, som gjør det krevende å se en tydelig retning for variabelen.

I beskrivelsen av figur 5.3 har vi sett på de 20 viktigste variablene for prediksjon av ROA ved anvendelse av XGBoost, samt disse variablenes omfang og retning. Det er noe lettere å tolke de variablene som har lineær sammenheng med ROA, men XGBoost identifiserer flere komplekse sammenhenger som vi ytterligere skal presentere ved PDP og ICE-kurver for de syv viktigste variablene. Disse er valgt ut fordi variablene har tetthet av observasjoner utenfor nullpunktet av SHAP-verdier. Disse plottene er i likhet med beeswarmplottet aggregert for alle periodene i *rolling window*.

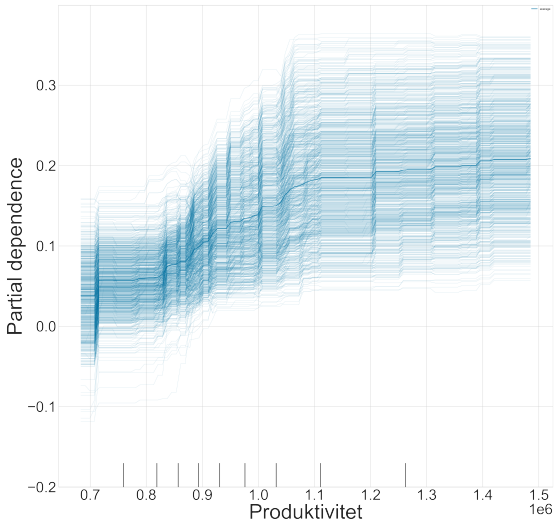
PDP og ICE-plott

I figur 5.4 ser vi den delvise avhengigheten mellom produktivitet og ROA. Dette plottet forsterker det som ble observert i beeswarmplottet til produktivitet, ved at det er en tilnærmet positiv lineær sammenheng mellom produktivitet og ROA. PDP gir et ytterligere bilde av grafens stigning. Fra starten av figuren ser man at det er svært lite økning i ROA med produktivitet mellom 700 000 og 800 000. PDP viser også at den bratte lineære kurven ikke fortsetter i uendeligheten. Det er en bratt positiv linearitet frem til omtrent 1 100 000 i produktivitet, hvor stigningen reduseres. De vertikale linjene på x-aksen viser distribusjonen av data til funksjonen. I PDP er de vertikale linjene tettere for produktivitet mellom omtrent 750 000 og 1 000 000. Den bratte stigningen i starten av plottet er derfor basert på flere observasjoner sammenlignet med når den flater ut i slutten av plottet. Dette impliserer at starten av plottet gir mer pålitelige resultater. I figur 5.5 kan man til dels se dette ved at ICE-

kurvene har mindre variasjon i starten av plottet, og at variasjonen øker for høyere verdier av produktivitet. Dette kan også indikere at det er individuelle forskjeller på helningen i plottet, men ICE-kurvene bekrefter antakelsen om at det er en tilnærmet positiv lineær sammenheng mellom produktivitet og ROA.



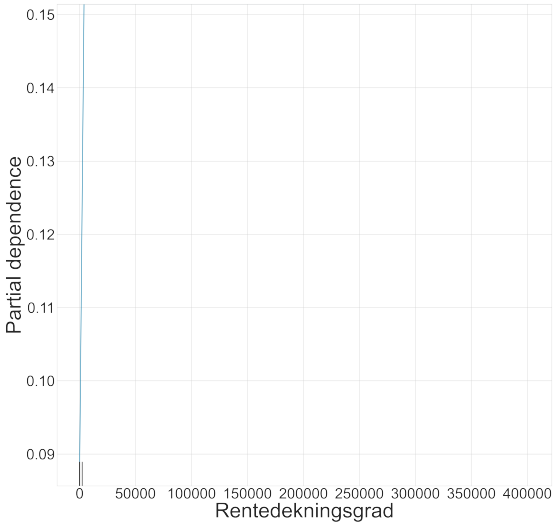
Figur 5.4: PDP for produktivitet



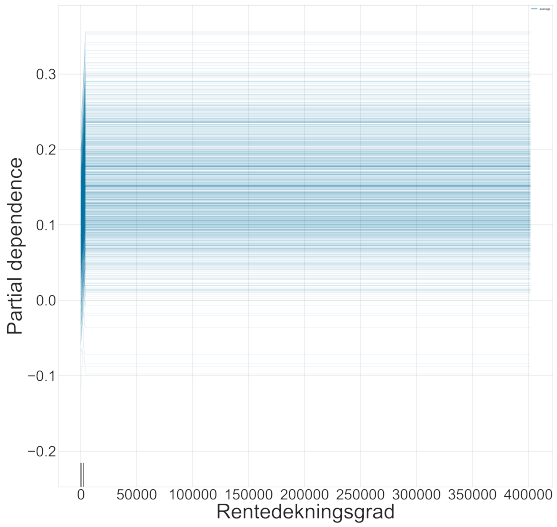
Figur 5.5: ICE-kurver for produktivitet

For rentedekningsgrad ser vi en derimot en annen trend. I figur 5.6 har grafen nesten vertikal linje, som antyder stor endring i ROA ved økning av lave verdier av rentedekningsgrad. Majoriteten av data befinner seg noe over null i rentedekningsgrad, som er illustrert med de vertikale linjene på x-aksen. Tabell 3.4 viser at medianen til rentedekningsgrad er 19.59, mens gjennomsnittet er 71 617, noe som indikerer stor spredning i verdiene. ICE-kurvene i figur 5.7 viser den samme signifikante økningen helt til venstre i plottet, men også at ICE-kurvene flater ut for høyere verdier av rentedekningsgrad. Det indikerer at rentedekningsgrad har positiv effekt på ROA frem til en viss terskelgrense. Etter dette har rentedekningsgrad liten til ingen påvirkning på ROA. Variasjonen i ICE-kurvene både over og under gjennomsnittet, styrker antakelsen om at det finnes en terskelgrense, men at verdien på denne grensen varierer for de individuelle observasjonene. Det er krevende å tolke hva terskelverdien til PDP er, som kan bety at det er noen uteliggere i dataen, selv etter winsorizing.

Figur 5.8 viser den delvise avhengigheten mellom lønnskostnad i % av driftsinntekt og ROA. I likhet med beeswarmplottet i 5.3, viser det også en tilnærmet positiv lineær sammenheng med ROA. De vertikale linjene på x-aksen viser at distribusjonen er jevnt fordelt i plottet der de fleste observasjonene har lønnskostnad i % av driftsinntekt mellom 20 % og 50 %. Det er omtrent ingen økning i ROA for verdier mellom omtrent 22 % og 27 %. Deretter er det derimot bratt stigning dersom lønnskostnad i % av driftsinntekt øker fra omtrent 28 % til 30 %. Ifølge PDP vil forhold over 50 % ikke føre til ytterligere økninger i ROA. ICE-kurvene i figur 5.9 viser stort sett det samme som PDP, men det er generelt mer variasjon i de individuelle observasjonene

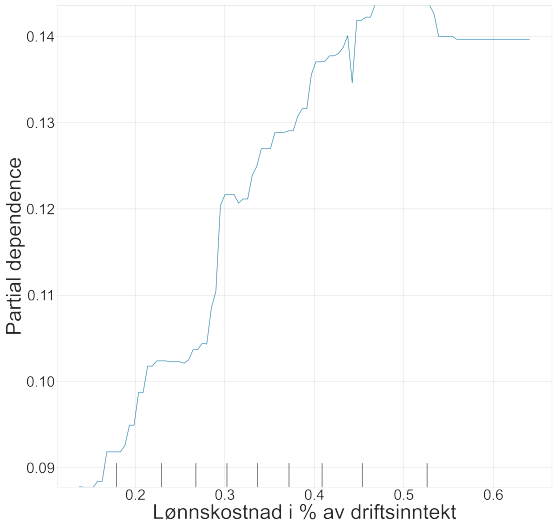


Figur 5.6: PDP for rentedekningsgrad

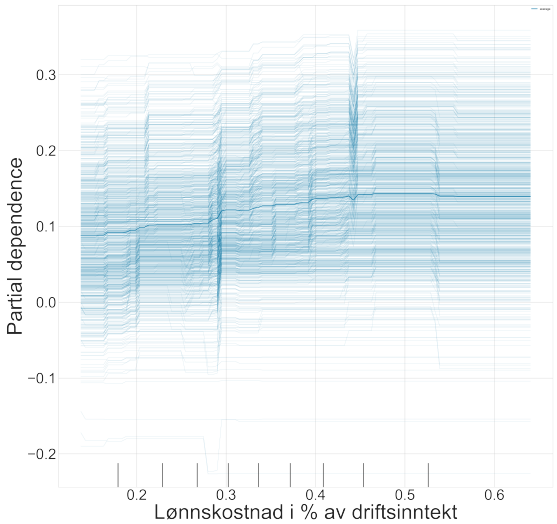


Figur 5.7: ICE-kurver for rentedekningsgrad

for denne variabelen enn det var for både produktivitet og rentedekningsgrad. Noen av ICE-kurvene viser til og med at lønnskostnad i % av driftsinntekt kan ha negativ påvirkning på ROA. Denne variasjonen er jevn i hele ICE-plottet, og kan bety at effekten til denne variabelen på ROA kan være avhengig av andre forklaringsvariabler i modellen.



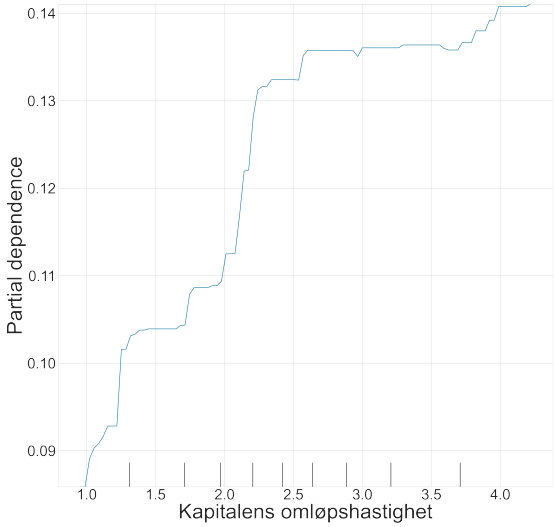
Figur 5.8: PDP for lønnskostnad i % av driftsinntekt



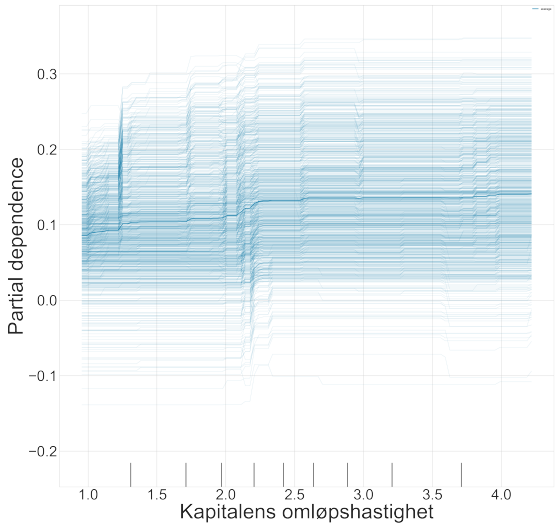
Figur 5.9: ICE-kurver for lønnskostnad i % av driftsinntekt

Kapitalens omløpshastighet viser også en tilnærmet positiv lineær sammenheng med ROA. I figur 5.10 finner vi den delvise avhengigheten mellom kapitalens omløpshastighet og ROA, gitt at de andre variablene er stabile. Fra en verdi på omtrent

1 og fram til 2.5, er det en bratt stigningskurve der høyere verdi av kapitalens omløpshastighet gir høyere ROA. Økningen er særlig bratt fra 2 til 2.5, hvor kapitalens omløpshastighet har gjennomsnittlig påvirkning på ROA med ca. 0.03. Verdier fra 2.5 og utover har moderat til svak økning av ROA. De vertikale linjene på x-aksen viser at observasjonene i datasettet er distribuert relativt jevnt i hele plottet. Ved å se på ICE-kurvene i figur 5.11, ser vi også at det er variasjon for individuelle observasjoner. Et fåtall observasjoner viser at kapitalens omløpshastighet har negativ påvirkning på ROA, men de aller fleste viser positiv sammenheng. Vi ser at det er tettere linjer over 0 på y-aksen. Begge figurene viser derfor at høyere verdier av kapitalens omløpshastighet har sammenheng med høyere prediksjon av ROA opp til en verdi på omtrent 2.5.



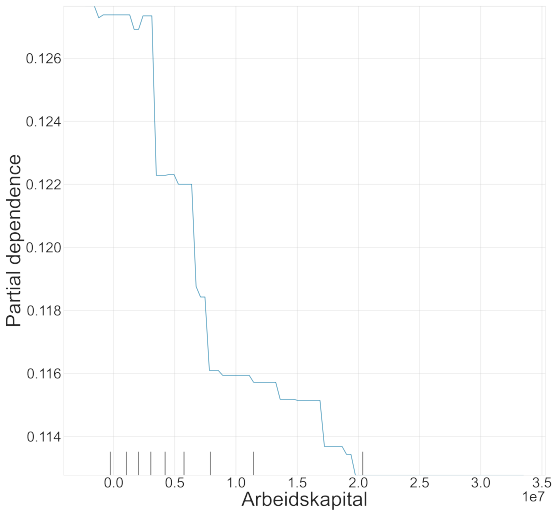
Figur 5.10: PDP for kapitalens omløpshastighet



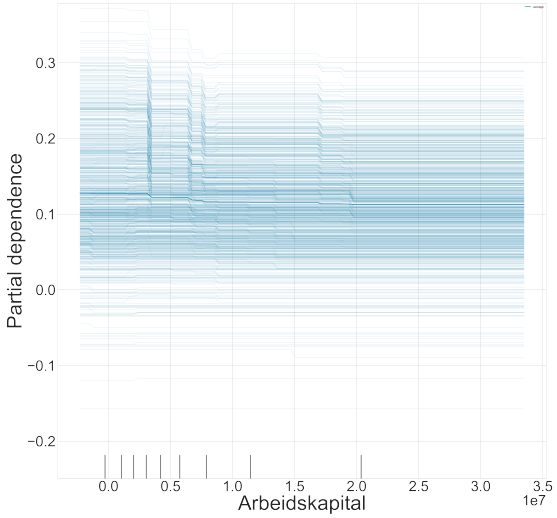
Figur 5.11: ICE-kurver for kapitalens omløpshastighet

I figur 5.12 finner vi PDP for arbeidskapital, som i likhet med figurene over, viser delvis avhengighet. I denne grafen finner vi at høyere verdier for arbeidskapital har en tilnærmet negativ sammenheng med ROA. Det er en nedadgående trend som inkluderer noen bratte partier. Det er et bratt parti som kan minne om et fossefall for arbeidskapital på omtrent 3 500 000 til 4 000 000 der ROA gjennomsnittlig synker med omtrent 0.004. Et tilsvarende «fossefall» finnes ved arbeidskapital på omtrent 6 000 000 - 7 000 000, hvor ROA synker med omtrent 0.006. Deretter følger en moderat til svak negativ lineær kurve. Fra de vertikale linjene på x-aksen ser man at de fleste observasjonene har arbeidskapital på mellom 0 og 10 000 000. PDP viser derfor at arbeidskapital har negativ effekt på ROA for de fleste observasjonene. ICE-kurvene for arbeidskapital, i figur 5.11, viser variasjon i individuell respons. Det er både ICE-kurver som har negativt stigningstall, i likhet med PDP, men også ICE-kurver som er relativt flate som tilsier at arbeidskapital ikke påvirker ROA for disse observasjonene. ICE-kurvene viser at man ikke nødvendigvis kan generalisere at arbeidskapital påvirker ROA i negativ retning, selv om det er tilfellet for flere

observasjoner.



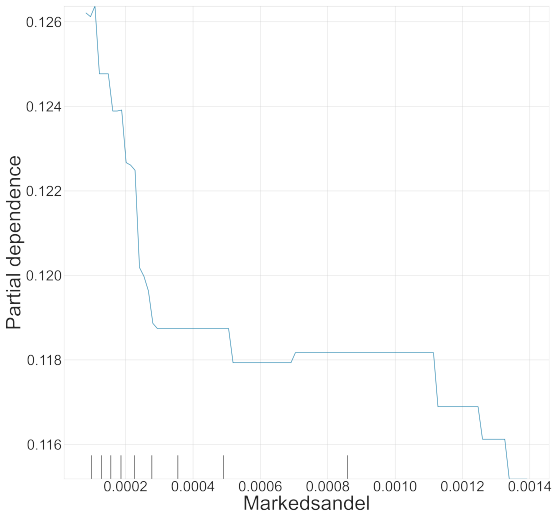
Figur 5.12: PDP for arbeidskapital



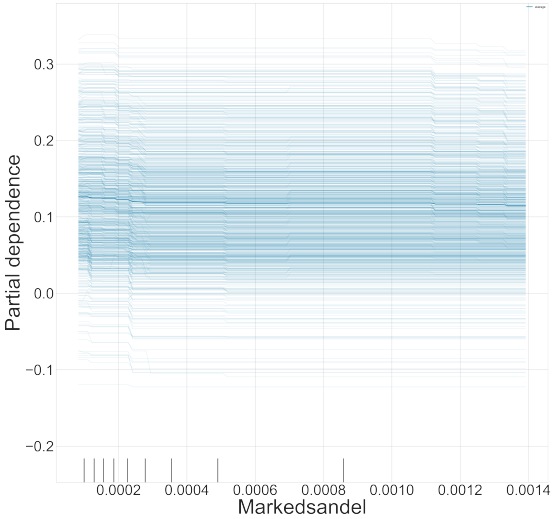
Figur 5.13: ICE-kurver for arbeidskapital

Markedsandelen er ytterligere presentert ved delvis avhengighetsplott i figur 5.14. I beskrivelsen av SHAP-plottet i figur 5.3, så man at arbeidskapital og markedsandel hadde svært like beeswarmplott. I PDP og ICE-kurvene ser man imidlertid at de er noe ulike. Markedsandel har, i likhet med arbeidskapital, en nedadgående trend. Vi ser et «fossefall» i starten av figur 5.14 som viser at en økning i lave verdier av markedsandel indikerer lavere prediksjon av ROA, og lønnsomhetsmålet synker med omtrent 0.007 i dette partiet. Den negative helningen flater derimot ut for markedsandel på omtrent 0.00025, og dette partiet varer lenger sammenlignet med arbeidskapital i figur 5.12. Markedsandel mellom 0.00025 til 0.0011 har liten til moderat påvirkning på ROA, etterfulgt av ytterligere nedadgående trend. Fra de vertikale linjene på x-aksen ser man at de fleste observasjonene befinner seg i den nedre delen av skalaen for markedsandel, fra omtrent 0.0001 til 0.0003. Dette stemmer med tabell 3.4 som viser at medianen for markedsandel er 0.000196. ICE-kurvene i figur 5.15 viser signifikant variasjon i y-verdiene for markedsandel lik null i de individuelle observasjonene. De fleste kurvene har nedadgående trend i starten av plottet, noe som styrker indikasjonen på at en økning i lave verdier av markedsandel kan bety lavere predikert ROA.

