Aalborg University Business School

Kandidatafhandling

Machine Learning og Tværsnittet af Skandinaviske Aktiemarkedsafkast

Emilie Krogsgård Rom (20184802)

Vejleder: Peter Pommergård Lind

1. Juni 2023

Resumé

This thesis investigates whether machine learning techniques such as random forest and gradient boosting improve cross-sectional asset return forecasting statistically and economically compared to traditional OLS regression. The analysis deviates from extant literature in three ways: (1) addressing the U.S. bias by analyzing Scandinavian stocks, (2) including fewer variables to get an eye-level comparison, and (3) excluding micro caps. Surprisingly, this thesis finds OLS exhibiting superior predictive performance that translates into economic profitability relative to machine learning techniques. However, OLS is only marginally better than gradient boosting and all three portfolios generate significant Sharpe ratios. Reasons for the underperformance might include limitations in the data size and potential over-/underfitting. Nevertheless, there is moderate support for the results when looking at the deviations in the methodology (1-3) and when seeing the existing results in the literature through a critical lens.

AALBORG UNIVERSITET
STUDENTERRAPPORT

Titel:

Machine Learning og Tværsnittet af Skandinaviske Aktiemarkedsafkast

Projektperiode:

Forårssemesteret 2023

Forfatter:

Emilie Krogsgård Rom

<erom18@student.aau.dk>

Vejleder: Peter Pommergård Lind

Sidetal: 62,8

Afleveringsdato: 1. Juni 2023

Aalborg University Business School

https://www.business.aau.dk/

Indholdsfortegnelse

Kapitel 1 Indledning **1**

Kapitel 2 Litteraturgennemgang **4**

2.1 Aktivprissætning med Machine Learning 5

2.1.1 Den første strøm 5

2.1.2 Den anden strøm 5

2.2 Fortolkningen af Machine Learning i finanslitteraturen 7

2.3 Bidrag til litteraturen 8

Kapitel 3 Metoder og databeskrivelse **9**

3.1 Videnskabsteoretisk og metodisk overvejelse 9

3.2 Empiri 10

3.2.1 Softwareværktøjer 11

3.2.2 Databeskrivelse 11

Kapitel 4 Algoritmer og metodologi **15**

4.1 Machine Learning 15

4.2 Problemopsætning 17

4.3 Algoritmer 19

4.3.1 Simpel Lineær Regression 19

4.3.2 Et beslutningstræ 20

4.3.3 Random Forest 24

4.3.4 Gradient Boosting 28

4.3.5 Kombination af RF og GB 32

4.4 Hypertuning (Træning, Validering og Test) 34

4.5 Evaluering 36

4.6 Porteføljesammensætning 38

4.6.1 Risikojusterede afkast (alfa) . . . 39

Kapitel 5 Empiriske resultater **40**

5.1 In-Sample test . . . 40

5.1.1 Variables betydning . . . 40

5.2 Out-of-sample test . . . 42

5.2.1 Prædikteringsevne . . . 42

5.2.2 Porteføljepræstation . . . 45

5.3 Robusthedsanalyse . . . 55

5.3.1 Porteføljepræstation høj-risiko versus lav-risiko måneder . . . 56

Kapitel 6 Diskussion **58**

6.1 Kritisk diskussion af litteraturen . . . 58

6.2 Machine Learning i finans . . . 60

6.2.1 Lav signal-til-støj ratio . . . 60

6.2.2 Inkorporering af økonomisk teori . . . 61

6.3 Mulige forbedringer . . . 62

Kapitel 7 Konklusion **64**

Litteratur **66**

Appendiks A Datastream **73**

Appendiks B Definition af variable **77**

Appendiks C Det skandinaviske aktiemarked **82**

Appendiks D Hypertuning **84**

Appendiks E Variables betydning **86**

Appendiks F Supplerende til: Empiriske resultater **89**

F.1 Robusthedstest . . . 89

Appendiks G Supplerende til: Algoritmer og metodologi **91**

G.1 OLS-antagelser . . . 91

G.2 Faktormodeller . . . 92

Kapitel 1

Indledning

Machine Learning (ML) har oplevet en bemærkelsesværdig stigning i popularitet i løbet af det seneste årti og har udvist hidtil uset succes på tværs af forskellige videnskabelige domæner. Da søgen efter faktorer til at forklare tværsnittet af aktieafkast har frembragt en faktor-zoo[1], synes ML at være ideelt egnet til at tæmme denne dimensionalitetsudfordring, når faktormodeller er anvendt til at prædiktere aktieafkast. Derfor ser det ud til, at ML har potentialet til ligeledes at etablere dominans inden for finansområdet. I en nylig akademisk undersøgelse har Gu et al. (2020) gennemført en bemærket komparativ analyse af ML-metoder til prædiktering af aktieafkast ved at syntetisere litteraturen om kapitalaktivers prisfastsættelse inden for ML-området. De giver bevis for en afvisning af det ordinære mindste kvadraters (OLS) benchmark til fordel for ML-metoder i form af den statistiske ydeevne og den økonomiske præstation.