Figur 5.16 illustrerer den delvise avhengigheten mellom finansieringsgrad 1 og ROA. Figuren viser at økning av finansieringsgrad 1 har en positiv effekt på lønnsomhet, men at effekten foregår i bolker. Dette stemmer med beeswarmplottet for finansieringsgrad 1 i 5.3 som viser at det er en kompleks sammenheng. PDP i figur 5.16 viser at en økning fra null gir høyere ROA, og at påvirkningen deretter flater ut frem til finansieringsgrad 1 på omtrent 2000. Deretter er det en bratt økning for verdier noe over 2000. Når finansieringsgrad 1 når omtrent 3000 vil derimot høyere verdier for finansieringsgrad 1, ikke lenger ha effekt på prediksjon av ROA, som ses ved at grafen flater ut. Dette er likevel en ekstremt høy verdi for finansieringsgrad 1,

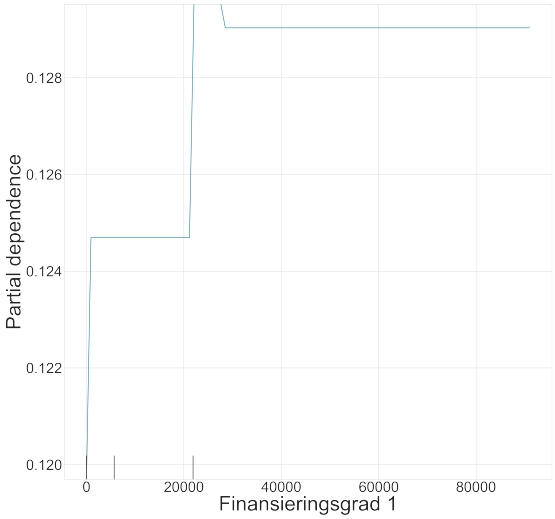


Figur 5.14: PDP for markedsandel

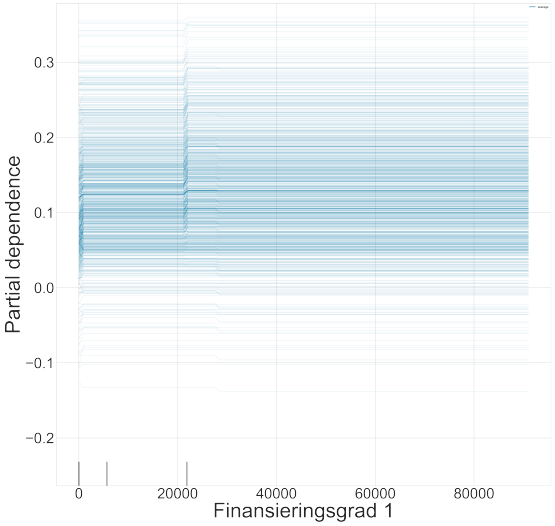


Figur 5.15: ICE-kurver for markedsandel

og vil være relevant i svært få tilfeller. Det er generelt ekstremt høye verdier for finansieringsgrad 1 i PDP, som indikerer uteliggere i dataen. I figur 5.17 viser ICE-kurvene også at det er økning i bolker, men at det er signifikant variasjon i de individuelle kurvene. Noen ICE-kurver er horisontale som antyder at for disse observasjonene, har finansieringsgrad 1 ingen påvirkning på ROA.



Figur 5.16: PDP for finansieringsgrad 1



Figur 5.17: ICE-kurver for finansieringsgrad 1

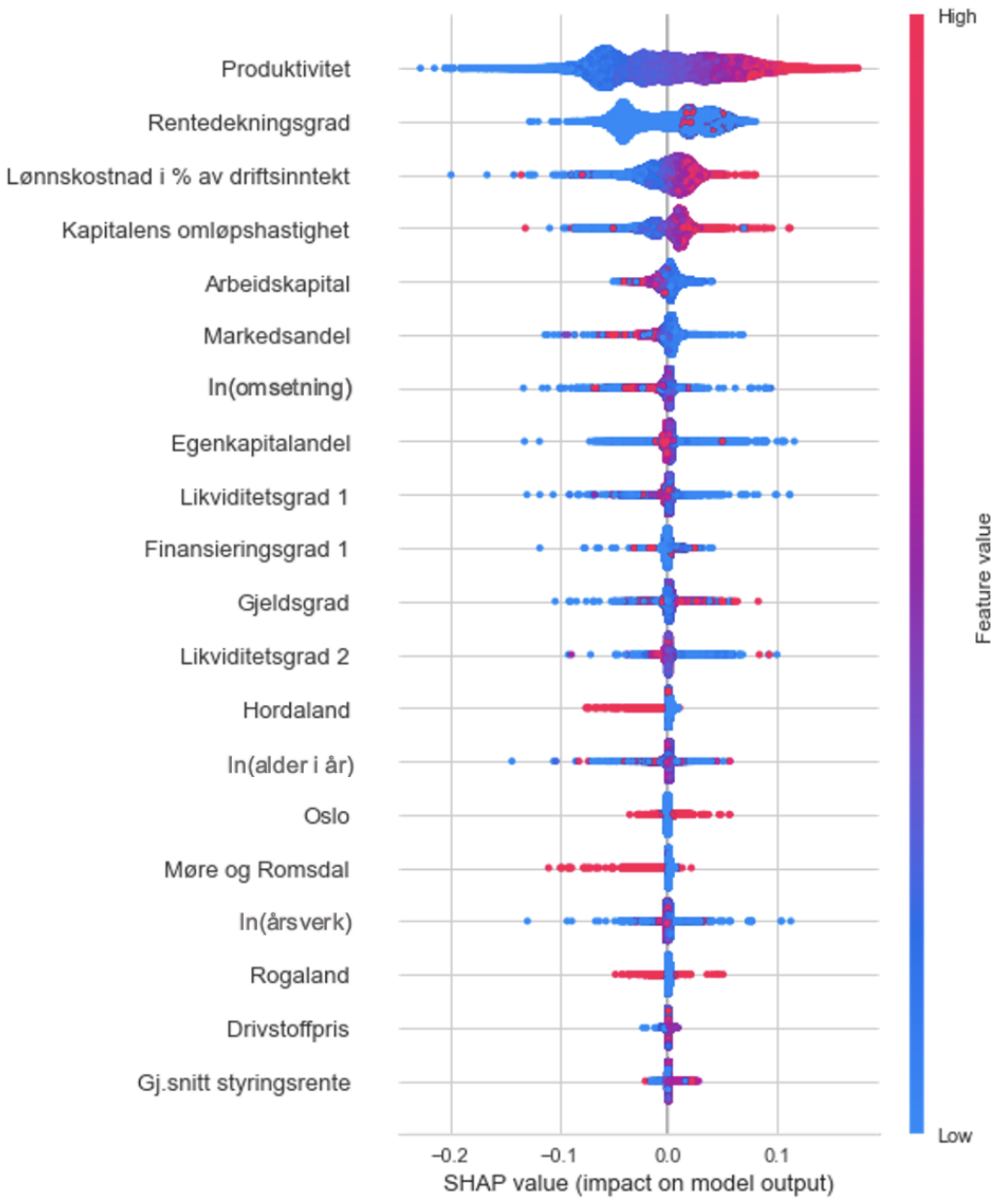
5.2.2 PyCaret prediksjon av ROA

Videre i oppgaven presenteres resultatet fra PyCaret. PyCaret tester for flere metoder, men foreslo metoden Gradient Boosting Regressor i alle periodene. Denne metoden har likhetstrekk med XGBoost, fordi den også er basert på beslutningstrær og boosting, men bruker sklearn-biblioteket i Python. Det er derfor ikke overraskende at resultatene er relativt like som for XGBoost. I tabell 5.3 er evalueringsmålene for alle periodene i *rolling window*. Tabellen inkluderer kun evalueringsmål for testsett, fordi modellen blir trent ved sekvensielle pipelines som gjør resultatene for treningssett mer utfordrende å hente ut.

For RMSE varierer testsettet fra 0.137 og 0.153, og for MSE mellom 0.019 til 0.023. MAE varierer fra 0.099 til 0.111. Disse resultatene er tilsvarende som for XGBoost, men det predikeres noe bedre på siste periode med testsett 2022 som i XGBoost hadde 0.113. R^2 har verdier mellom 0.218 til 0.271 for de ulike periodene. For MAPE varierer testsettene fra 151 til 8745, som også her viser at noen observasjoner har betydelige prediksjonsfeil som øker gjennomsnittet, særlig for testår 2018. MdAPE varierer fra 50 til 58 som betyr at median absolutt prediksjonsfeil er mellom 50 % og 58 %. Når det gjelder toleransegrense, viser PyCaret at mellom 24 % og 28 % av observasjonene har 10 % eller mindre prediksjonsfeil.

Tabell 5.3: Evalueringsmål for alle periodene for prediksjon av ROA i PyCaret

	2014	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**	**2022**
RMSE testsett	0.139	0.137	0.137	0.137	0.137	0.148	0.148	0.148	0.153
MSE testsett	0.019	0.019	0.019	0.019	0.019	0.022	0.022	0.022	0.023
MAE testsett	0.102	0.099	0.101	0.099	0.099	0.107	0.107	0.107	0.111
R^2 testsett	0.264	0.221	0.245	0.255	0.271	0.227	0.236	0.218	0.243
MAPE testsett	169	721	421	290	8745	183	151	190	409
MdAPE testsett	51	50	53	56	57	55	55	57	58
10 % toleranse testsett	0.24	0.25	0.25	0.28	0.26	0.26	0.24	0.25	0.27



Figur 5.18: Aggregert SHAP beeswarmplott for PyCaret prediksjon av ROA for alle periodene i *rolling window*

I figur 5.18 presenteres SHAP beeswarmplott for PyCaret, aggregert for alle periodene i *rolling window*. Tolkningen av dette plottet er tilsvarende som for beeswarmplottet til XGBoost, og har flere likheter. I denne delen er derfor fokuset på det som er ulikt. De seks viktigste variablene er identiske med XGBoost. Produktivitet, lønnskostnad i % av driftsinntekt og kapitalens omløpshastighet har en tilnærmet positiv lineær sammenheng med ROA, illustrert ved gradvise fargeendringer. En ulikhet er at for både lønnskostnad i % av driftsinntekt og kapitalens omløpshastighet, er det noen røde prikker til venstre i plottet. I XGBoost er lave verdier kun assosiert med at modellen vil predikere lavere ROA året etter, men i PyCaret er det en mer kompleks sammenheng.

For rentedekningsgrad er det flere merkbare røde prikker sentrert mot midten, noe til høyre i plottet, enn det er i XGBoost. Det kan bety at høye verdier for rentedekningsgrad indikerer en noe positiv retning for ROA. Det er likevel mest blått, som viser at det er en overvekt av lave verdier for rentedekningsgrad i datasettet.

Arbeidskapital og markedsandel har også blå prikker til høyre og viser at lave verdier er assosiert med prediksjon av høyere ROA. Det er også både røde og blå prikker til venstre for disse variablene. En differanse er at markedsandel er bredere i dette plottet, som kan tyde på at for enkelte observasjoner har markedsandel større påvirkning på ROA i PyCaret enn i XGBoost.

I beeswarmplottet til PyCaret, vist ved figur 5.18, er ln(omsetning) den syvende viktigste variabelen for å predikere kontinuerlig ROA. Modellene til XGBoost og PyCaret er uenig i rekkefølgen for de viktigste variablene fra og med den syvende viktigste variabelen. For ln(omsetning) viser plottet at det i hovedsak er blå verdier til høyre som indikerer at lave verdier er assosiert med høyere verdier av ROA. På venstresiden er det imidlertid både røde og blå prikker, som indikerer en mer kompleks sammenheng mellom ln(omsetning) og ROA. Variabelen er relativt bred med SHAP-verdier fra omtrent -0.12 til 0.1, i motsetning til omtrent -0.04 og 0.03 i XGBoost-plottet. Dette kan indikere at ln(omsetning) har større påvirkning på predikert ROA for enkelte observasjoner. Tettheten av observasjoner er sentrert noe til høyre for nullpunktet på x-aksen, og disse observasjonene er hovedsakelig lilla. Det indikerer at de fleste observasjonene har gjennomsnittlig verdi av ln(omsetning), og at dette påvirker ROA i noe positiv retning.

Videre har egenkapitalandel, likviditetsgrad 1, likviditetsgrad 2 og ln(årsverk) røde verdier rundt midten, noe til venstre for null i SHAP-verdi på x-aksen. De blå verdiene er derimot spredt både til høyre og venstre i plottet, som kan indikere at lave verdier kan påvirke ROA både i positiv og negativ retning, men at høye verdier indikerer moderat til noe negativ retning.

Finansieringsgrad 1 blir valgt som den syvende viktigste variabelen av XGBoost, men av PyCaret er det den tiende viktigste variabelen. Majoriteten av observasjonene er blå, men de røde prikkene befinner seg både til høyre og venstre i beeswarmplottet. Det viser en kompleks sammenheng mellom ROA og finansieringsgrad 1. Tettheten er konsentrert rundt null i SHAP-verdi, som indikerer at for de fleste observasjonene, har variabelen svært lite påvirkning på ROA.

Når det gjelder gjeldsgrad, befinner de fleste røde prikkene seg til høyre, og de fleste

blå til venstre i plottet. For SHAP-verdi mellom -0.25 og 0.25 på x-aksen befinner det seg likevel noen prikker av motsatt farge. Det kan indikere at i flere tilfeller så vil høy verdi av gjeldsgrad predikere høyere ROA, og for lav verdi av gjeldsgrad vil det predikere lavere ROA, men at sammenhengen er kompleks. Tettheten befinner seg svakt til høyre i plottet som indikerer at gjeldsgrad påvirker ROA i svak positiv retning for de fleste observasjonene.

PyCaret velger de samme fire fylkene blant de 20 viktigste variablene. Forskjellen er at PyCaret viser både positive og negative effekter av lokalisering i Oslo, men XGBoost indikerer i større grad positiv ROA. Når det gjelder makrovariablene, valgte XGBoost kun drivstoffpris blant de 20 viktigste variablene. PyCaret velger derimot også gjennomsnittlig årlig styringsrente. Variablene er rangert som henholdsvis nr. 19 og 20 i viktighet. Bredden er kort som indikerer at de påvirker ROA i liten grad.

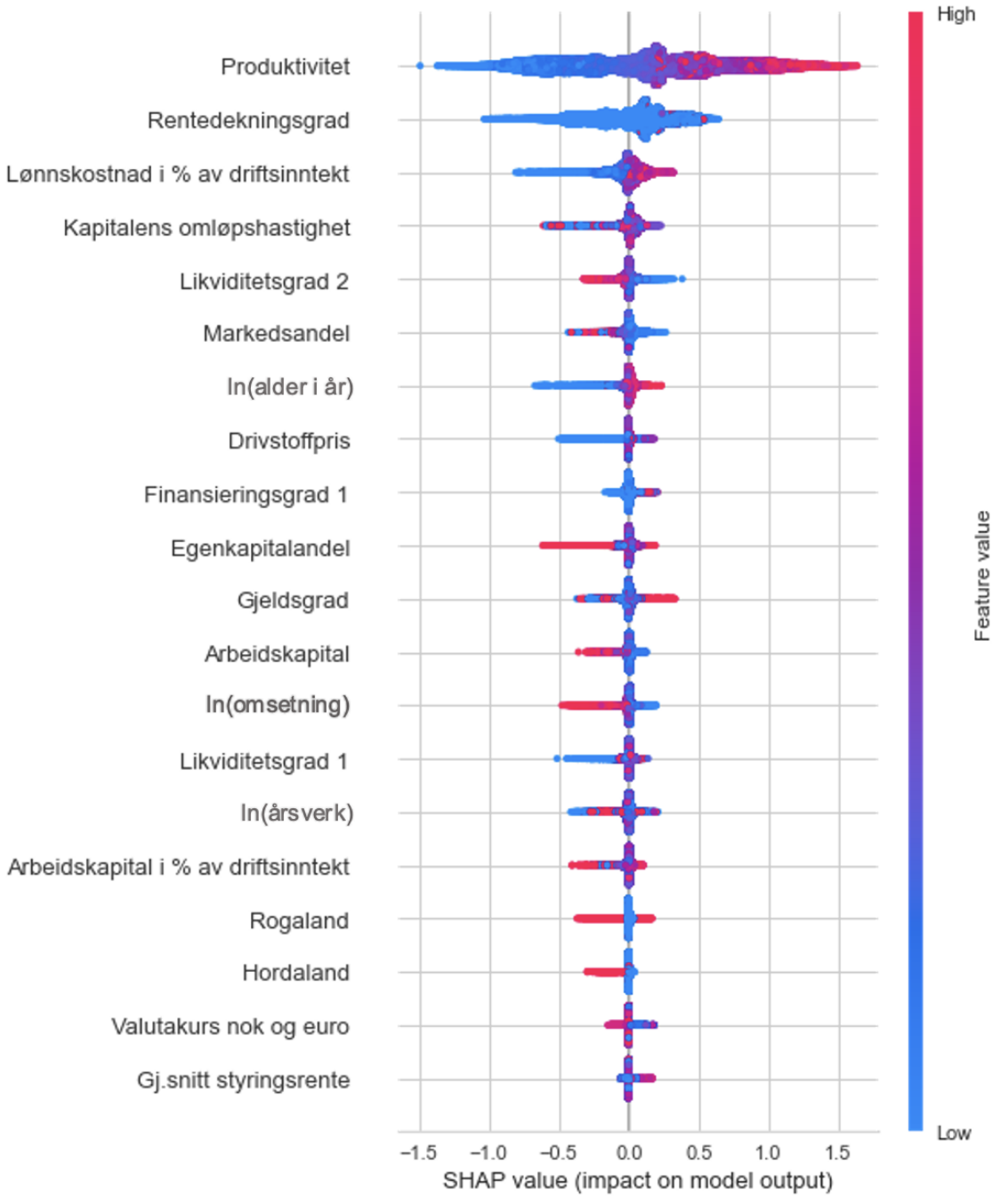
5.3 Klassifisering av ROA

I dette underkapittelet presenteres resultatene til XGBoost for klassifisering av ROA ett år frem i tid. Også her presenteres først evalueringsmålene, og videre presenteres de viktigste variablene ved bruk av SHAP, PDP og ICE-kurver

Tabell 5.4 viser en oversikt over klassifiseringsmodellens ytelse for alle periodene, både for trenings- og testsett. For treningssettet i perioden ligger AUC-verdiene mellom 0.771 og 0.794, med relativt små variasjoner i periodene. For testsettet varierer AUC-verdiene fra 0.729 til 0.759, som er litt lavere enn verdiene for treningssettet. Brier score for treningssettene ligger mellom 0.126 og 0.141, og for testsettene mellom 0.129 og 0.16. Testsettene har noe høyere verdier enn treningssettet.

Tabell 5.4: AUC og Brier score for XGBoost klassifisering av ROA

	2014	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**	**2022**
AUC treningssett	0.788	0.779	0.790	0.794	0.781	0.794	0.779	0.789	0.771
AUC testsett	0.747	0.739	0.737	0.738	0.759	0.735	0.748	0.729	0.737
Brier score treningssett	0.134	0.138	0.126	0.139	0.131	0.139	0.139	0.141	0.137
Brier score testsett	0.129	0.139	0.140	0.160	0.144	0.147	0.138	0.146	0.151



Figur 5.19: Aggregert SHAP beeswarmplott for klassifisering av ROA for alle periodene i *rolling window*

Tilsvarende som tidligere i oppgaven, presenterer figur 5.19 et SHAP beeswarmplott for klassifisering av bedrifters lønnsomhet basert på ROA i to klasser, «lønnsom» og «ikke lønnsom». Plottet viser at produktivitet er den viktigste variabelen, med SHAP-verdier som strekker seg fra -1.5 til 1.7. Beeswarmplottet har fargeendringer fra blå, til lilla, til rød, hvor tettheten av observasjoner ligger til høyre for nullpunktet med i hovedsak lilla og røde prikker. Det viser at moderat til høye verdier av produktivitet øker sannsynligheten for at bedriften klassifiseres som «lønnsom», og at lave produktivitetsverdier assosieres med klassen «ikke lønnsom».

For rentedekningsgraden er tettheten noe til høyre for nullpunktet av SHAP-verdier, noe som tyder på at variabelen generelt er assosiert med høyere sannsynlighet for kategori «lønnsom». Dette er uavhengig av verdien til rentedekningsgrad, og ettersom nesten hele plottet er blått, antyder det at lave verdier kan påvirke klassifiseringen mot både klassen «lønnsom» og «ikke lønnsom». Det er imidlertid noen røde prikker på høyresiden, som kan indikere at høye verdier av rentedekningsgrad kan bidra til høyere sannsynlighet for kategori «lønnsom».

Lønnskostnad i % av driftsinntekt har i hovedsak blå farge på venstresiden av nullpunktet, som øker sannsynligheten for kategori «ikke lønnsom» ved lave verdier. Variabelen er blå frem til nullpunktet, og deretter blir den lilla og rød på høyresiden, som kan bety at middels til høye verdier øker sannsynligheten for at en bedrift blir klassifisert som «lønnsom». Når det gjelder kapitalens omløpshastighet er det en blanding av alle fargene i hele omfanget, noe som gjør det krevende å trekke konklusjoner om hvordan ulike verdier av variabelen påvirker klassifiseringen. Tettheten av observasjoner befinner seg svakt til høyre i plottet, som indikerer at for de fleste observasjonene vil kapitalens omløpshastighet bidra i retning «lønnsom». Bredden til plottet strekker seg derimot lenger mot venstre som antyder at i enkelte tilfeller kan variabelen i større grad bidra til kategori «ikke lønnsom».

De fire viktigste variablene for klassifiseringen har tettheten av observasjoner utenfor nullpunktet på x-aksen. Det tydeliggjør at disse variablene påvirker lønnsomhetsklassifiseringen for majoriteten av observasjoner i utvalget. De resterende variablene har tettheten av observasjoner hovedsakelig plassert rundt nullpunktet på x-aksen, som derimot indikerer at disse variablene generelt har liten innflytelse på lønnsomhetsklassifiseringen for de fleste observasjonene. Det er likevel interessant i det følgende å vite noe om variablenes omfang og hvordan ulike verdier potensielt påvirker klassifiseringen.

Likviditetsgrad 2 har flest røde verdier på venstresiden av plottet, som indikerer at høye verdier assosieres med større sannsynlighet for klassifisering som «ikke lønnsom». Det er en gradvis overgang i fargespekter med lilla i midten og deretter blå til høyre i plottet, som kan tyde på at lavere verdier gir større sannsynlighet for klasse «lønnsom». Når det gjelder markedsandel, er det røde prikker til venstre i plottet, som kan indikere at høy markedsandel er assosiert med økt sannsynlighet for klassifisering som «lønnsom». Det er derimot en blanding av røde og blå punkter på venstresiden, som viser at det er en mer kompleks sammenheng mellom lav markedsandel og lønnsomhetsklassifisering.

Ln(alder i år), drivstoffpris og likviditetsgrad 1 har til felles at de nesten kun

har blå farge i breddene og rødt i midten. Dette indikerer at lave verdier for disse variablene påvirker klassifiseringen mer enn det høye verdier gjør. For ln(alder i år) er bredden størst til venstre, som kan tyde på at dersom bedriften er ung kan det i enkelte tilfeller øke sannsynligheten for at bedriften blir klassifisert som «ikke lønnsom». En eldre bedrift vil derimot ha mindre påvirkning på klassifiseringen, men påvirker noe mot «lønnsom». Drivstoffpris og likviditetsgrad 1 har også størst bredde til venstre, og kan tolkes på samme måte som ln(alder i år), der lave verdier kan øke sannsynligheten for klasse «ikke lønnsom». For disse to variablene har høye verdier liten påvirkning.

Finansieringsgrad 1 har lave verdier både til venstre og rundt midten, men har også rødt på høyresiden. Lave verdier påvirker i hovedsak lite, men kan gå mot begge klassene. Høye verdier assosieres derimot med høyere sannsynlighet for klasse «lønnsom». Når det gjelder egenkapitalandel, har plottet rødt både til venstre og høyre, med noen blå observasjoner i midtpartiet. Høy egenkapitalandel vil derfor i enkelte tilfeller påvirke klassifiseringen i større grad, og som regel i retning «ikke lønnsom». Lav egenkapitalandel har mindre effekt på klassifiseringen.

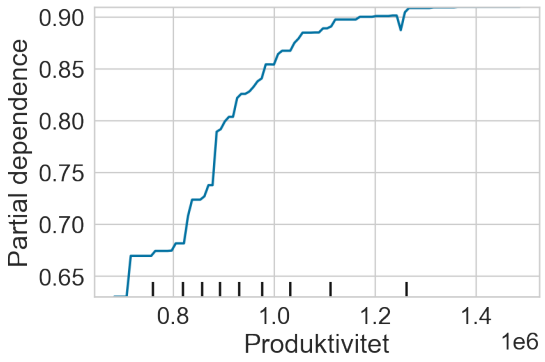
Gjeldsgrad, ln(årsverk) og arbeidskapital i % av driftsinntekt har likheter med kapitalens omløpshastighet ved at plottene har blanding av alle fargene i hele omfanget. For disse variablene er det krevende å assosiere ulike verdier med lønnsomhetsklassifiseringen. Når det gjelder arbeidskapital og ln(omsetning), har begge størst bredde til venstre, og denne siden er rød. Det indikerer at høye verdier av disse variablene er assosiert med høyere sannsynlighet for klasse «ikke lønnsom». Høyresidene er kortere og blå som indikerer at lave verdier påvirker klassifiseringen mot klassen «lønnsom», men i mindre grad.

For lokaliseringsvariablene, velger XGBoost klassifiseringsmodellen kun ut Rogaland og Hordaland. Beliggenhet i Hordaland er assosiert med økt sannsynlighet for klasse «ikke lønnsom», mens det for Rogaland kan øke sannsynligheten for begge klassene i ulike tilfeller. Til slutt er også makrovariablene valutakurs nok og euro og gjennomsnittlig styringsrente valgt ut blant de 20 viktigste variablene for klassifisering. Det er lite omfang av SHAP-verdier for disse makrovariablene, og de har derfor lite innflytelse på klassifiseringen. Det kan likevel tyde på at høy valutakurs gir noe økt sannsynlighet for klasse «ikke lønnsom», og motsatt effekt for lav valutakurs. For høy gjennomsnittlig styringsrente er det noe høyere sannsynlighet for klasse «lønnsom», og motsatt effekt for lav styringsrente.

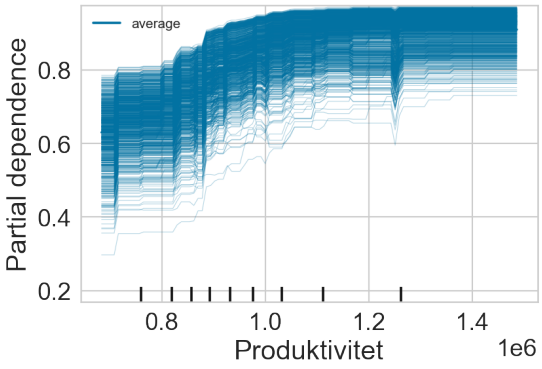
PDP og ICE

I denne delen skal vi presentere PDP og ICE-kurver for de fire viktigste variablene i beeswarmplottet til klassifiseringsmodellen med XGBoost. Variablene det gjelder er produktivitet, rentedekningsgrad, lønnskostnad i % av driftsinntekt og kapitalens omløpshastighet, ettersom disse har tettheten av observasjoner utenfor midtpunktet i 5.19.

For produktivitet, figur 5.20, er det en tilnærmet positiv lineær sammenheng i det delvise avhengighetsplottet. Høyere verdier for produktivitet øker derfor sannsynligheten for at modellen klassifiserer bedriften som «lønnsom». PDP viser likevel ikke en perfekt lineær sammenheng. Kurven har brattere helning på omtrent 900 000 i produktivitet, og etter omtrent 1 200 000 i produktivitet flater kurven ut, som antyder et metningspunkt. I figur 5.21 ser man at ICE-kurvene for produktivitet viser den samme tilnærmede positive lineariteten som PDP. Det er lite variasjon i kurvene, som viser at det er en robusthet i at høyere produktivitet har sammenheng med kategori «lønnsom». De fleste bedriftene har produktivitet mellom 800 000 - 1 100 000.

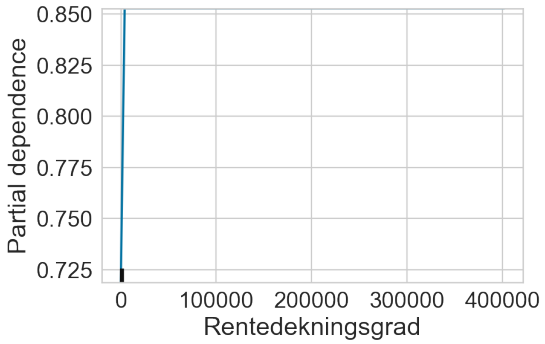


Figur 5.20: PDP for produktivitet klassifisering av ROA

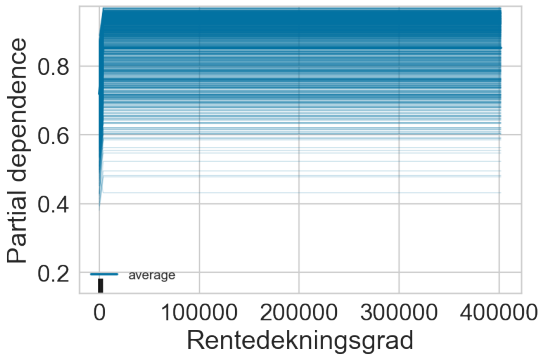


Figur 5.21: ICE-kurver for produktivitet klassifisering av ROA

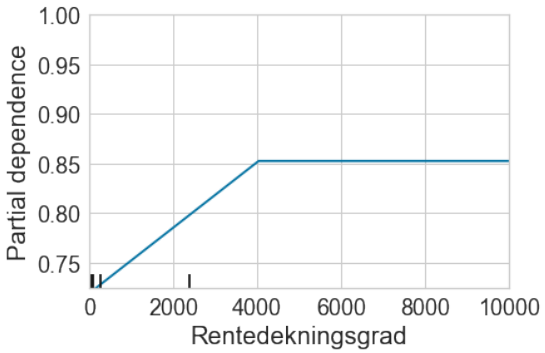
I figur 5.22 observeres en bratt kurve med positivt stigningstall, som indikerer at høyere verdier for rentedekningsgrad assosieres med økt sannsynlighet for lønnsom klassifisering. Denne effekten er sammenfallende med PDP for prediksjon av kontinuerlig ROA i figur 5.6. ICE-kurvene i figur 5.23 viser også at det er en terskelverdi hvor effekten rentedekningsgrad har på klassifiseringen forsvinner. Denne terskelverdien er ytterligere undersøkt i de forstørrede visningene, av henholdsvis PDP og ICE for rentedekningsgrad, i figur 5.24 og 5.25. Disse figurene viser at terskelverdien ligger på rentedekningsgrad lik 4000, og at høyere verdier enn dette ikke vil påvirke klassifiseringen. Dette er likevel en svært høy verdi for rentedekningsgrad og det vil være svært få tilfeller som opplever ingen effekt av rentedekningsgrad. ICE-kurvene i figur 5.25 viser at denne terskelverdien på 4000 er konsekvent for de individuelle observasjonene, men at det derimot er ulik helningsgrad på kurvene før denne terskelverdien.



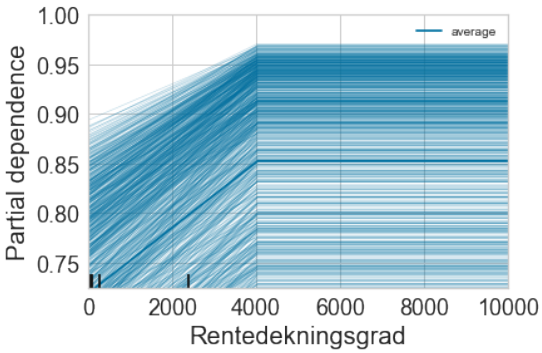
Figur 5.22: PDP for rentedekningsgrad klassifisering av ROA



Figur 5.23: ICE-kurver for rentedekningsgrad klassifisering av ROA



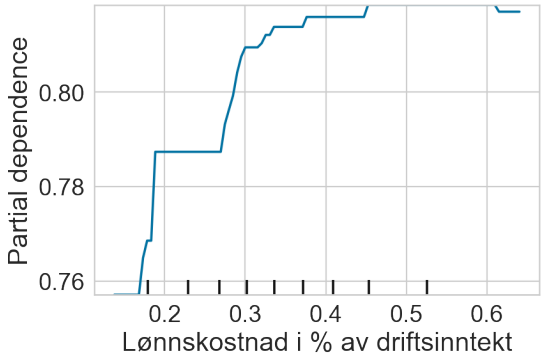
Figur 5.24: Nærmere undersøkelse av terskelverdi for PDP rentedekningsgrad klassifisering av ROA



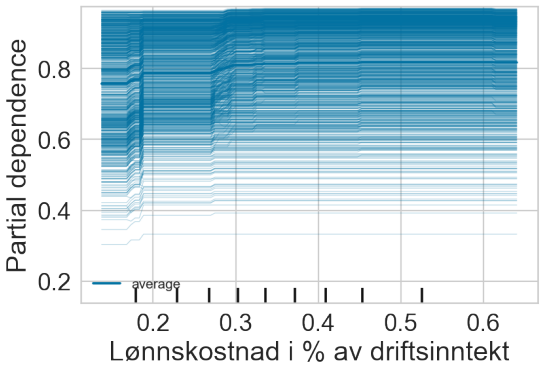
Figur 5.25: Nærmere undersøkelse av terskelverdi for ICE-kurver rentedekningsgrad klassifisering av ROA

I figur 5.26 illustreres PDP for lønnskostnad i % av driftsinntekt. Grafen har positiv helning, som indikerer at det er en sammenheng mellom høyere verdier av lønnskostnad i % av driftsinntekt og «lønnsom» klassifisering. Økninger på rundt 20 % gir økt sannsynlighet for å bli klassifisert som «lønnsom», før den flater ut og deretter øker igjen for verdier mellom omtrent 27 %- 30 %. Fra 30 % og utover på x-aksen påvirker lønnskostnad i % av driftsinntekt klassifiseringen i moderat grad. ICE-kurvene i figur 5.27 viser også at det er en positiv helning i kurvene, men at disse flater ut på rundt 30 %. Det er likevel variasjon i hvordan lønnskostnad i % av driftsinntekt påvirker klassifiseringer for individuelle observasjoner, som kan ses ved spredte og noe ulike helninger på ICE-kurvene.