Den traditionelle litteratur omkring prædiktering af aktieafkast anvender statistiske modeller til at estimere fremtidige afkast langs to dimensioner: tidsserie- (TS) og tværsnitsprædiktering (CS). TS-modeller anvender typisk tidsserieregression af aggregerede porteføljeafkast på en række makroøkonomiske variable, herunder værdiansættelsesforhold og renter, samt tekniske indikatorer (Cochrane 2011). I modsætning hertil forklarer CS-modeller typisk forskelle i forventede afkast på tværs af aktier som en funktion af karakteristika på aktieniveau, såsom størrelse, værdi og momentum (Fama & French 1993, Jegadeesh & Titman 1993, Lewellen 2015). Denne tilgang anvender typisk en Fama & MacBeth (1973) regression. Men som følge af underpræstationen af OLS-benchmarken i deres undersøgelse, anbefaler Gu et al. (2020) anvendelsen af ML-teknikker til at overvinde begrænsningerne ved almindeligt anvendte metoder. Siden da har der været en spirende interesse for at generalisere konklusionen til forskellige ML-metoder og aktieunivers.

[1] I løbet af de seneste årtier er faktorerne blevet så mange, at Cochrane (2011) omtalte samlingen som en zoo.

Flere studier bekræfter sammenhængene observeret af Gu et al. (2020) mens andre forholder sig mere kritiske. Avramov et al. (2022) konkluderer eksempelvis, at:

> "Deep learning signals extract profitability from difficult-to-arbitrage stocks and during high limits-to-arbitrage market states. In particular, excluding microcaps, distressed stocks, or episodes of high market volatility considerably attenuates profitability. Machine learning-based performance further deteriorates in the presence of reasonable trading costs due to high turnover and extreme positions."

Ud over den førnævnte forskning fremhæver en anden undersøgelse foretaget af De Nard et al. (2022) et ønske om at vende tilbage til den lineære metode ved at inkorporere en teknik kaldet **VA**riabel **S**ubsample **A**ggregation (VASA). Denne tilbøjelighed stammer til dels fra de betydelige datakrav, der er forbundet med ML. Utilstrækkelige data kan resultere i ustabilitet, hvilket igen kan begunstige vedtagelsen af den enklere metode frem for den mere komplekse. Ydermere har Li et al. (2022) vægt på et almindeligt overblik i litteraturen, nemlig negligeringen af investorperspektivet. I et forsøg på at løse dette, foreslår de at gøre strategierne både investerbare og fortolkelige. Fordelene påvist ved ML er hermed mere beskedne i forhold til litteraturen.

Denne afhandling har til formål at bygge videre på forskningen udført i litteraturen ved at afvige fra deres forskningsdesign på flere væsentlige aspekter med henblik på at undersøge generaliteten af konklusionen om at forkaste OLS-benchmarket til fordel for ML i tværsnitsprædiktering. Mere specifikt er fokus på: (1) det uudforskede skandinaviske marked[2], (2) at bruge et mere begrænset sæt af variable, der tillader en sammenligning i øjenhøjde, og (3) at ekskludere mikrocaps (beholde et større investorperspektiv). Alle er uddybet i afsnit 2.3: *Bidrag til litteraturen.*

Problemformulering

Hermed er afhandlingens problemformulering:

Hvordan sammenligner træbaserede machine-learing-algoritmer sig med den lineære OLS ved (tværsnits-) prædiktering af skandinaviske aktiers merafkast?

Med henblik på at opnå en fyldestgørende besvarelse af ovenstående problemformulering opstilles nedenstående hypoteser:

[2]Specifikt Danmark, Norge og Sverige.

Hypotese 1: Det er muligt med statistisk nøjagtighed at prædiktere aktiers merafkast.