PDP for kapitalens omløpshastighet finnes i figur 5.28. Grafen viser en bratt helning for økninger rett over 1 i kapitalens omløpshastighet, hvor sannsynligheten for lønnsom klassifisering øker med omtrent 0.04. Deretter har kapitalens omløpshastighet marginal effekt på klassifiseringen opp til en verdi på rett over 2, hvor sannsynligheten øker med omtrent 0.01. Effekten på klassifiseringen flater ut frem til omtrent

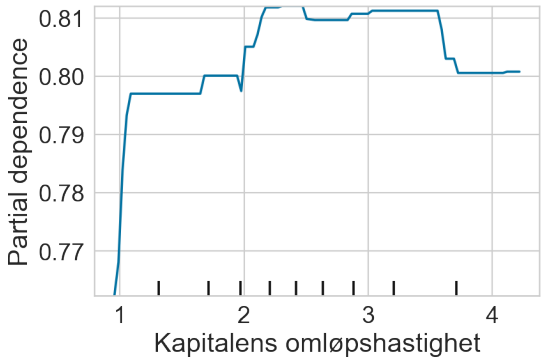


Figur 5.26: PDP for lønnskostnad i % av driftsinntekt klassifisering av ROA

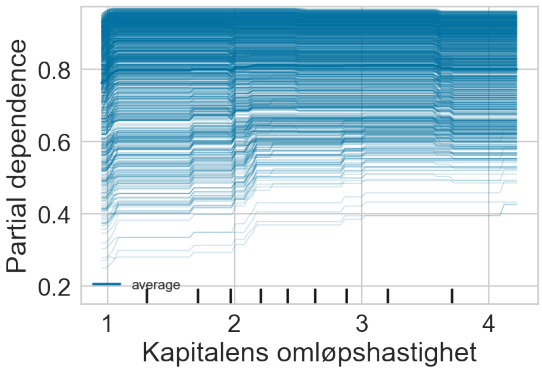


Figur 5.27: ICE-kurver for lønnskostnad i % av driftsinntekt klassifisering av ROA

3.6, hvor sannsynligheten reduseres med omtrent 0.01, tilbake til samme nivå som intervallet rett over 1 til 2. I figur 5.29 viser ICE-kurvene at det er variasjon i hvordan og hvorvidt kapitalens omløpshastighet påvirker klassifiseringen. Det er ICE-kurver, som i likhet med PDP, har positiv helning for økning av lave verdier av kapitalens omløpshastighet, samtidig som det er individuelle observasjoner som har positiv helning også for høyere verdier, i tillegg til observasjoner uten helninger.



Figur 5.28: PDP for kapitalens omløpshastighet klassifisering av ROA



Figur 5.29: ICE-kurver for kapitalens omløpshastighet klassifisering av ROA

6 Diskusjon

For å danne et helhetlig bilde av hvilke faktorer som påvirker lønnsomheten til små- og mellomstore bedrifter innen bygg- og anleggsbransjen, har vi sett på variabler på tre forskjellige nivåer; bedriftsspesifikt, bransjespesifikt og makroøkonomisk nivå. Resultatene som ble presentert i kapittel 5 tyder på at det er en rekke uavhengige variabler som bidrar til å forklare og predikere lønnsomhet blant bedriftene i utvalget. I dette kapittelet skal vi først diskutere resultatene fra modellene, og deretter de viktigste uavhengige variablene.

6.1 Modellene

OLS og fixed effects

OLS har en forklaringsgrad på 24.1 %, som er andelen av variansen i ROA som blir forklart av de uavhengige variablene i modellen. For fixed effects er forklaringsgraden noe lavere (20.5 %). Dette indikerer at evnen til å forklare variansen i ROA blir redusert når man kontrollerer for individuelle bedriftseffekter. Selv om modellene fanger opp en betydelig del av variansen i ROA, er det fremdeles en stor del som ikke forklares av de uavhengige variablene i modellene. Dette kan blant annet skyldes at det er andre viktige forklaringsvariabler som ikke er inkludert i modellene, eller at det er variabler med ikke-lineære effekter som er krevende å fange opp i lineære modeller. Lorentzen og Bergander (2022), som også måler lønnsomhet gjennom ROA, har relativt like resultater med en forklaringsgrad for OLS på 20.8 %, og 25.8 % for fixed effects. Resultatene indikerer at også de har hatt utfordringer med å forklare store deler av variansen. Dette understreker kompleksiteten i å fange opp alle relevante variabler som kan påvirke bedriftenes lønnsomhet. Det kan være andre variabler som er krevende å tallfeste, som for eksempel bedriftskultur eller de ansattes engasjement, som kan ha vesentlig innvirkning på bedrifters lønnsomhet.

Regnskapsårene viser årlige effekter på ROA sammenlignet med referanseåret, som er 2010. I OLS og fixed effects viser de senere regnskapsårene en generelt positiv effekt på ROA. Dette indikerer at ROA, kontrollert for andre faktorer, generelt sett er høyere for årene etter 2010. Koeffisientene er noe høyere for OLS enn for fixed effects for de samme regnskapsårene. Dette viser at effekten til regnskapsårene blir redusert når det kontrolleres for enhetsspesifikke egenskaper. Det er ikke uventet at regnskapsårene har positive koeffisienter ettersom 2010 har lavere gjennomsnittlig ROA enn de senere regnskapsårene i utvalget, som vist i figur 3.3.

XGBoost og PyCaret

For prediksjonsmodellen med XGBoost, viser evalueringsmålene at det er relativt like resultater mellom trenings- og testsettene, og treningssettene gjør det noe bedre enn testsettene. Det er likevel et unntak for 10 % toleransegrense, hvor to perioder i XGBoost gjør det noe bedre i testsettet enn for treningssettet. Det er noe uvanlig, men kan muligens skyldes tilfeldigheter. Generelt sett tyder resultatene likevel på at modellen verken er overtilpasset eller undertilpasset, ref. *The Bias-Variance Trade-Off*. MAE for prediksjonsmodellen med XGBoost er for treningssettet mellom 0.092 og 0.099, og for testsettene mellom 0.098 og 0.113. Det vil si at modellen er noe uforutsigbar i prediksjonen av ROA, fordi en feilmargin på mellom 0.098 og 0.113 er en relativt stor endring når det kommer til lønnsomhet. Særlig når medianen er 0.09, så kan denne gjennomsnittlige feilmarginen for eksempel predikere at en relativt lønnsom bedrift, ikke er lønnsom.

For prediksjon av kontinuerlig ROA med både XGBoost og PyCaret, er det som vi så i tabell 5.2 og 5.3, store forskjeller i prediksjonsfeil mellom MAPE og MdAPE, som kan tyde på at enkelte observasjoner har veldig store absolutte prediksjonsfeil som drar opp gjennomsnittet. For testperiode 2018 er de absolutte prediksjonsfeilene 7180 % i XGBoost og 8745 % for PyCaret, som er en ekstremt stor feil og indikerer noe galt i dataen for dette teståret. Dette kan muligens skyldes at det er noen større uteliggere i dette året, ettersom vi har sett noen indikasjoner på større uteliggere for eksempel i figur 5.7 for rentedekningsgrad. Dette tyder derfor på at MdAPE er et bedre evalueringsmål for denne modellen, og XGBoost og PyCaret for kontinuerlig ROA viser at median absoluttprediksjonsfeil er mellom 50 % og 58 %. Det antyder at minst halvparten av prediksjonene avviker fra de faktiske verdiene med minst 50 %. Det viser at man må ha et noe kritisk blikk på prediksjonen ettersom det kan være betydelig feil når det gjelder lønnsomhetsnivå. For en toleranse på 10 % prediksjonsfeil viser imidlertid testsettene at modellene har god nøyaktighet for mellom 23 % og 27 % av observasjonene.

RMSE, MSE og MAE for både prediksjonsmodellen med XGBoost og PyCaret viser til en generell stabilitet over perioden. Det er imidlertid en økning i feilraten mot slutten av perioden, fra 2020 til 2022. Det kan se ut til at modellene i disse periodene sliter med å tilpasse seg nye endringer i økonomiske forhold eller mønstre som påvirker ROA. Det kan muligens være at covid-19 er en av grunnene til at modellene presterer dårligere de siste årene. For PyCaret har evalueringsmålene RMSE og MSE hatt en merkbar økning fra 2018 til 2019, med videre økninger frem til 2022. Dette tyder på at PyCaret er mindre nøyaktig i de siste årene av perioden. Toleransetestsettene på 10 % viser imidlertid at PyCaret kan komme med konsistente prediksjoner innenfor denne feilmarginen.

Vi har brukt AUC og Brier score for å evaluere klassifiseringsmodellen med XGBoost. AUC-verdiene bør ligge i nærheten av 1, som indikerer en perfekt modell (James mfl., 2023). Resultatene viser at AUC-verdiene for testsettet er noe lavere enn treningssettet, med AUC-verdier for testsettet mellom 0.729 og 0.759, og mellom 0.771 og 0.794 for treningssettet. Dette indikerer at modellen generelt har god pre-

diktiv evne, og gjør det bedre enn tilfeldig tildeling av klasse. Imidlertid er 79.1 % av bedriftene i utvalget definert som «lønnsom». Det viser at modellen også kunne fått god prediktiv evne ved å klassifisere alle bedriftene som «lønnsom». Modellen predikerer likevel også en del av observasjonene som «ikke lønnsom», forvirringsmatrise ligger i vedlegg F, og det er muligens mer interessant å kunne predikere hvilke bedrifter som er i fare for å være «ikke lønnsom» ett år frem i tid. For Brier score skal verdiene helst ligge nærme 0, som indikerer en nøyaktig modell med lav feilrate og god prediktiv evne (Dash, 2020). Brier scoren for testsettet ligger mellom 0.129 og 0.160, og treningssettet mellom 0.126 og 0.141. Brier scoren var særlig høy for perioden med testsett i 2017, som hadde en Brier score på 0.160. Det kan tyde på at det var mer krevende å predikere riktig klasse i denne perioden.

Det er ikke overraskende at klassifiseringsmodellen viser til bedre evalueringsmål enn prediksjonsmodellene ettersom det er mer krevende å predikere akkurat lønnsomhetsnivå enn om bedriften vil være «lønnsom» eller «ikke lønnsom». Dette gjelder særlig ettersom dataen, selv etter winsorizing, viser tegn til at det eksisterer ekstremverdier. Dette vil i større grad slå ut som feil for modellene som predikerer kontinuerlig ROA. Det er heller ikke overraskende at vi ikke oppnår særlig høy forklaringsgrad i prediksjonene. Det var som forventet krevende å predikere lønnsomhet kun basert på årsregnskap og statistiske variabler uten å gå mer i dybden på bedriftene, som for eksempel ved å analysere bedriftenes strategi.

6.2 Lønnsomhetsdriverne

Vi vil først diskutere de bedriftsspesifikke forholdene, hvor vi begynner med å se på variabler knyttet til Porters kostnadsdrivere, etterfulgt av en gjennomgang av likviditet og soliditet. Videre blir bransjespesifikke og makroøkonomiske variabler diskutert.

6.2.1 Bedriftsspesifikke variabler

Selskapsstørrelse

Variablene ln(årsverk), ln(omsetning) og markedsandel har blitt brukt for å måle hvorvidt selskapsstørrelse påvirker lønnsomheten til små- og mellomstore bedrifter i bygg- og anleggsbransjen.

LASSO trakk frem ln(årsverk) som en viktig forklaringsvariabel, med positiv koeffisient. Dette tyder på at en økning i årsverk kan føre til økt lønnsomhet, målt ved ROA. Resultatene fra prediksjonsmodellen med XGBoost og PyCaret, viser imidlertid at ln(årsverk) har liten til ingen påvirkning på lønnsomheten for de fleste observasjonene, og at det er en kompleks sammenheng mellom variabelen og lønnsomhet. Likevel viser resultatene til noen mønstre. Resultatene fra prediksjonsmodellen med XGBoost kan indikere at flere årsverk indikerer lavere lønnsomhet for noen bedrifter, og færre årsverk kan indikere en økning i ROA. Det er i tråd med funnene til Føyen og Danielsen (2020), som observerte at en økning i antall årsverk hadde negativ ef-

fekt på lønnsomheten. En mulig forklaring på den negative sammenhengen mellom antall årsverk og lønnsomhet, kan være at bedrifter med flere ansatte enn nødvendig, ikke vil oppnå tilstrekkelige produktivitetsgevinster til å dekke økningen i kostnader. I tillegg peker Porter (1985) på at større bedrifter kan oppleve mindre motiverte ansatte. Dette kan muligens også gjelde de mellomstore bedriftene i utvalget. Resultatene fra PyCaret gir indikasjoner på at effekten til ln(årsverk) kan variere ved at både høye og lave verdier kan ha både negativ og positiv påvirkning på ROA. For klassifiseringsmodellen er det vanskelig å si noe om hvordan variabelen påvirker sannsynligheten for å predikere en av klassene.

Ln(omsetning) blir ikke trukket frem som en viktig forklaringsvariabel av LASSO, men blir rangert som nummer åtte og syv i viktighet av prediksjonsmodellen med XGBoost og PyCaret. For klassifiseringsmodellen blir variabelen ansett som mindre viktig, og ln(omsetning) er rangert som nummer 13 i viktighet. Beeswarmplottene for prediksjonsmodellen med XGBoost viser at variabelen påvirker ROA i liten grad, men at det er tendenser til at høye verdier, som er omsetning på over kr 50 000 000, kan ha negative effekter på ROA, mens lave verdier, omsetning på mindre enn kr 50 000 000, kan føre til høyere ROA for noen bedrifter. For klassifiseringsmodellen indikerer resultatene at lave verdier øker sannsynligheten for at modellen predikerer bedriften som «lønnsom», mens høye verdier øker sannsynligheten for å bli klassifisert som «ikke lønnsom».

Funnene indikerer at høyere omsetning ikke nødvendigvis fører til økt lønnsomhet, og antyder at det ikke er omsetningsvekst alene som driver lønnsomheten. Dette kan indikere at vekst i omsetning uten tilsvarende kostnadskontroll eller effektivitetsforbedringer, kan redusere lønnsomheten. Dette står i sammenheng med rapporten til Dalsegg og Lidsheim (2023) som beskriver at i perioden fra 2021 til 2022 var det en omsetningsvekst fra 8.9 % til 13.3 % i 2022, men at driftsmarginen i samme periode sank med 0.6 prosentpoeng. De begrunnet omsetningsveksten med høy inflasjon og prisvekst, og mener at tallene for endringer i omsetning og driftsmargin viser at bedriftene ikke har lykkes i å ta ut økninger i kostnadene gjennom salgsprisene. Funnene i denne oppgaven kan derfor muligens tyde på at dette kan være tilfelle, også for flere perioder med omsetningsøkning. Funnene står derimot i kontrast til Akintoye og Skitmore (1991), som observerte at selskapsstørrelse, målt gjennom omsetning, var signifikant og positivt korrelert med lønnsomhet for bedrifter innen bygg og anlegg. En potensiell forklaring på forskjellen i resultat kan forklares ved at deres forskning utelukkende fokuserte på store aktører, mens vi har sett på små- og mellomstore bedrifter. Beeswarmplottet til PyCaret viser derimot en mer kompleks sammenheng mellom ln(omsetning) og ROA.

Markedsandel blir også trukket frem av LASSO til å ha negativ påvirkning på lønnsomheten, noe som kan indikere at en økning i markedsandel ikke vil føre til økt lønnsomhet. Dette kan skyldes, som diskutert av Dalsegg og Lidsheim (2023), at en økning i markedsandel kan føre til priskonkurranse og høyere kostnader som vil undergrave fortjenesten. En annen mulig forklaring på at økning i markedsandeler kan føre til redusert lønnsomhet, er at det kan være dyrt å kapre markedsandeler, med økte kostnader knyttet til eksempelvis markedsføringskampanjer og utvidelser

av kapasiteten. Funnene står i kontrast til Capon mfl. (1990) sine funn, som kom frem til at vekst i markedsandel vil føre til høyere lønnsomhet. Beeswarmplottene til prediksjonsmodellen med XGBoost og PyCaret viser antydning til at lave verdier av markedsandel kan indikere høyere lønnsomhet, og høye verdier for markedsandel kan knyttes til redusert lønnsomhet. Det kan tyde på at bedrifter ikke nødvendigvis vil dra nytte av å øke markedsandelen. Beeswarmplottet for klassifiseringsmodellen kan indikere at lave verdier for markedsandel øker sannsynligheten for at modellen predikerer bedriften som «lønnsom», men at det for høye verdier er en noe mer kompleks sammenheng.

Samlet sett antyder en økning av selskapsstørrelse, målt gjennom disse tre variablene, en kompleks effekt på lønnsomheten, med både positive og negative effekter. Økt selskapsstørrelse forbindes ofte med stordriftsfordeler ved at bedrifter kan fordele kostnader over høyere omsetning eller ved å drive mer rasjonelt ved høyere volum, og funnene samsvarer med antakelsen til Porter (1985), om at stordriftsfordeler kan gi både negative og positive effekter på lønnsomheten.

Kapasitetsutnyttelse

Variablene produktivitet og lønnskostnad i % av driftsinntekt måler om kapasitetsutnyttelse påvirker lønnsomhet. Produktivitet, målt som verdiskaping per årsverk, blir av LASSO valgt som den viktigste forklaringsvariabelen med positiv koeffisient. Variabelen har positive koeffisienter i både OLS og fixed effects, som antyder at en økning i produktivitet vil predikere høyere lønnsomhet. Produktivitet blir også trukket frem som den viktigste forklaringsvariabelen av begge XGBoost-modellene, og av PyCaret. Beeswarmplottene til prediksjonsmodellen med XGBoost og PyCaret, viser en tendens til at produktivitet har en lineær sammenheng med ROA, ved at lave verdier for produktivitet indikerer lavere ROA, mens høyere verdier indikerer høyere ROA. Beeswarmplottet til klassifiseringsmodellen indikerer at høyere verdier av produktivitet øker sannsynligheten for at modellen klassifiserer en bedrift som «lønnsom», og at lave verdier øker sannsynligheten for at en bedrift klassifiseres som «ikke lønnsom». PDP til prediksjonsmodellen bekrefter den lineære sammenhengen mellom produktivitet og ROA. I tillegg viser det at verdier for produktivitet fra 800 000 til 1 100 000, har bratt positiv lineær sammenheng, og at det deretter er en svakere positiv økning i ROA for verdier over 1 100 000. Funnene støtter antakelsen om at produktivitet blant de ansatte bidrar til å skape konkurransefortrinn. Dette er i tråd med Porter (1985), som hevder at bedrifter kan oppnå konkurransefortrinn dersom de er flinke til å utnytte kapasiteten, og funnene til Føyen og Danielsen (2020) som fant at kapasitetsutnyttelse, målt gjennom produktivitet, har positiv sammenheng med lønnsomhet for bedrifter innen bygg og anlegg.

Lønnskostnad i % av driftsinntekt blir også trukket frem av LASSO som en viktig forklaringsvariabel, med negativ koeffisient. Koeffisientene er også negative i både OLS og fixed effects, som vil si at økning av denne variabelen generelt tenderer mot å redusere ROA. Imidlertid viser beeswarmplottene til prediksjonsmodellen med XGBoost og PyCaret en motstridende tendens, ved at lave verdier indikerer lavere ROA,

mens høye verdier indikerer høyere ROA. Klassifiseringsmodellen støtter dette, og viser at høye verdier er assosiert med økt sannsynlighet for at en bedrift klassifiseres som «lønnsom». Dette gjelder særlig for økninger av lønnskostnad i % av driftsinntekt opp til 30 %, som sett i figur 5.26. Dette kan gi mening ettersom bygg og anlegg er en arbeidsintensiv bransje, og lave verdier for denne variabelen kan tyde på at bedrifter har for lav investering i menneskelige ressurser, som kan føre til redusert lønnsomhet. Fra PDP for prediksjonsmodellen med XGBoost, ser vi at det er en tilnærmet positiv lineær sammenheng mellom variabelen og lønnsomhet. Det er en kraftig økning i lønnsomhet for verdier inntil 50 %, og for verdier etter dette vil det ikke være noen ytterligere økning i lønnsomhet. Dette kan tyde på at det er en terskelverdi, hvor lønnsomheten øker frem til lønnskostnadene utgjør 50 % av driftsinntektene, og at en ytterligere økning etter dette ikke vil øke lønnsomheten.

Erfaring

Porter (1985) hevder bedrifter kan redusere kostnadene gjennom erfaring, ved at aktiviteter gjennomføres mer effektivt. For å undersøke om dette er tilfelle, har bedriftens alder, målt gjennom ln(alder i år), blitt brukt som forklaringsvariabel for å måle om erfaring påvirker lønnsomheten. Ln(alder i år) blir trukket frem som en relativt viktig forklaringsvariabel av LASSO med positivt fortegn, selv om koeffisienten er lav. På grunn av problemer med multikollinearitet, ble ikke denne forklaringsvariabelen med i OLS eller fixed effects. Resultatene fra beeswarmplottene til prediksjonsmodellen med XGBoost og PyCaret, indikerer at lang levetid for bedrifter ikke nødvendigvis påvirker lønnsomheten. Imidlertid er resultatene uregelmessig for de yngre bedriftene, og lønnsomheten kan bli påvirket både positivt og negativt dersom bedriften er ung. Unge bedrifter assosieres likevel oftere med lavere ROA, enn høyere. Dette bekreftes også av beeswarmplottet til klassifiseringsmodellen, som indikerer at unge bedrifter kan øke sannsynligheten for å bli klassifisert som «ikke lønnsom», men at eldre bedrifter har liten påvirkning på klassifiseringen, men påvirker noe mot «lønnsom». Det at unge bedrifter oftere assosieres med lavere ROA, kan for eksempel ha sammenheng med at unge bedrifter enda ikke har etablert en lojal kundebase og en kjent merkevare, eller at de har begrensede økonomiske ressurser.

Resultatene er i tråd med Føyen og Danielsen (2020) som kom frem til at langvarig erfaring ikke nødvendigvis ga store utslag på lønnsomheten. Jolly Cyril og Singla (2020) fant heller ingen sammenheng mellom bedrifters alder og lønnsomhet. Det kan dermed se ut til at langvarig erfaring ikke nødvendigvis gir konkurransefortrinn, noe som kan skje dersom det er læringslekkasjer, der bedrifter i samme bransje tar lærdom av hverandre (Porter, 1985). Når det gjelder læringslekkasjer, kan innleie av arbeidskraft spille en rolle i hvordan kunnskap og erfaring spres mellom bedriftene i bransjen. Ved innleie av arbeidskraft kan de midlertidige ansatte bringe videre innsikt og praksiser fra tidligere arbeidsgivere. Dette kan skape en uformell kunnskapsoverføring som kan redusere bedrifters konkurransefortrinn.

Lokalisering

Både Whittington mfl. (2020) og Porter (1985) mener bedrifters lokalisering vil kunne påvirke lønnsomheten, og for å undersøke denne påstanden, har det blitt inkludert dummyvariabler for fylkene i Norge.

Blant alle dummyvariablene for lokalisering, var det kun Oslo og Hordaland som ble pekt ut av LASSO til å være viktige forklaringsvariabler, med positiv koeffisient for Oslo og negativ for Hordaland. Disse funnene blir bekreftet av beeswarmplottene til prediksjonsmodellen med XGBoost og PyCaret, som antyder at bedrifter i Oslo kan oppleve høyere lønnsomhet. Dette gir mening dersom vi ser på figur 3.1 som viser gjennomsnittlig ROA per fylke, og viser at Oslo er ett av fylkene med høyest gjennomsnittlig ROA. En av grunnene til at Oslo opplever høyere lønnsomhet, kan skyldes at de drar nytte av den gode infrastrukturen, og at leverandører og kunder er samlet på ett sted (Dalsegg & Lidsheim, 2023). Det er samtidig også noe overraskende ettersom Oslo er det fylket med flest bedrifter, ref. figur 3.2. Forventningen var derfor at det kunne være hard konkurranse i Oslo, og at dette kunne påvirke lønnsomheten til bedriftene negativt.

Videre indikerer beeswarmplottene for prediksjonsmodellen med XGBoost og PyCaret at lokalisering for bedrifter i Hordaland kan indikere lavere lønnsomhet. Det understøttes av figur 3.1, som viser at Hordaland er ett av fylkene med lavest gjennomsnittlig ROA. Klassifiseringsmodellen bekrefter også disse funnene, og resultatene fra beeswarmplottet indikerer at lokalisering i Hordaland øker sannsynligheten for at en bedrift blir klassifisert som «ikke lønnsom». Ifølge Dalsegg og Lidsheim (2023) kan dette reflektere at selv om sentrale steder har høyere aktivitetsnivå og flere prosjekter på grunn av høyere befolkningstetthet, kan den økte konkurransen i disse områdene likevel slå negativt ut på lønnsomheten. Hordaland er også ett av fylkene med flest bedrifter, som kan tyde på at det er hard konkurranse som kan redusere lønnsomheten.

Prediksjonsmodellen med XGBoost og PyCaret trekker også frem Møre og Romsdal og Rogaland som viktige forklaringsvariabler. Resultatene peker på at bedrifter lokalisert i Møre og Romsdal ofte har lavere ROA. Dette gir også mening når man ser på figur 3.1, som viser at Møre og Romsdal er det fylket med lavest gjennomsnittlig ROA. Det lave antallet bedrifter i fylket indikerer at det nødvendigvis ikke er så hard konkurranse i dette fylket, men den lave lønnsomheten kan potensielt skyldes lav befolkningstetthet og færre prosjekter. En annen mulig forklaring på den lave lønnsomheten i Møre og Romsdal kan være knyttet til demografiske utfordringer, som for eksempel utflytting av yngre arbeidskraft og en aldrende befolkning (Whittington mfl., 2020). For Rogaland er resultatene utydelige og det er vanskelig å fastslå en tydelig effekt på lønnsomheten for denne variabelen alene. Hvis vi ser dette i sammenheng med aktivitetsnivået, påpekte Dalsegg og Lidsheim (2023) at høye energipriser i 2023 medførte fortsatt høyt aktivitetsnivå i de områdene som er knyttet til olje- og gassnæringen. Det kan derfor tenkes at energiprisen i større grad påvirker aktivitetsnivået i Rogaland, og at endringer i energipris kan gjøre lønnsomheten mer volatil.

Likviditet

Variablene arbeidskapital, arbeidskapital i % av driftsinntekter, likviditetsgrad 1, likviditetsgrad 2 og kapitalens omløpshastighet har blitt brukt for å undersøke forholdet mellom likviditet og lønnsomhet for bedriftene i utvalget.

Blant forklaringsvariablene som måler likviditet, blir kapitalens omløpshastighet pekt ut av LASSO til å være den mest signifikante, med positiv koeffisient. Dette er i tråd med funnene til Lorentzen og Bergander (2022) om at kapitalens omløpshastighet har en positiv effekt på lønnsomhet. Som nevnt i kapittel 3, er ikke kapitalens omløpshastighet et direkte mål på likviditet, men den gir innsikt i hvor effektive bedrifter er til å konvertere aktiva til inntekter. Høye verdier av kapitalens omløpshastighet kan bidra til å forbedre likviditeten ved at bedriften raskt omgjør kapital til inntekter, noe som frigjør midler til mer lønnsomme aktiviteter. Funnene viser en positiv sammenheng mellom kapitalens omløpshastighet og lønnsomhet, og det styrker hypotesen om at bedrifter som utnytter kapitalen effektivt, oppnår bedre lønnsomhet. Overraskende nok viser kapitalens omløpshastighet en positiv sammenheng med ROA i OLS, mens den i fixed effects viser en negativ sammenheng. Forskjellen i resultatene kan skyldes at OLS ikke skiller mellom individuelle bedriftsforskjeller, og dermed ikke klarer å isolere forklaringsvariabelens effekt i like stor grad som fixed effects-modellen.

Beeswarmplottene for prediksjonsmodellen med XGBoost og PyCaret, viser at høye verdier av kapitalens omløpshastighet, over 2.5, påvirker ROA i positiv retning. Resultatene understøttes av PDP og ICE-kurven, henholdsvis vist i figur 5.10 og 5.11, som begge viser en positiv lineær sammenheng mellom kapitalens omløpshastighet og ROA. Dette gjelder særlig for verdier mellom 1 og 2.5, som viser en bratt stigning i lønnsomhet ved økning i kapitalens omløpshastighet. Beeswarmplottet til klassifiseringsmodellen viser en kompleks sammenheng, og det er derfor krevende å tolke hvordan ulike verdier av variabelen påvirker klassifiseringen ved denne alene. PDP for klassifiseringsmodellen i figur 5.28, viser til en ikke-lineær sammenheng mellom kapitalens omløpshastighet og klassifisering, og at verdier marginalt bedre enn 1 er assosiert med økt sannsynlighet for at en bedrift klassifiseres som «lønnsom», men at høyere verdier enn det påvirker lønnsomheten i mindre grad.

XGBoost-modellene og PyCaret har identifisert likviditetsgrad 1 og 2 som moderat viktige forklaringsvariabler. Beeswarmplottene til prediksjonsmodellen med XGBoost og PyCaret indikerer at begge disse variablene har begrenset påvirkning på ROA, men med tendenser til å påvirke ROA i ulike retninger. For likviditetsgrad 1 kan resultatene fra beeswarmplottet til prediksjonsmodellen indikere at lave verdier, under 1.5, predikerer høyere lønnsomhet, men at høye verdier synes å ha mindre påvirkning på lønnsomheten. Beeswarmplottet til PyCaret viser derimot at både lave og høye verdier av likviditetsgrad 1 kan påvirke lønnsomheten i både positiv og negativ retning, og at det kan være en moderat til noe negativ påvirkning for høye verdier, over 1.5, for likviditetsgrad 1 på lønnsomheten. Lorentzen og Bergander (2022) fant en positiv sammenheng mellom likviditetsgrad 1 og lønnsomhet i sine regresjonsmodeller. Resultatene i denne oppgaven viser imidlertid at det er en mer kompleks sammen-

heng, og at verdier opp til 1.5 i likviditetsgrad 1 kan indikere høyere lønnsomhet, men at høyere verdier enn dette gir liten til noe negativ effekt på lønnsomheten. Det kan derfor vise til at det er en balanse mellom det å ha tilstrekkelig likviditet, uten at det påvirker lønnsomheten negativt.

For likviditetsgrad 2 viser resultatene fra prediksjonsmodellen med XGBoost og PyCaret at lave verdier, under 1.37, har ulik effekt på lønnsomheten, og at høye verdier imidlertid kan ha negativ påvirkning. Tolkningen er derfor i stor grad lik som likviditetsgrad 1. Beeswarmplottet til klassifiseringsmodellen indikerer at høye verdier øker sannsynligheten for å bli klassifisert som «ikke lønnsom», og lave verdier gir økt sannsynlighet for klasse «lønnsom». Knardal og Sending (2022) hevder likviditetsgrad 2 burde være minst 1, men resultatene antyder at dersom verdien er en del høyere enn dette kan det også være ufordelaktig for lønnsomhet.

Arbeidskapital blir også pekt ut av LASSO til å være en viktig forklaringsvariabel med negativ koeffisient, som tyder på at en økning i arbeidskapital vil ha negativ påvirkning på lønnsomheten. Resultatene fra beeswarmplottene til prediksjonsmodellen med XGBoost og PyCaret, tyder på at lave verdier for arbeidskapital, under 8.1 millioner, kan predikere høyere ROA. Det er vanskelig å tolke i hvilken retning de høye verdiene påvirker ROA, og det tyder på at det er en mer kompleks sammenheng mellom arbeidskapital og ROA. Beeswarmplottet til klassifiseringsmodellen antyder at høye verdier av arbeidskapital kan øke sannsynligheten for klasse «ikke lønnsom», mens lave verdier påvirker klassifiseringen i liten grad. PDP til arbeidskapital i 5.12 illustrerer at høyere verdier av arbeidskapital har negativ sammenheng med lønnsomhet, med en noe slakkere nedadgående linje for arbeidskapital over syv millioner.

Arbeidskapital i % av driftsinntekt blir pekt ut av LASSO til å være en viktig forklaringsvariabel med negativ koeffisient, som tyder på at en økning av denne variabelen vil redusere den predikerte lønnsomheten. Begge XGBoost-modellene rangerer arbeidskapital i % av driftsinntekt lavt og resultatene tyder på at dette er en variabel som i liten grad påvirker ROA. PyCaret velger ikke arbeidskapital i % av driftsinntekter blant de 20 viktigste forklaringsvariablene. Dette kan muligens skylde at vi har to mål på arbeidskapital i modellene.

Samlet sett indikerer resultatene at variablene som måler likviditet er viktige for å forklare lønnsomhet, noe som er i tråd med funnene til Jolly Cyril og Singla (2020) og Lorentzen og Bergander (2022), som mener god likviditetsstyring er viktig for lønnsomhet. Funnene tyder på at høye verdier for likviditetsgrad 1 og 2, samt arbeidskapital og arbeidskapital i % av driftsinntekter, er forbundet med dårligere lønnsomhet. Dette virker i utgangspunktet lite intuitivt ettersom god likviditet ofte anses å være positivt, fordi det sikrer at bedriften kan møte sine kortsiktige forpliktelser. En mulig forklaring på dette kan være at bedrifter med likviditet over gjennomsnittet i utvalget, holder av en betydelig del av sine ressurser i likvide midler med lav avkasting, slik som kontanter, i stedet for å investere ressursene i mer inntektsgenererende aktiviteter.

Soliditet

Egenkapitalandel, finanseringsgrad 1, rentedekningsgrad og gjeldsgrad benyttes for å se på forholdet mellom soliditet og lønnsomhet for bedriftene i utvalget.

Rentedekningsgrad blir pekt ut av LASSO til å være en viktig forklaringsvariabel, med positiv koeffisient. Koeffisienten til rentedekningsgrad er positiv i både OLS og fixed effects, noe som indikerer at en økning av denne variabelen er positivt assosiert med lønnsomhet. En høy rentedekningsgrad tilsier at bedriften genererer tilstrekkelig inntekt til å enkelt dekke sine finanskostnader, og indikerer god økonomisk helse. Resultatene fra beeswarmplottene til prediksjonsmodellen med XGBoost og PyCaret, tyder på at høye verdier, over gjennomsnittet på 71 617, for rentedekningsgrad indikerer høyere lønnsomhet, mens lave verdier verken tydelig indikerer høyere eller lavere lønnsomhet. Dette kommer trolig av at gjennomsnittet til rentedekningsgrad er svært høyt, men med median på 19.59. Det tyder på en kraftig skjevhet i dataen, og de fleste observasjonene befinner seg under gjennomsnittet som lave verdier i beeswarmplottene. I klassifiseringsmodellen kan beeswarmplottet også vise at for de fleste observasjonene, vil høyere rentedekningsgrad øke sannsynligheten for klasse «lønnsom».

Gjeldsgrad blir også pekt ut til å være en relativt viktig forklaringsvariabel av LASSO, med en positiv koeffisient. I OLS og fixed effects er koeffisientene for gjeldsgrad positive, noe som indikerer at en økning av denne variabelen er positivt korrelert med lønnsomhet. For prediksjonsmodellen med XGBoost og PyCaret, tyder beeswarmplottene på at gjeldsgrad påvirker ROA i liten grad, og at det ikke foreligger en lineær sammenheng. Likevel kan det være tendenser til at høye verdier av gjeldsgrad kan predikere høyere ROA. Dette understøttes av beeswarmplottet til klassifiseringsmodellen, som antyder at høyere verdier kan assosieres med høyere sannsynlighet for klasse «lønnsom». Dette står i tråd med funnene til Jolly Cyril og Singla (2020) som hevder høyere andeler av gjeld har en positiv sammenheng med lønnsomheten, men det går imot funnene til Ab. Halim mfl. (2014) som fant et negativt forhold mellom høyere andeler av gjeld og lønnsomhet. En mulig forklaring på den positive sammenhengen mellom gjeldsgrad og lønnsomhet kan være dersom investeringsavkastningen overstiger gjeldskostnaden ved opptak av gjeld, slik som Jolly Cyril og Singla (2020) argumenterer for i sin undersøkelse. De trekker også frem at rentekostnadene reduserte skatten, som reduserte kostnaden av gjeld sammenlignet med egenkapital.

LASSO peker ut egenkapitalandel som den nest viktigste forklaringsvariabelen, og den har positiv koeffisient. Variabelen har også positive koeffisienter i både OLS og fixed effects, som indikerer at en økning i egenkapitalandelen kan indikere høyere ROA. En mulig forklaring på den positive sammenhengen kan være at dersom bedrifter har høy ROA, vil det indikere at bedriften genererer overskudd fra sine aktiva. Dersom bedriften holder tilbake dette overskuddet (ikke utdelt som utbytte) vil dette øke egenkapitalen, og dermed også egenkapitalandelen. Egenkapitalandel blir også pekt ut som en viktig forklaringsvariabel av begge XGBoost-modellene og PyCaret. Resultatene fra beeswarmplottene til prediksjonsmodellen med XGBoost kan indike-

re at lave verdier, under 0.3, er assosiert med høyere lønnsomhet. Beeswarmplottet til PyCaret indikerer at lave verdier er veldig volatile og kan påvirke i både positiv og negativ retning, og at høye verdier indikerer moderat til noe negativ retning for lønnsomhet. For klassifiseringsmodellen er det vanskelig å tolke hvordan ulike verdier av egenkapitalandel påvirker klassifiseringen. Selv om en høy egenkapitalandel ofte blir sett på som en indikator på økonomisk stabilitet, antyder funnene at en høy egenkapitalandel ikke nødvendigvis assosieres med høyere lønnsomhet. De motstridende signalene fra gjeldsgrad og egenkapitalandel kan indikere at mens bedrifter med høy egenkapitalandel kan oppleve økonomisk stabilitet, kan de imidlertid gå glipp av muligheter som bedrifter med høyere gjeldsgrad drar nytte av. I lys av dette, kan det se ut som strategisk bruk av gjeld kan bidra til økt lønnsomhet.