Hypotese 2: Det er muligt at udnytte prædikteringerne til at sammensætte økonomisk rentable porteføljer, der kan generere (statistisk signifikant) alfa, det vil sige risikojusteret afkast.

Hypotese 3: ML-algoritmerne udkonkurrerer OLS statistisk og økonomisk (hypotese **1** og **2**). Med træbaserede ML-algoritmer henvises til Random Forest (RF) og Gradient Boosting (GB).

Accepten af hypotese **1** betegner algoritmernes evne til at prædiktere tværsnitsaspektet med en vis præcision, hvilket er det indledende skridt mod at kunne definere en handelsstrategi. Hypotese **2** fremsætter, at prædikteringerne giver økonomisk rentabilitet, hvilket ikke nødvendigvis er garanteret. Sidstnævnte opsummerer den komparativ analyse og statuerer, baseret på litteraturen, at ML-algoritmerne udkonkurrerer OLS, som er det endelige mål for afhandlingen.

Foruden opstillingen af ovenstående hypoteser har det ligeledes været nødvendigt at foretage en række afgrænsninger, derfor er forskellige valg og fravalg løbende diskuteret og argumenteret.

Fremgangsmåde

Resten af denne afhandling er struktureret som følger. Afsnit 2 gennemgår den eksisterede litteratur om kapitalaktivers prisfastsættelse med vægt på ML-modellers rolle inden for området. Afsnit 3 introducerer til de metodiske overvejelser, herunder følger en grundig databeskrivelse af, hvordan dataene er håndteret. Afsnit 4 introducerer til ML (og nøglebegreber heri), opsætter problemet og introducerer algoritmerne, der er anvendt i den komparative analyse. Endvidere indeholder afsnittet en introduktion til anvendte evalueringsmål, både statistiske og økonomiske. Afsnit 5 giver en detaljeret fortolkning af de empiriske resultater fra den komparative analyse med hensyn til prædiktering af fremtidigt merafkast (hypotese **1**). Afsnittet søger også at besvare hypotese **2**, der sammensætter porteføljer baseret på prædikteringerne, og kontrollerer for bredt accepterede risikofaktorer. Afsnit 6 vil diskutere resultaterne af den empiriske analyse og opsamle et kritisk perspektiv på den teoretiske rammesætning. Afsnit 7 vil sammenstille analyse og diskussion samt empiri og teori til at besvare afhandlingens problemformulering og foreslår relevante aspekter, som fremtidig forskning bør overveje at tage i betragtning.

Kapitel 2

Litteraturgennemgang

Prædiktering af aktieafkast er ikke et nyt område, men historisk har undersøgelserne fokuseret på lineære modeller (se f.eks. Haugen & Baker (1996), Lewellen (2015) og Drobetz et al. (2019))[1] ved at følge tilgangen introduceret af Fama & French (1993). Denne litteratur fokuserer ofte på identifikation af prissatte risikofaktorer og kvantificering af tilsvarende risikopræmier. Selvom der ikke er konsensus om det nøjagtige sæt af prissatte risikofaktorer, har forskere identificeret hundredvis af anomalier, der angiveligt forudsiger afkast på en måde, der ikke kan forklares med risikoeksponeringer for bredt accepterede risikofaktorer (se f.eks. Green et al. (2017)).

Denne afhandling bidrager til den spirende (og meget aktive) litteratur omkring kapitalaktivers prisfastsættelse (tværsnitsprædiktering) ved brug af ML-metoder, der forsøger at håndtere den multidimensionelle udfordring, som formuleret af Cochrane (2011) i hans AFA-præsidenttale[2]. De fleste artikler kan tildeles en af de følgende to strømme. Den første retning tager et sæt af karakteristikasorterede eller -projekterede porteføljer som givet og anvender dimensionalitetsreduktion med hensyn til at opnå en stokastisk diskonteringsfaktor (SDF) eller et mindre antal af faktorer, hvor forskellige versioner af latente faktormodeller er blandt de mest populære tilgange. Den anden del af litteraturen modellerer påvirkningen af karakteristika på afkast direkte uden at pålægge en underliggende risikomodel eller betingelse omkring arbitrage. Litteraturen inden for området er endog meget overlappende (se f.eks. Nagel (2021) for en mere omfattende litteraturgennemgang). Litteraturen er i det følgende præsenteret med et hovedfokus på sidstnævnte

[1] De præsenterer omfattende beviser for, at inkorporeringen af flere karakteristika giver en stærk prædikteringsevne for aktieafkast. De udfordrer også hypotesen om det efficiente marked ved at konkludere, at aktier med højere forventet (og realiseret) afkast ikke er mere risikable end deres modparter med lavere afkast. Til gengæld viser de, at prædikteringer af aktieafkast kan bruges til at implementere rentable investeringsstrategier.

[2] Cochrane (2011) efterlyste (specifikt) metoder ud over tværsnitsregression og -porteføljesortering.

strøm, som er tættest denne afhandling, men indleder med at introducere førstnævnte. Dernæst præsenteres væsentlige problematikker/diskussioner vedrørende anvendelsen af ML i finans.

2.1 Aktivprissætning med Machine Learning

2.1.1 Den første strøm

En række artikler forsøger ved hjælp af ML at forbedre modellerne for kapitalaktivers prisfastsættelse ved at identificere relevante faktorer til at forklare tværsnittet af aktieafkast. Feng et al. (2018) udvikler en automatiseret faktorsøgningsalgoritme, der søger over virksomhedskarakteristika med formålet at minimere prissætningsfejlene. Kelly et al. (2019) anvender Instrumented Principal Components Analysis (IPCA) til at udføre dimensionalitetsreduktion af det faste karakteristikarum. Denne metode udvider den projicerede PCA ved at lade aktivernes belastning af latente faktorer afhænge af en vektor af karakteristika. Hvis karakteristika/forventet afkastforhold er drevet af kompensation for eksponering mod latente risikofaktorer, vil IPCA identificere de tilsvarende latente faktorer. Hvis der ikke findes sådanne faktorer, udleder IPCA, at den karakteristiske effekt er kompensation uden risiko, og allokerer den til en anomali-skæring. Blandt en lang række karakteristika, der er udforsket i litteraturen, er kun ti statistisk signifikante på 1 pct.-niveauet i IPCA-specifikationen og er ansvarlige for næsten 100 pct. af modellens nøjagtighed. Feng et al. (2020) og Freyberger et al. (2020) fokuserer på LASSO-estimering og finder tegn på væsentlig redundans blandt tværsnitsprædiktorerne for aktieafkast. Hvor sidstnævnte finder, at kun 13 ud af 62 karakteristika giver øget information til tværsnittet af forventet afkast. Gu et al. (2021) udvider IPCA-tilgangen i Kelly et al. (2019) ved hjælp af et autoencoder-netværk, der tillader faktorbelastningerne at afhænge ikke-lineært af karakteristika. Dette fører til mindre prissætningsfejl sammenlignet med førende faktormodeller.

2.1.2 Den anden strøm

Kun en mindre andel af litteraturen anvender ML til aktieprædiktering og -udvælgelse med henblik på at opstille en handelsstrategi. Den tidlige litteratur analyserer ofte kun én ML-metode. Et første eksempel på brugen af neurale netværk (NN) findes i Messmer (2017). Heri præsenteres et dybt Feedforward NN (FNN) til at prædiktere det (månedlige) amerikanske tværsnit af aktieafkast. Dette FNN er trænet på en række selskabsspecifikke variable og genererer attraktive risikojusterede afkast sammenlignet med den lineære benchmark, hvilket bekræfter vigtigheden af ikke-linearitet i modeller for kapitalaktivers prisfastsættelse. Messmer er endog forsigtig i sin konklusion om, at det (dybe) NN er den optimale metode til at inkludere denne ikke-linearitet.

Moritz & Zimmermann (2016) foreslår træbaserede porteføljesorteringer for at omgå begrænsningerne ved den lineære model og for at redegøre for interaktionen mellem mange forklarende variable. De prædikterer selskabers aktieafkast ved hjælp af alle en-måneds afkast over en toårig årrække før porteføljedannelsen. En prædikteringsbaseret lang-kort portefølje giver overnormale afkast i forhold til Fama & French (1993)'s model plus momentum.