Finansieringsgrad 1 blir ikke trukket frem som en viktig forklaringsvariabel av LASSO, men begge XGBoost-modellene og PyCaret trekker den frem som en relativt viktig forklaringsvariabel. For både prediksjonsmodellen med XGBoost og PyCaret er påvirkningen krevende å tolke, og antyder en kompleks sammenheng mellom finansieringsgrad 1 og ROA. Imidlertid kan beeswarmplottet for prediksjonsmodellen med XGBoost antyde at lave verdier av finansieringsgrad 1, under gjennomsnittet på 231 477, assosieres med lavere lønnsomhet. På en side virker dette fornuftig, fordi lave verdier for finansieringsgrad 1 tilsier at store deler av omløpsmidlene er langsiktig finansiert, og det medfører høyere finansieringskostnader enn ved kortsiktig finansiering. Derimot er gjennomsnittet til likviditetsgrad 1 ekstremt høyt, og lave verdier i modellen er ikke det samme som det man generelt anser som lave verdier for finansieringsgrad 1. Knardal og Sending (2022) sier at finansieringsgrad 1 bør være rundt 2, og dette er mye lavere enn gjennomsnittet til denne oppgaven. Medianen er imidlertid på 2.78, og det er derfor en ekstrem skjevfordeling i dataen. Ekstremverdier drar opp gjennomsnittet kraftig, samtidig som medianen er lav og indikerer at de fleste observasjonene har mye lavere finansieringsgrad 1 enn gjennomsnittet. På grunn av dette, er det utfordrende å tolke den reelle påvirkningen som finansieringsgrad 1 har på lønnsomheten. For klassifiseringsmodellen kan resultatene imidlertid indikere at verdier over gjennomsnittet kan øke sannsynligheten for å klassifiseres som «lønnsom». Dette virker mindre fornuftig ettersom høye verdier for finansieringsgrad 1 vil tilsi at anleggsmidlene er kortsiktig finansiert, noe som er et tegn på usunn finansiering og således en indikasjon på at virksomheten kan stå overfor utfordringer.

Samlet sett indikerer funnene at soliditet, målt gjennom rentedekningsgrad, egenkapitalandel, gjeldsgrad og finansieringsgrad 1, spiller en betydelig rolle i å forklare lønnsomhet blant bedriftene i utvalget. Høye verdier for rentedekningsgrad er positivt korrelert med høyere lønnsomhet, noe som understøtter viktigheten av god økonomisk helse. Selv om en høy egenkapitalandel ofte blir ansett som en indikator på økonomisk stabilitet, viser resultatene at det nødvendigvis ikke henger sammen med økt lønnsomhet. Høyere gjeldsgrad er assosiert med høyere lønnsomhet, noe som tyder på at strategisk bruk av gjeld kan føre til økt lønnsomhet ved å utnytte muligheter som gir høyere avkastning. Effekten finansieringsgrad 1 har på lønnsomheten er uklar, og det er behov for en dypere analyse for å forstå sammenhengen mellom denne variabelen og lønnsomhet.

6.2.2 Bransjespesifikke variabler

Produksjonsindeks og markedsandel har blitt brukt for å undersøke hvordan bransjespesifikke faktorer påvirker lønnsomheten til bedriftene i utvalget.

Produksjonsindeks blir ikke trukket frem som en viktig forklaringsvariabel i noen av modellene. Dette antyder at produksjonsindeks, som måler aktivitetsnivået i bransjen, ikke har en vesentlig påvirkning på lønnsomheten for bedriftene i utvalget. Dette funnet står i kontrast til Lorentzen og Bergander (2022), hvor produksjonsindeks var en signifikant forklaringsvariabel for den ene modellen.

Markedsandel ble brukt som et mål på selskapsstørrelse for de bedriftsspesifikke forholdene, men også for å analysere konkurranseintensiteten i bransjen. Tabell 3.4 viser at gjennomsnittlig markedsandel er 0.0348 %. Dette indikerer at bedriftene i utvalget har relativt små markedsandeler og dermed også liten individuell markedsmakt. Dette er på ingen måte overraskende ettersom vi har avgrenset til små- og mellomstore bedrifter. Ifølge Porter (2008) kan en høy grad av intern rivalisering, kjent ved et stort antall konkurrenter med lik størrelse og makt, ofte føre til redusert lønnsomhet. Den lave markedsandelen blant bedriftene i utvalget kan bidra til økt konkurranseintensitet, der det vil være krevende for SMB å påvirke eller dominere markedsforholdene. Dalsegg og Lidsheim (2021) trekker frem at i bygg- og anleggsbransjen, der produktdifferensieringen er lav, økes konkurranseintensiteten ytterligere. I slike markeder, argumenterer de for at det er vanskelig å oppnå konkurransefortrinn på andre måter enn å tilby lavest pris. Det kan derfor tenkes at den negative sammenhengen mellom markedsandeler og lønnsomhet skyldes at bedriftene, i forsøk på å øke sine markedsandeler, engasjerer seg i aggressive prisstrategier, som kan redusere lønnsomheten og potensielt lede til priskriger.

6.2.3 Makroøkonomiske variabler

For å undersøke om makroøkonomiske faktorer kan forklare lønnsomhetsforskjeller for bedrifter innen bygg og anlegg, har forklaringsvariablene gjennomsnittlig årlig styringsrente, endring i styringsrente gjennom året, valutakurs mellom euro og nok, drivstoffpriser og BNP blitt brukt.

De makroøkonomiske variablene ble ikke valgt ut av LASSO til å være blant de 13 viktigste forklaringsvariablene. Dette antyder at de makroøkonomiske variablene generelt sett har liten påvirkning på lønnsomheten til bedriftene i utvalget. Prediksjons- og klassifiseringsmodellen med XGBoost, peker likevel ut drivstoffpris som en relativt viktig forklaringsvariabel. Resultatet fra beeswarmplottet for klassifiseringsmodellen viser at variabelen i liten grad påvirker klassifiseringen. Imidlertid er det tendenser til at lave verdier for drivstoffpris kan øke sannsynligheten for at bedrifter blir klassifisert som «ikke lønnsom», mens høye verdier ikke påvirker klassifiseringen. For prediksjonsmodellen med XGBoost indikerer resultatene at lave verdier for drivstoffpris kan predikere lavere ROA, mens høye verdier påvirker ROA verken positivt eller negativt. Selv om det virker lite intuitivt, tyder resultatene altså på at lave drivstoffpriser kan føre til redusert lønnsomhet.

Blant makrovariablene, velger PyCaret ut drivstoffpris og gjennomsnittlig styringsrente, og rangerer dem henholdsvis som nummer 19 og 20 i viktighet, noe som understreker deres marginale rolle i å påvirke lønnsomheten. Videre viser resultatene fra PyCaret at det heller ikke foreligger noen lineær sammenheng mellom disse variablene og ROA, og at sammenhengen er kompleks. Klassifiseringsmodellen med XGBoost trekker frem valutakurs og gjennomsnittlig styringsrente som forklaringsvariabler, og rangerer dem som nummer 19 og 20 i viktighet. Dette indikerer at de påvirker klassifiseringen i svært liten grad.

Siden tidligere litteratur peker på at det er de bedriftsspesifikke faktorene som er av størst betydning, er det ikke overraskende at de makroøkonomiske variablene var mindre viktige i modellene. Whittington mfl. (2020) påpeker også at makroøkonomiske faktorer ikke nødvendigvis gir så god innsikt i lønnsomhetsvariasjoner innad i en bransje, men at de heller forklarer hvorfor noen bransjer oppnår høyere lønnsomhet enn andre. Det som var overraskende var at BNP, som indikator på den økonomiske utviklingen, ikke ble ansett som en viktig forklaringsvariabel av modellene, til tross for at Whittington mfl. (2020) fremhever den økonomiske utviklingen som særlig viktig.

I forklaringsmodellene benyttes verdier for makrovariablene i det samme regnskapsåret som vi beregner lønnsomheten, mens i de prediktive modellene er makrovariablene lagget ett år bak i tid. Det at makrovariablene ikke blir ansett som særlig viktige kan muligens skyldes at effekten de har på lønnsomheten tar lengre eller kortere tid enn det som er modellert. Ifølge Dalsegg og Lidsheim (2023) kan en økning i styringsrenten føre til redusert aktivitetsnivå i bransjen, men påpeker likevel at det er en treghet i forhold til forventninger i markedet. Det kan derfor hende at for eksempel justeringer i styringsrenten har en effekt på lønnsomheten til bedriftene i utvalget, men at det er en treghet, som hadde krevd flere laggede variabler. En annen grunn kan være at effekten som makrovariablene har på lønnsomheten er mer indirekte og derfor mindre synlig i modellene.

Samlet sett tyder dette på at selv om makroøkonomiske variabler som drivstoffpris, gjennomsnittlig styringsrente og valutakurs kan ha en marginal effekt på ROA, er deres innflytelse svært begrenset. Dette antyder at de makroøkonomiske variablene ikke er avgjørende for å evaluere lønnsomhet blant små- og mellomstore bedrifter i bygg- og anleggsbransjen.

Denne siden er blank med hensikt

7 Konklusjon

I denne masteroppgaven har vi undersøkt problemstillingen: *Hva er de viktigste lønnsomhetsdriverne for norske SMB i bygg- og anleggsbransjen?* Formålet er å utforske lønnsomhet slik at det kan bidra til bedre økonomisk forståelse og færre konkurser blant bygg- og anleggsbedrifter. I studien benyttes forklaringsmodellene OLS og fixed effects i sammenheng med LASSO, for å analysere lønnsomhetsdriverne i det samme året som lønnsomhetsmålet ROA. I tillegg ble det utviklet prediktive modeller der XGBoost og PyCaret predikerer kontinuerlig ROA, samt at XGBoost benyttes til å klassifisere om en bedrift kommer til å være «lønnsom» eller «ikke lønnsom» det neste året.

Vi har fokusert på bedriftsspesifikke, bransjespesifikke og makroøkonomiske variabler, og kommer frem til at det er de bedriftsspesifikke variablene som best forklarer lønnsomhet blant små- og mellomstore bedrifter i bransjen. Blant de viktigste lønnsomhetsdriverne, har produktivitet, som mål på kapasitetsutnyttelse, særlig vist seg å være en viktig forklaringsvariabel. Den har en tydelig positiv påvirkning på lønnsomhet, noe som understreker viktigheten av produktivitet blant de ansatte. Lønnskostnad i % av driftsinntekt blir også fremhevet som en viktig forklaringsvariabel for kapasitetsutnyttelse. Resultatene peker på at en høyere andel lønnskostnad i forhold til driftsinntekt kan føre til økt lønnsomhet, noe som antyder at høyere investeringer i menneskelige ressurser kan føre til økt lønnsomhet i en arbeidsintensiv bransje. Kapitalens omløpshastighet har også pekt seg ut som en viktig forklaringsvariabel og har positiv sammenheng med lønnsomhet, som tyder på at bedrifter som er effektive i å omgjøre sine aktiva til inntekter vil være mer lønnsomme.

Likviditetsgrad 1 og 2, samt arbeidskapital måler likviditet, og vi kommer frem til at høye verdier, som er verdier over gjennomsnittet, er forbundet med lavere ROA. Dette kan tyde på at det er en balanse mellom å ha tilstrekkelig likviditet uten at det skal gå på bekostning av lønnsomheten. Når det gjelder variablene som måler soliditet, finner vi at rentedekningsgrad, egenkapitalandel og gjeldsgrad er viktige forklaringsvariabler for lønnsomhet. Forholdet mellom egenkapitalandel og gjeldsgrad er særlig interessant, og resultatene tyder på at strategisk bruk av gjeld kan øke lønnsomheten, og at høye egenkapitalandeler ikke nødvendigvis er assosiert med høyere lønnsomhet.

Videre tyder funnene på at geografisk plassering kan ha en innvirkning på lønnsomheten, og modellene viser at bedrifter som er lokalisert i Oslo er assosiert med høyere lønnsomhet, men at bedrifter som er lokalisert i Hordaland og Møre og Romsdal er assosiert med lavere lønnsomhet. Funnene viser i tillegg at erfaring, målt gjennom bedriftens alder, ikke nødvendigvis påvirker lønnsomheten, men at unge bedrifter er mer volatile når det kommer til lønnsomhet, og vil derfor oftere assosieres med

dårligere lønnsomhet.

Resultatene viser at selv om mange av de bedriftsspesifikke variablene har signifikante effekter på lønnsomheten, varierer betydningen mellom modellene. Eksempelvis, har variablene antall årsverk og omsetning, som måler selskapsstørrelse, vist seg å ha både positive og negative påvirkninger på lønnsomheten, avhengig av hvilken modell som blir brukt. Det understreker at det er komplekse sammenhenger mellom selskapsstørrelse og lønnsomhet. Markedsandel har vist en negativ sammenheng med lønnsomhet, noe som kan tyde på at en økt markedsandel kan lede til priskonkurranse og høyere kostnader, noe som kan redusere lønnsomheten.

Resultatene viser at de bransjespesifikke og makroøkonomiske variablene generelt hadde lite påvirkning på lønnsomheten, sammenlignet med de bedriftsspesifikke variablene. Produksjonsindeksen viser ingen signifikant påvirkning på lønnsomheten, men makroøkonomiske variabler som drivstoffpris, valutakurs og gjennomsnittlig styringsrente har marginal effekt på lønnsomheten.

7.1 Begrensninger ved studien

I kvantitative oppgaver som benytter mye data vil det alltid være en utfordring knyttet til om dataen er korrekt. I denne studien benyttes årsregnskap for norske AS, der de fleste årsregnskapene trolig har blitt revidert. Det er likevel alltid en mulighet for at noen av tallene ikke er korrekte. Datasettet inneholder også data for årene som ble preget av covid-19, og dette har det ikke blitt tatt hensyn til i oppgaven, med unntak av at vi benytter *rolling window* for å trene på data som er nærmere testsettene. Når det gjelder forklaringsmodellene, er en begrensning knyttet til valg av straffeparameter λ i LASSO. Det ble testet for ulike verdier av λ for å redusere antallet variabler til mellom 10-15, hvor vi valgte å benytte de 13 viktigste variablene. Dette var en grense som vi selv valgte, og det er derfor en begrensning i studien ved at det er uvisst om det å velge flere variabler kunne ha gitt bedre modeller, selv om disse variablene ville blitt ansett som mindre viktige.

En viktig begrensning i oppgaven er knyttet til håndtering av uteliggere. Vi valgte å gjennomføre en winsorizing for 1 % og 99 % persentil for å unngå å slette observasjoner. Dette ble beskrevet i kapittel 3. Ettersom dataen består av årsregnskap som trolig er revidert, antok vi også at verdiene i hovedsak er korrekte, selv ved høye og lave verdier i enkelte regnskapsposter. Begrensningen er likevel i hovedsak tilknyttet beregning av de regnskapsbaserte nøkkeltallene som beregnes ved divisjon. Ved beregning av nøkkeltall der nevneren er null, får nøkkeltallet verdien uendelig eller minus uendelig. Dette tok vi imidlertid hensyn til ved at de regnskapsbaserte nøkkeltallene med verdier uendelig eller minus uendelig ble satt til maksimum og minimumsverdi for de variablene det gjelder i det samme året, før winsorizingen.

Begrensninger ved uteliggere er særlig knyttet til nøkkeltallet finansieringsgrad 1, hvor omtrent 9400 observasjoner hadde null i langsiktig gjeld, og rentedekningsgrad som har omtrent 2500 observasjoner med null i sum finanskostnader. Det betyr at en større andel av observasjonene for disse forklaringsvariablene, ble satt til maksimum-

eller minimumsverdi. Dette er flere observasjoner enn de som ble justert i winsorizingen. Dette har skjevfordelt disse variablene, som man kan observere i tabell 3.4. Finansieringsgrad 1 har median på 2.78, og gjennomsnitt på 231477, mens rentedekningsgrad har median på 19.59 og gjennomsnitt på 71617. Finansieringsgrad 1 har anleggsmidler i teller, og har derfor kun hatt problemer med uendelige verdier. Rentedekningsgrad kan ha hatt både uendelige og minus uendelige, men den deskriptive statistikken viser at det i hovedsak har vært problemer knyttet til de aller høyeste verdiene, altså at det har vært flere uendelige verdier. Medianen viser imidlertid fornuftige verdier, som viser at de fleste observasjonene ikke er uteliggere. Denne skjevfordelingen er likevel en større begrensning i oppgaven.

En annen begrensning i oppgaven er knyttet til beregningenes kompleksitet. I maskinlæringsmodellene ble det gjennomført to forenklinger knyttet til begrensninger i beregningskraft og tid, i henholdsvis XGBoost og PyCaret. I XGBoost-modellene er hyperparametere tilpasset ved bruk av *RandomSearch*, med 100 tilfeldig genererte iterasjoner. Vedlegg E gir en oversikt over hyperparametere valgt for alle periodene i *rolling window*. I disse tabellene kan vi se at det generelt er valgt ulike hyperparametere for de ulike periodene. Det kan muligens indikere at rutenettet for tilpasningen er for stort for antallet iterasjoner, og at anvendelse av *GridSearch* eller flere iterasjoner i *RandomSearch* hadde gitt bedre resultater. Det er likevel utført en avveining i denne oppgaven i forhold til beregningskraft og tid, særlig med tanke på at vi har utført modeller for ni ulike perioder både i prediksjon av kontinuerlig ROA, og klassifiseringen.

Det er også en mulig svakhet i beregningenes kompleksitet ved anvendelse av PyCaret. Ved utvelgelse av metode for de ulike periodene i *rolling window*, har vi valgt å benytte standardoppsett for PyCaret, som ekskluderer de mest beregningstunge metodene. Det er derfor mulig at beregningstunge metoder, som f.eks. nevrale nettverk ville gitt bedre resultater ved bruk av PyCaret i noen av periodene. I oppgaven testet vi med alle metodene, inklusive de beregningstunge, for nesten hele datasettet som en test. Denne testen valgte gradient boosting regressor som den beste metoden, og vi valgte derfor å ekskludere de beregningstunge metodene da vi utviklet modeller for alle periodene i *rolling window*.