Gu et al. (2020) er de første til omfattende anvendelse og sammenligning af ML-metoder i tværsnitsprædiktering (i forsøget på at nedbryde kløften mellem de to discipliner) uden stillingtagen til risiko-afkast forhold. Ikke desto mindre er denne artikel blevet den primære benchmark for senere artikler, der anvender ML, både for den anvendte praksis (database, stikprøveopdeling, etc.) og opnåede resultater (R^2, variabel betydning, etc.). Metoderne, de sammenligner, er ordinære lineære modeller, penaliserede lineære modeller, lineære modeller med dimensions reduktion via principal komponenter (PCR) og partielle mindste kvadraters (PLS), regressionstræer (RF samt GB) og forskellige FNN'er. De inkluderer 94 selskabsspecifikke karakteristika og interaktionen af hvert karakteristika med otte makroøkonomiske tidsserier fra Welch & Goyal (2008) samt 74 industri (dummy) variable, det vil sige sammenlagt +900 variable. De prædikterer næste måneds merafkast for alle aktier på de store amerikanske børser over en 30-årig periode (1987-2016). Konklusionen er, at ML som helhed har potentialet til at forbedre og forenkle den empiriske forståelse af aktivafkast. For det anvendte data klarer FNN sig bedst, tæt forfulgt af RF og GB. ML-metoderne heri er generelt enige med hensyn til de mest indflydelsesrige variable, som hører til de tre kategorier: (1) momentum, (2) likviditet og (3) volatilitet. På makroniveau er den vigtigste variabel den aggregerede book-to-market ratio. Den prædikteringsbaserede handelsstrategi, der genererer den højeste Sharpe-Ratio, er et ensemble af alle anvendte metoder. Den umiddelbare implikation er endog, at ML kan hjælpe med at løse praktiske investeringsproblemer, såsom markedstiming, porteføljevalg og risikostyring. Det skyldes, at ML-porteføljerne genererer Sharpe-Ratioer, der er mere end en fordobling af præstationen fra førende regressionsbaserede strategier i litteraturen.

Efter denne skelsættende undersøgelse er der opstået en række artikler, der bidrager til litteraturen ved at undersøge almenheden af konklusionen. Dette er opnået ved anvendelse af samme forskningsdesign på andre aktiemarkeder (se f.eks. Tobek & Hronec (2021), Drobetz & Otto (2021) samt Rubesam (2022) for international anvendelse og applikation til individuelle regioner eller lande) eller andre aktivklasser (se f.eks. Bianchi et al. (2020) for applikation til obligationsafkast). Eksempelvis anvender Tobek & Hronec (2021) hundrede af aktieanomalier, registreret i den videnskabelige litteratur, til at demonstrere, at ML-metoder, der kombinerer alle anomalier

i ét enkelt fejlprissignal, er værdifulde over hele verden og kan forblive i et likvidt aktieunivers. Drobetz & Otto (2021) finder desuden, at en klassifikationsbaseret porteføljedannelse, der ikke estimerer det forventede afkast på aktieniveau, præsterer endnu bedre end den bedste ML-model. En anden del af litteraturen (se f.eks. Leippold et al. (2022) og Chen et al. (2023)) er dedikeret til at forbedre algoritmerne, der er undersøgt i Gu et al. (2020).

Avramov et al. (2022) viser endog, at afkastprædikteringer fra Deep-Learning (DL)-metoder, i USA, primært trækker deres profitabilitet fra aktier og markedsstadier, der er svære at arbitrere fra. De argumenterer ligeledes for, at prædikteringerne forstyrres yderligere, hvis mikrocap selskaber ekskluderes, og hvis rimelige transaktionsomkostninger er overvejet[3]. Leung et al. (2021) finder, at de økonomiske gevinster af GB, for udviklede aktiemarkeder, tenderer til at være mere begrænset og kritisk afhængig af evnen til at påtage risiko samt implementere handlerne efficient. Derudover dokumenterer Cakici et al. (2022), at ML-strategier fungerer bedst for mindre aktier, såvel som i lande med mange listede selskaber og høj idiosynkratisk risiko.