7.2 Videre forskning

I denne studien har vi begrenset forskningen til bygg- og anleggsbransjen, og fokusert på kun små- og mellomstore bedrifter. Det kunne derfor vært interessant å undersøkt om funnene er representative også for andre sektorer, både i Norge og internasjonalt. Det kunne også vært interessant å utvidet eller redusert avgrensningen i forhold til størrelse, som f.eks. å avgrense studien ytterligere til å kun inkludere mellomstore bedrifter. Det kunne trolig også vært nyttig å gjøre tilsvarende analyser som inkluderer andre type variabler, eksempelvis ikke-finansiell informasjon som alder på daglig leder, andel kvinner i ledelsen osv. Dette strekker seg også til utvelgelse av andre regnskapsbaserte nøkkeltall enn de vi benytter i denne oppgaven.

I forbindelse med klassifiseringsmodellen, er et forslag til videre forskning å se på ulike avgrensninger i klasser. Vi har valgt å definere bedrifter med ROA over 2 % som «lønnsom», og bedrifter under denne grensen som «ikke lønnsom». I videre forskning kunne man utført en sensitivitetsanalyse knyttet til denne avgrensningen i klasser. Det hadde kanskje særlig vært interessant å klassifisere om en bedrift vil få veldig bra lønnsomhet eller ikke, ettersom vi i tidlige tester fant at de fleste bedriftene ble klassifisert som enten dårlig eller veldig bra med inndeling i fire klasser. Et forslag er også å teste ytterligere med inndeling i flere klasser som dårlig, moderat, bra og veldig bra lønnsomhet.

I oppgaven har vi blant annet sett at de prediktive modellene gjør det noe svakere i de senere årene, og disse inkluderer år som var påvirket av covid-19. I videre forskning kan man derfor utforske effektene av covid-19 på bedrifters lønnsomhet. Dette temaet kan også knyttes til hvordan de statlige støtteordningene påvirket lønnsomheten under og etter pandemien. Vi har også begrensninger i studien knyttet til, særlig finansieringsgrad 1, siden mange verdier manglet langsiktig gjeld. Det kunne derfor vært interessant å utføre en grundigere analyse av dette nøkkeltallet, og undersøke hvorfor flere bedrifter ikke har langsiktig gjeld.

Vi finner også forskjeller i lønnsomhet for ulike lokaliseringer, men går ikke i dybden på hvorfor. I videre forskning hadde det derfor vært interessant å analysere hva som er grunnene til ulik lønnsomhet, ved å eksempelvis benytte befolkningstetthet, vær, tilgang på arbeidskraft osv. som variabler for å fange opp ulikheter i lokalisering. Vi finner også at makrovariabler har liten effekt på lønnsomhet, men ser kun på makroverdier med ett års *lag* i de prediktive modellene. Videre forskning kan derfor utvikle prediktive modeller som har flere *lag* i makrovariabler, for å analysere om eksempelvis styringsrenten påvirker lønnsomheten med lengre tidshorisont. Det kunne eventuelt også ha blitt gjennomført en sensitivitetsanalyse med språkbehandling for å se om oppfatninger og forventninger til markedet vil gi mer nøyaktige resultater for prediksjon av lønnsomhet.

Referanser

Ab. Halim, M. S., Jusoh, M., Osman, N., & Amlus, H. (2014). Determining the Financial Performance Factors Among Bumiputera Entrepreneurs in Malaysian Construction Industry. *Australian Journal of Basic and Applied Sciences*, *8*, 18–24.

Akintoye, A., & Skitmore, M. (1991). Profitability of UK construction contractors. *Construction Management and Economics*, *9*. https://doi.org/10.1080/01446199100000025

Altinn. (2024, 1. januar). *Hva koster en arbeidstaker*. Hentet 19. mars 2024, fra https://info.altinn.no:443/starte-og-drive/arbeidsforhold/ansettelse/hva-koster-en-arbeidstaker/

Barney, J. B. (2006). *Strategic management and competitive advantage: concepts and cases*. Pearson/Prentice Hall.

Berg, T. (2021). *Grunnleggende økonomistyring* (3. utgave.). Cappelen Damm akademisk. Hentet 8. mars 2024, fra https://www.nb.no/search?q=oaiid:%22oai:nb.bibsys.no:999920149477502202%22&mediatype=b%C3%B8ker

Bhattacharya, A. (2022). *Applied machine learning explainability techniques: make ML models explainable and trustworthy for practical applications Using LIME, SHAP, and more*. Packt Publishing.

Brownlee, J. (2021, 19. januar). *Regression metrics for machine learning* [MachineLearningMastery.com]. Hentet 8. mars 2024, fra https://machinelearningmastery.com/regression-metrics-for-machine-learning/

Brønnøysundregistrene. (2024 februar). *Bedrifts- og foretaksstatistikk* [Brønnøysundregistrene]. Hentet 18. mars 2024, fra https://www.brreg.no/produkter-og-tjenester/statistikk/bedrifts-og-foretaksstatistikk/

Bygballe, L. E., Grimsby, G., Engebretsen, B. E., & Reve, T. (2019). En verdiskapende bygg-, anlegg- og eiendomsnæring (BAE): Oppdatering 2019 [Accepted: 2019-11-20T08:25:24Z Publisher: Handelshøyskolen BI]. *2*. Hentet 12. mars 2024, fra https://biopen.bi.no/bi-xmlui/handle/11250/2629396

Capon, N., Farley, J., & Hoenig, S. (1990). A Meta-Analysis of Financial Performance. *Management Science*, *36*, 1143–1159. https://doi.org/10.1287/mnsc.36.10.1143

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Çorbacıoğlu, Ş. K., & Aksel, G. (2023). Receiver operating characteristic curve analysis in diagnostic accuracy studies: A guide to interpreting the area under

the curve value. *Turkish Journal of Emergency Medicine*, *23*(4), 195–198. https://doi.org/10.4103/tjem.tjem_182_23

Dalsegg, H., & Lidsheim, T. (2021). BDO - Bygg og anleggsanalysen. Hentet 22. mars 2024, fra https://issuu.com/konsis/docs/bygg-_og_anleggsanalysen

Dalsegg, H., & Lidsheim, T. (2023 oktober). *Bygg- og anleggsanalysen 2023*. Hentet 16. januar 2024, fra https://issuu.com/konsis/docs/bygg-_og_anleggsanalysen_2023

Dash, S. (2020 desember). Brier Score - How to measure accuracy of probablistic predictions. Hentet 15. april 2024, fra https://www.machinelearningplus.com/statistics/brier-score/

Deloof, M. (2003). Does Working Capital Management Affect Profitability of Belgian Firms? [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1468-5957.00008]. *Journal of Business Finance & Accounting*, *30*(3-4), 573–588. https://doi.org/10.1111/1468-5957.00008

Ding, K., Lev, B., Peng, X., Sun, T., & Vasarhelyi, M. A. (2020). Machine learning improves accounting estimates: Evidence from insurance payments. *Review of Accounting Studies*, *25*(3), 1098–1134. https://doi.org/10.1007/s11142-020-09546-9

Directorate-General for Internal Market, Industry, Entrepreneurship and SMEs (European Commission). (2015). *User guide to the SME definition*. Publications Office of the European Union. Hentet 18. mars 2024, fra https://data.europa.eu/doi/10.2873/620234

Finansdepartementet. (2011, 25. januar). *NOU 2011: 1* [Regjeringen.no] [Publisher: regjeringen.no]. Hentet 19. mars 2024, fra https://www.regjeringen.no/no/dokumenter/nou-2011-1/id631151/

Fortmann-Roe, S. (2012 juni). *Understanding the Bias-Variance Tradeoff*. Hentet 28. april 2024, fra https://scott.fortmann-roe.com/docs/BiasVariance.html

Friedman, J. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*, *29*, 1189–1232. https://doi.org/10.2307/2699986

Føyen, L. J., & Danielsen, T. L. (2020). *Lønnsomhet i bygg- og anleggsbransjen : en studie av lønnsomhetsdrivere i store norske byggog anleggsselskap* [Masteroppgave] [Accepted: 2020-09-21T09:55:35Z]. Hentet 22. februar 2024, fra https://openaccess.nhh.no/nhh-xmlui/handle/11250/2678738

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation [Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10618600.2014.90 *Journal of Computational and Graphical Statistics*, *24*(1), 44–65. https://doi.org/10.1080/10618600.2014.907095

Hargrave, M. (2024). Return on Assets (ROA): Formula and 'Good' ROA Defined. Hentet 3. april 2024, fra https://www.investopedia.com/terms/r/returnonassets.asp

Hellesnes, P. (udatert). *Postnummer (Norge)* [BEDREINNSIKT]. Hentet 18. mars 2024, fra http://www.bedreinnsikt.no/datasett-postnummer.html

Hill, R. C. (2012). *Principles of econometrics* (4th ed.). Wiley.

IBM. (udatert). *What is machine learning? | IBM*. Hentet 19. februar 2024, fra https://www.ibm.com/topics/machine-learning

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). Statistical Learning. I *An Introduction to Statistical Learning: with Applications in Python* (s. 25–27). Springer International Publishing. https://doi.org/10.1007/978-3-031-38747-0_1

Jolly Cyril, E., & Singla, H. K. (2020). Comparative analysis of profitability of real estate, industrial construction and infrastructure firms: evidence from India [Publisher: Emerald Publishing Limited]. *Journal of Financial Management of Property and Construction*, *25*(2), 273–291. https://doi.org/10.1108/JFMPC-08-2019-0069

Knardal, P. S., & Sending, A. (2022). *Finansregnskap med analyse* (1. utgave.). Fagbokforlaget. Hentet 11. mars 2024, fra https://www.nb.no/search?q=oaiid:%22oai:nb.bibsys.no:999920180189702202%22&mediatype=b%C3%B8ker

Langli, J. C. (2010). *Årsregnskapet* (9. utg.). Gyldendal akademisk. Hentet 18. mars 2024, fra https://urn.nb.no/URN:NBN:no-nb_digibok_2011062905132

Long, J. S., & Ervin, L. H. (2000). Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model [Publisher: [American Statistical Association, Taylor & Francis, Ltd.]]. *The American Statistician*, *54*(3), 217–224. https://doi.org/10.2307/2685594

Lorentzen, M., & Bergander, H. J. (2022 mai). *Hva bygger lønnsomhet? En gjennomgang av lønnsomhetsfaktorene i bygg- og anleggsbransjen* [Masteroppgave, UiT Norges arktiske universitet] [Accepted: 2022-10-19T05:34:05Z]. Hentet 1. mars 2024, fra https://munin.uit.no/handle/10037/27067

Lundberg, S., & Lee, S.-I. (2017, 24. november). A Unified Approach to Interpreting Model Predictions. Hentet 23. februar 2024, fra http://arxiv.org/abs/1705.07874

McGahan, A. M., & Porter, M. E. (2002). What Do We Know about Variance in Accounting Profitability? [Publisher: INFORMS]. *Management Science*, *48*(7), 834–851. Hentet 1. mars 2024, fra https://www.jstor.org/stable/822694

Moez, A. (udatert). *Pycaret/tutorials/tutorial - regression.ipynb at master · pycaret/pycaret* [GitHub]. Hentet 26. april 2024, fra https://github.com/pycaret/pycaret/blob/master/tutorials/Tutorial%20-%20Regression.ipynb

Moez, A. (2020 april). *PyCaret: An open source, low-code machine learning library in Python* [PyCaret version 1.0.0]. https://www.pycaret.org

Moez, A. (2021 november). *PyCaret tutorial: A beginner's guide for automating ML workflows using PyCaret*. Hentet 29. februar 2024, fra https://www.datacamp.com/tutorial/guide-for-automating-ml-workflows-using-pycaret

Mudadla, S. (2023, 28. november). *Bias variance tradeoff in machine learning.* [Medium]. Hentet 7. mars 2024, fra https://medium.com/@sujathamudadla1213/bias-variance-tradeoff-in-machine-learning-a1856b55f6a9

NHO. (2024, 6. februar). *Tall og fakta om SMB*. Hentet 18. mars 2024, fra https://www.nho.no/tema/sma-og-mellomstore-bedrifter/tall-og-fakta-om-smb/

Nielsen, D. (2016). *Tree Boosting With XGBoost - Why Does XGBoost Win "Every" Machine Learning Competition?* [Master thesis]. NTNU [Accepted: 2017-03-13T07:58:50Z]. Hentet 6. mars 2024, fra https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/2433761

Norges Bank. (2024a, 25. januar). *Endringer i styringsrenten*. Hentet 18. mars 2024, fra https://www.norges-bank.no/tema/pengepolitikk/Styringsrenten/Styringsrenten-Oversikt-over-rentemoter-og-endringer-i-styringsrenten-/

Norges Bank. (2024b, 14. mars). *Valutakurser*. Hentet 18. mars 2024, fra https://www.norges-bank.no/tema/Statistikk/Valutakurser/

O'brien, R. M. (2007). A Caution Regarding Rules of Thumb for Variance Inflation Factors. *Quality & Quantity*, *41*(5), 673–690. https://doi.org/10.1007/s11135-006-9018-6

Paraschiv, F., Schmid, M., & Wahlstrøm, R. R. (2023, 20. september). Bankruptcy prediction of privately held SMEs using feature selection methods. https://doi.org/10.2139/ssrn.3911490

Patil, A. P., Deepshika, M. P., Mittal, S., Shetty, S., Hiremath, S. S., & Patil, Y. E. (2017). Customer churn prediction for retail business. *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, 845–851. https://doi.org/10.1109/ICECDS.2017.8389557

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Porter, M. E. (1979). How Competitive Forces Shape Strategy [ISSN: 0017-8012 Issue: 2 Num Pages: 137- Place: Boston Publication Title: Harvard business review Volume: 57].

Porter, M. E. (1985). *Competitive advantage: creating and sustaining superior performance*. Free Press.

Porter, M. E. (2008). The Five Competitive Forces That Shape Strategy [Publisher: Harvard Business School Publication Corp.]. *Harvard Business Review*, *86*(1), 78–93. Hentet 26. februar 2024, fra https://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=28000138&site=ehost-live

Sagi, O. (2024, 13. mars). *Model performance metrics for regression models | pecan help center*. Hentet 20. mars 2024, fra https://help.pecan.ai/en/articles/6456388-model-performance-metrics-for-regression-models

Sending, A. (2009). *Økonomistyring 1*. Fagbokforl. Hentet 9. april 2024, fra https://urn.nb.no/URN:NBN:no-nb_digibok_2020101407544

Shapley, L. S. (1951, 21. august). *Notes on the n-person game — II: The value of an n-person game*. RAND Corporation. Hentet 25. mars 2024, fra https://www.rand.org/pubs/research_memoranda/RM0670.html

Shmueli, G., Bruce Peter C., G., & Peter Patel, N. R. (2019). *Data Mining for Business Analytics: Concepts, Techniques and Applications in Python*. John Wiley Sons Inc.

SSB. (udatert-a). *09654: Priser på drivstoff (kr per liter) 1986M08 - 2024M01. Statistikkbanken* [SSB]. Hentet 18. mars 2024, fra https://www.ssb.no/system/

SSB. (udatert-b). *11417: Årslønn, etter næring (SN2007), statistikkvariabel og år. Statistikkbanken* [SSB]. Hentet 20. mars 2024, fra https://www.ssb.no/system/

SSB. (udatert-c). *13430: Produksjonsindeks for bygge- og anleggsvirksomhet, etter næring (SN2007) 2005M01 - 2024M01. Statistikkbanken* [SSB]. Hentet 18. mars 2024, fra https://www.ssb.no/system/

SSB. (udatert-d). *Alle endringer i de regionale inndelingene* [SSB]. Hentet 18. mars 2024, fra https://www.ssb.no/metadata/alle-endringer-i-de-regionale-inndelingene

SSB. (udatert-e). *Fakta om norsk økonomi* [SSB]. Hentet 18. mars 2024, fra https://www.ssb.no/nasjonalregnskap-og-konjunkturer/faktaside/norsk-okonomi

SSB. (udatert-f). Norges viktigste handelspartnere. Hentet 2. mai 2024, fra https://www.ssb.no/utenriksokonomi/utenrikshandel/statistikk/utenrikshandel-med-varer/artikler/norges-viktigste-handelspartnere

SSB. (2023 oktober). Næringenes økonomiske utvikling. Hentet 19. mars 2024, fra https://www.ssb.no/virksomheter-foretak-og-regnskap/virksomheter-og-foretak/statistikk/naeringenes-okonomiske-utvikling

Studenmund, A. H., & Johnson, B. K. (2017). *A Practical guide to using econometrics* (7th edition, global edition.). Pearson.

Tian, S., & Yu, Y. (2017). Financial ratios and bankruptcy predictions: An international evidence. *International Review of Economics & Finance*, *51*, 510–526. https://doi.org/10.1016/j.iref.2017.07.025

Tian, S., Yu, Y., & Guo, H. (2015). Variable selection and corporate bankruptcy forecasts. *Journal of Banking & Finance*, *52*, 89–100. 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

Visma. (udatert). Rentedekningsgrad. Hentet 11. april 2024, fra https://help.visma.net/no_no/financial-overview/content/online-help/kpi-interest-coverage-ratio.htm

Wahlstrøm, R. R. (2023, 17. november). Financial statements of companies in Norway [version: 4]. Hentet 15. mars 2024, fra http://arxiv.org/abs/2203.12842

Whittington, R., Johnson, G., Scholes, K., Angwin, D., & Regnér, P. (2020). *Exploring strategy* (Twelfth edition.). Pearson Education.