2.2 Fortolkningen af Machine Learning i finanslitteraturen

Med anvendelsen af ML-metoder i stort set alle anvendte videnskaber i de senere år, er det blevet afgørende at kunne fortolke resultater (og handelssignaler) fra sådanne modeller. Derfor er ML-metoder blevet analyseret særligt med henblik på dette. For eksempel præsenterer Lopez de Prado (2018) mange funktionsvigtige metoder, mens Li et al. (2022) opdeler det prædikterede afkast i tre komponenter; et lineært, et ikke-lineært og et interaktionskomponent gennem en ny metode kaldet Fingerprint. De foreslår desuden tre principper: Investability, Interpretability og Interesting til evaluering af den praktiske effektivitet af ML til aktieudvælgelse. Daul et al. (2022) dekomponerer ligeledes det prædikterede afkast i de førnævnte tre komponenter. De finder, at regressionstræer og NN'er er overlegne grundet deres stærke regulariseringsmekanisme og kapacitet til at fange interaktionseffekter. Det ikke-lineære komponent har på den anden side ingen prædikteringsevne. Givet denne metodologi formår de at isolere og studere interaktionskomponenten samt finde, at denne har en signifikant (langsigtet) prædikteringsevne uafhængigt af den lineære modellering og har været stabil gennem tiden. En række artikler forsøger ligeledes at fortolke ML-metoderne gennem såkaldte Shap Values, som indikerer vigtigheden af hver variabel (se f.eks. Hanauer et al. (2022)). Ofte udelader artiklerne den økonomiske motivation, fortolkning samt stabilitet i handelsstrategierne eller diskussionen vedrørende robustheden mod både

[3]En lovende retning er præsenteret i Bryzgalova et al. (2020) samt Cong et al. (2021), som estimerer optimale ML-porteføljer underlagt handelsbegrænsninger.

økonomiske og handelsrelaterede barrierer, der kan begrænse den praktiske implementerbarhed på de finansielle markeder.

2.3 Bidrag til litteraturen

Bidraget fra denne afhandling er tredelt. Første bidrag er til den spirende og hurtigt voksende litteratur om kapitalaktivers prisfastsættelse gennem anvendelsen af ML til afkastprædiktering. Antallet af artikler er accelereret de seneste to år og må forventes at fortsætte som følge af de lovende resultater, men der er imidlertid brug for både mere viden og større konsensus. Dernæst tilføjer afhandlingen til litteraturen ved at undersøge det uudforskede skandinaviske marked[4], specifikt Danmark, Norge og Sverige. Ved at fokusere på det skandinaviske marked adresserer afhandlingen adskillige tilbageværende spørgsmål vedrørende effektiviteten og anvendeligheden af ML-algoritmer i denne særlige markedssammenhæng, og dermed bygges bro over det eksisterende hul i litteraturen. Slutteligt fokuserer afhandlingen på et mere likvidt aktieunivers og vil i overensstemmelse med Hou et al. (2020) ekskludere mikrocaps. Dertil er litteraturen typisk afhængig af et meget stort sæt af variable. Eksempelvis er antallet af variable i Gu et al. (2020) +900, mens Leippold et al. (2022) har udvidet dette sæt til 1160 variable, hvor Tobek & Hronec (2021) har +300 variable med. Det er muligt at argumentere for, at ML kan håndtere irrelevante variable, mens OLS ikke kan, hvilket kan give ML-metoderne en fordel. For at opnå en sammenligning (mere) i øjenhøjde, er OLS-benchmarket og ML-metoderne givet det samme forholdsvis sparsomme sæt af variable. Disse to aspekter bringer et investorperspektiv og dermed større tillid til, at modellerne kan implementeres i praksis. Det ultimative mål er at bygge (rentable) porteføljer baseret på afkastprædikteringerne. Metodologisk er denne afhandling tættest Gu et al. (2020), der sammenligner den lineære OLS med modeller baseret på ML-algoritmer.

[4]Et potentielt problem i finanslitteraturen er det såkaldte U.S. bias (og dermed mangel på ekstern validitet). Andrew Karolyi (2016) finder, at kun 16% (23%) af alle artikler offentliggjort i top fire (fjorten) finanstidsskrifter undersøger andre markeder end det amerikanske, det vil sige en brøkdel, der er langt under målene, der afspejler deres økonomiske betydning, hvilket af flere grunde er yderst problematisk.

Kapitel 3

Metoder og databeskrivelse

3.1 Videnskabsteoretisk og metodisk overvejelse

Afhandlingen vil anskue verden ud fra et kritisk realistisk perspektiv, idet den anerkender, at det samfundsmæssige genstandsfelt vurderes at eksistere uafhængigt af forskeren, og at samfundsmæssige relationer vil ændre sig (Fuglsang & Bitsch-Olsen 2013). Den anerkender begrænsningerne af menneskelig opfattelse i fuldt ud at forstå virkeligheden, som den opfattes såvel kompleks som kontekstuel, da ændringer i afkast ofte undgår fuldstændig forklaring alene gennem rationelle midler. Følelser og intellektuelle mekanismer spiller en væsentlig rolle i udformningen af investorernes beslutninger. Ved at omfavne denne ramme har afhandlingen til formål at udvikle en nuanceret og omfattende model, der fanger underliggende strukturer og mekanismer ud over observerbare markedsdata.