Zahedi, L., Mohammadi, F. G., Rezapour, S., Ohland, M. W., & Amini, M. H. (2021, 29. april). Search Algorithms for Automated Hyper-Parameter Tuning. https://doi.org/10.48550/arXiv.2104.14677

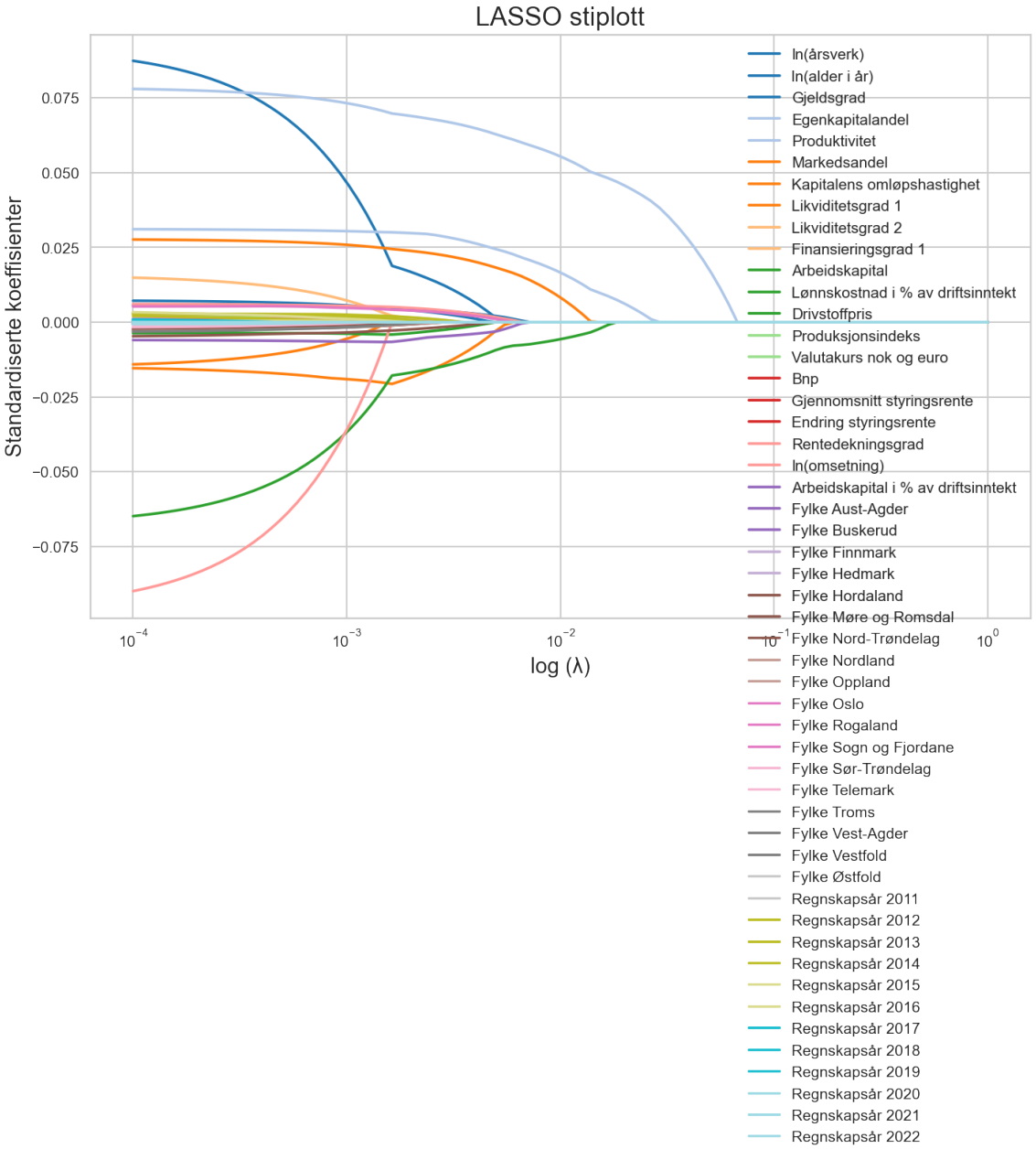
Denne siden er blank med hensikt

A Programvare

Programvare	Versjon
`python`	3.9.18
`stata`	18.0
`pandas`	1.4.2
`numpy`	1.21.5
`matplotlib`	3.5.1
`seaborn`	0.11.2
`statsmodels`	0.13.2
`scikit-learn`	1.0.2
`xgboost`	2.0.3
`scipy`	1.10.1
`shap`	0.44.1
`pycaret`	3.2.0

Denne siden er blank med hensikt

B LASSO stiplott for alle variabler



Figur B.1: LASSO stiplott for alle variabler

Denne siden er blank med hensikt

C Tester for regresjonsanalyse

Tabell C.1: VIF-test for de uavhengige variablene i OLS og fixed effects

Variabel	**VIF**
Produktivitet	7.361657
Lønnskostnad i % av driftsinntekt	4.985192
Egenkapitalandel	3.602805
Kapitalens omløpshastighet	5.782634
Gjeldsgrad	1.278594
Rentedekningsgrad	1.50274
Regnskapsår 2011	1.694065
Regnskapsår 2012	1.802539
Regnskapsår 2013	1.876985
Regnskapsår 2014	1.957567
Regnskapsår 2015	2.037798
Regnskapsår 2016	2.075267
Regnskapsår 2017	2.153080
Regnskapsår 2018	2.269118
Regnskapsår 2019	2.407179
Regnskapsår 2020	2.387107
Regnskapsår 2021	2.489161
Regnskapsår 2022	2.600619

Breusch Pagan viser at man forkaster nullhypotesen, og at det er oppdaget heteroskedastisitet. Det blir derfor kjørt robust regresjon.

Tabell C.2: Test for heteroskedastisitet ved Breusch Pagan

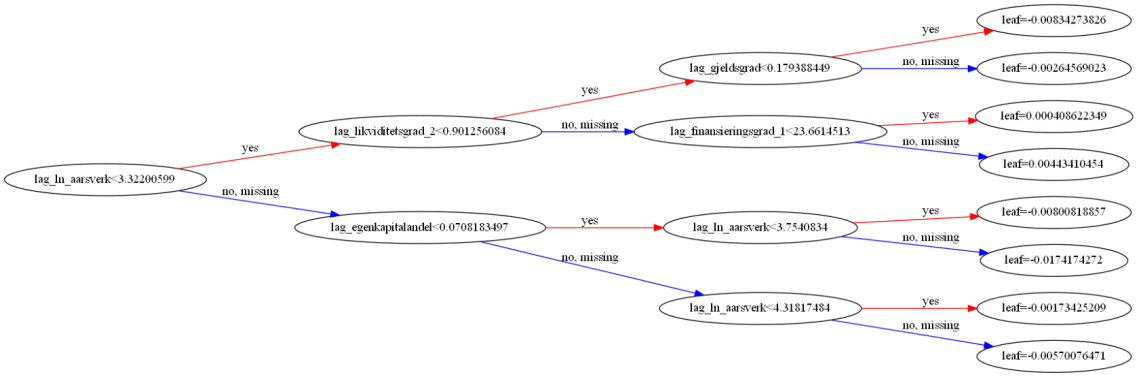
Lagrange-multiplikatorstatistikk	5932.4
Langrange-multiplikator p-verdi	0.0
F-verdi	393.6
P-verdi	0.0

Tabell C.3: Hausman-test

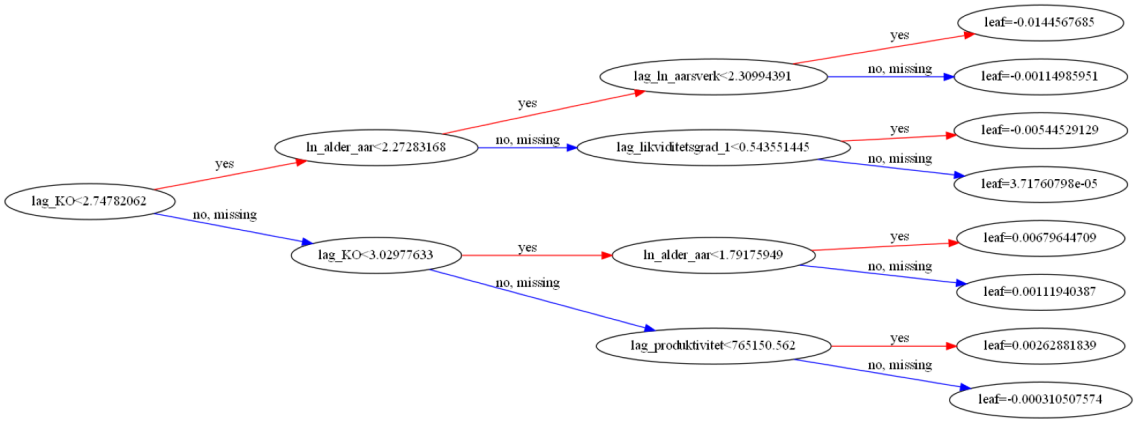
P-verdi	Tolkning
0.000	FE foretrekkes

D Eksempel på oppbygging av XGBoost

Figur D.1 og D.2 er et eksempel på hvordan XGBoost trener beslutningstrær ved hjelp av boosting i oppgaven. For perioden med treningssett fra 2014-2017 og testsett 2018 ser vi i D.1 beslutningstreet som er utgangspunktet, og deretter D.2 som er det siste treet, tre nr. 100, i perioden.



Figur D.1: Beslutningstre nr. 1 for prediksjon av ROA i periode med testsett i 2018



Figur D.2: Beslutningstre nr. 100 for prediksjon av ROA i periode med testsett i 2018

Denne siden er blank med hensikt

E Hyperparametere i XGBoost

Tabell E.1: Rutenett for hyperparametere i XGBoost modellen

Hyperparameter	**Rutenett**
`learning_rate`	[0.01, 0.1, 0.2]
`n_estimators`	[50, 100]
`max_depth`	[2, 3, 4, 5, 6]
`subsample`	[0.5, 0.75, 1]
`colsample_bytree`	[0.25, 0.5, 0.75, 1]
`min_child_weight`	[0, 1]
`reg_alpha`	[0, 0.1]
`min_split_loss/gamma`	[0, 0.1]
`seed`	123
`objective`	`reg:squarederror` og `binary:logistic`

Tabell E.2: Hyperparametere for XGBoost prediksjon av ROA

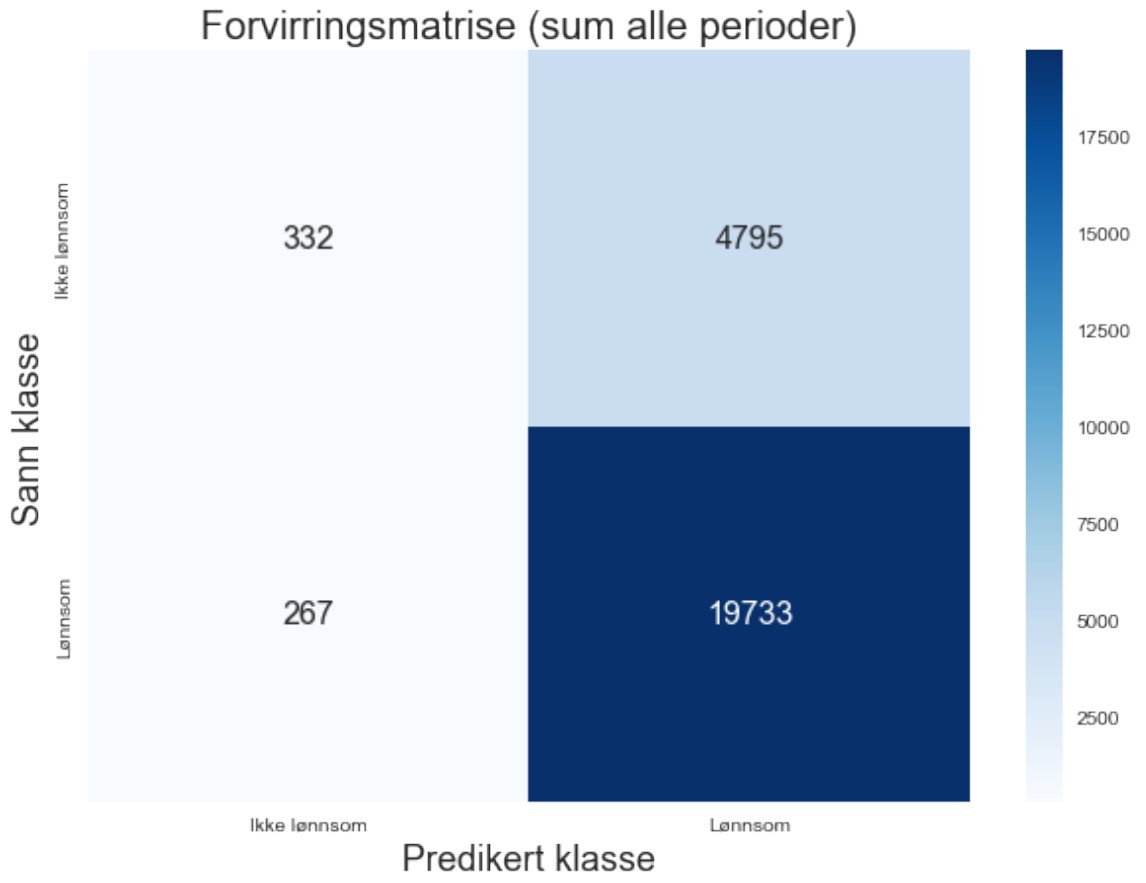
	2014	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**	**2022**
colsample_bytree	0.5	0.75	0.5	0.5	0.5	1	1	0.5	1
learning_rate	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
max_depth	5	5	5	5	3	5	5	4	5
min_child_weigth	0	1	0	0	0	0	0	1	0
min_split_loss	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
n_estimators	50	50	100	50	100	50	50	50	50
reg_alpha	0.1	0.1	0.1	0.1	0	0.1	0	0.1	0.1
subsample	0.75	0.5	1	0.75	0.75	1	0.75	1	1

Tabell E.3: Hyperparametere for XGBoost klassifisering av ROA

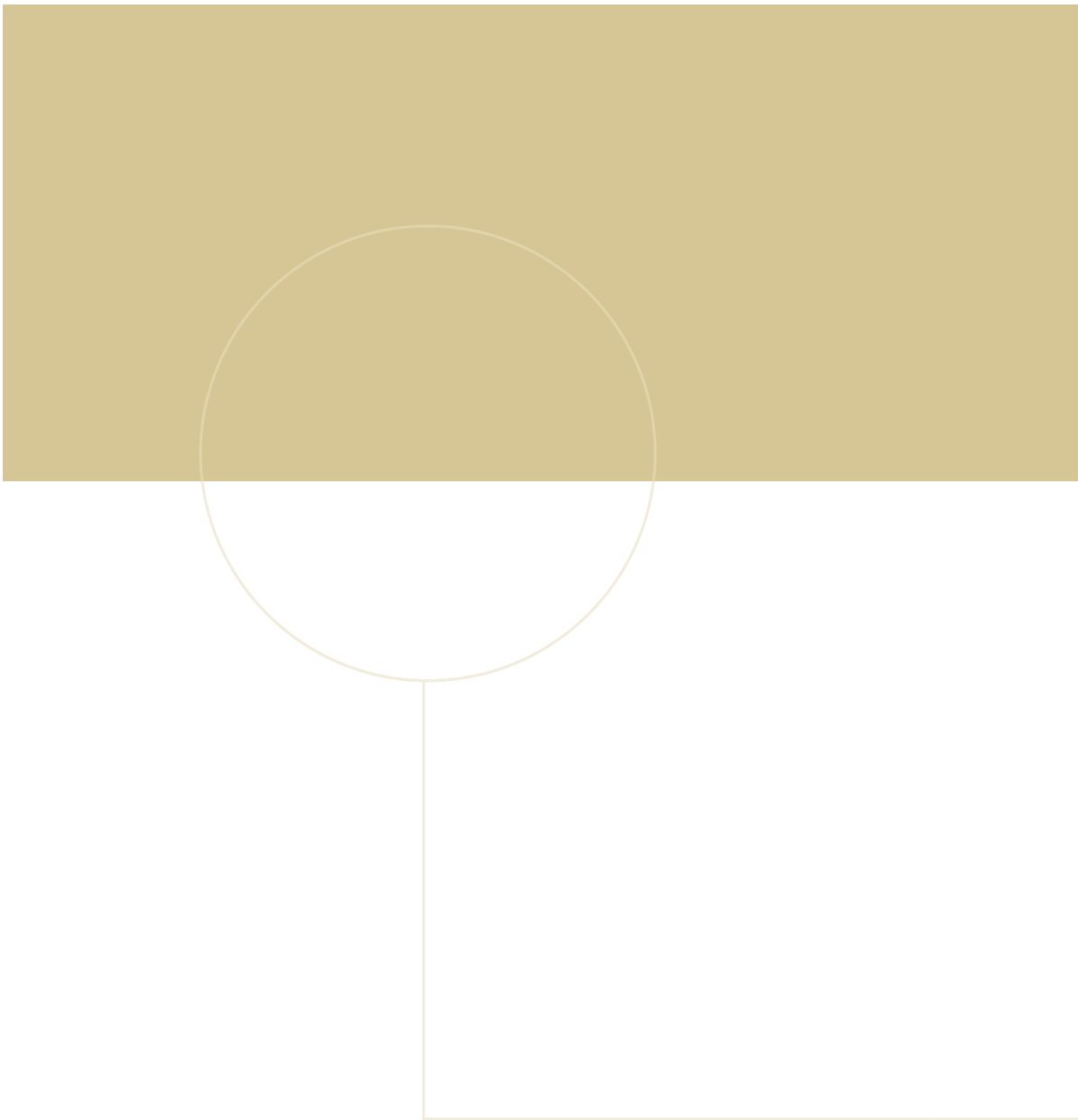
	2014	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**	**2022**
colsample_bytree	1	0.75	0.75	0.75	1	1	1	1	0.75
learning_rate	0.1	0.01	0.1	0.01	0.1	0.01	0.1	0.01	0.1
max_depth	3	4	3	5	3	5	3	5	3
min_child_weigth	1	0	1	0	1	0	1	0	1
n_estimators	50	100	50	50	50	100	50	100	50
reg_alpha	0	0.1	0	0	0	0	0	0	0
subsample	0.5	0.5	0.75	0.5	0.5	0.75	0.5	0.75	0.75
gamma	0	0	0	0.1	0	0.1	0	0.1	0.1

F Klassifiseringsplott

Forvirringsmatrisen i figur F.1 viser at de aller fleste observasjonene blir klassifisert som «lønnsom». Dette stemmer overens med at de fleste observasjonene i treningssettene er definert som «lønnsom». I forvirringsmatrisen er 332 observasjoner i testsettene riktig klassifisert som «ikke lønnsom», og 19733 riktig klassifisert som «lønnsom». Modellen predikerer 267 observasjoner som «ikke lønnsom» når den sanne klassen er «lønnsom», i tillegg til at modellen predikerer 4795 observasjoner som «lønnsom» når den sanne klassen er «ikke lønnsom».



Figur F.1: Forvirringsmatrise for XGBoost klassifiseringsmodell summert for alle testsettene i *Rolling Window*



NTNU

Kunnskap for en bedre verden