Afhandling er udarbejdet på en kombination af såvel den deduktive som den induktive tilgang. Det skyldes, at der gennem processen er en kontinuerlig iteration mellem den teoretiske ramme og den empiriske observation, som sikrer, at afhandlingen vil gå fra en beskrivende til en mere dybdegående forståelse af problemstillingen (Darmer et al. 2010). Målet er objektivt at vurdere kompatibiliteten mellem teori og virkelighed gennem streng empirisk testning, og derved udvide gyldighedens omfang for eksisterende teori. Den deduktive tilgang er indledningsvist anvendt, da afhandlingen har sit udgangspunkt i eksisterende teori, som er undersøgt for det skandinaviske aktiemarked. Hermed er resultaterne udledt fra en teoretisk forståelse ved at teste denne, hvor hypotesen om, hvorvidt ML-modeller sammenlignet med den lineære OLS kan generere statistisk og økonomisk relevans, er forsøgt be- eller afkræftet. Det betyder, at afhandlingen anvender en komparativ analyse med relevante anomalier, der (teoretisk og empirisk) kan forklare tværsnit-

tet, for at bestemme det fremtidige merafkast. Antagelsen er, at de forskellige modeller danner grobund for afkastprædiktering, og at de kan omsættes til rentabilitet på aktiemarkedet. Hertil indsamles empiri og teori samt kvantitativt og kvalitativt data, som analyseres og diskuteres. Den induktive tilgang hjælper så med at generalisere mod kausaliteter baseret på empiriske observationer, og derved bygge bro mellem teori og virkelighed. Den induktive inspiration afspejler de observerede mønstre og potentielle mangler i den teoretiske og empiriske ramme, hvilket fører til en dybere forståelse af underliggende mekanismer og mulige kausaliteter.

3.2 Empiri

Såvel kvantitativt som kvalitativt data er anvendt for at kunne opnå en nuanceret konklusion, der sikrer en kvalificeret og grundig analyse. En forudsætning er, at anvendelsen og fortolkningen sker hensigtsmæssigt med fokus på at sikre objektivitet. Det kvalitative data er anvendt til at underbygge eller modsige argumentationen men også til at analysere og fortolke den empiriske analyse ud fra. Desuden er særligt diskussionen centreret omkring kvalitativt data for at kunne nuancere og frembringe et kritisk perspektiv. Det kvantitative data er udarbejdet med fokus på validitet og realibilitet, som vil sikre afhandlingens repræsentativitet (og troværdighed)[1].

Samlet set vurderes afhandlingen at have et højt validitetsniveau, primært grundet brugen af variable, der er forankret i tidligere forskning, som har vist deres effektivitet i tværsnitsprædiktering. Dermed er variablenes relevans tæt knyttet til anvendeligheden på den aktuelle problemstilling. Endvidere forventes brugen af forskellige evalueringsmål at styrke resultaterne. Slutteligt anvender afhandlingen triangulering som et middel til at øge validiteten af dens resultater og til at kompensere for potentielle begrænsninger ved de valgte metoder. Dette opnås ved at sammenstille afhandlingens resultater med eksisterende litteratur, herunder analyser af det samme fænomen. Dermed implementeres både kvalitative og kvantitative metoder for at gennemføre undersøgelsen. Ved at anvende triangulering valideres og underbygges resultaterne, hvilket øger afhandlingens generaliserbarhed samtidig med, at det giver større dybde og flere perspektiver.

Den overordnede reliabilitet af afhandlingen er vurderet moderat høj, hvilket bl.a. fremkommer ved brugen af en datakilde, der er yderst anvendt i litteraturen, samt indsatsen for at løse potentielle mangler, som beskrevet i afsnit 3.2.2. Et aspekt, der endog i moderat grad kan påvirke afhandlingens reliabilitet, vedrører den relative mængde af manglende data, da tilstedeværelsen heraf kan give anledning til bekymring med hensyn til fuldstændigheden af de analyserede data.

[1] Selve databeskrivelsen er gennemgået i afsnit 3.2.2.

3.2.1 Softwareværktøjer

Al kodning (og dataklargøring) i denne afhandling er fortaget i Python. Python har et stort udvalg af bibliotekter, og anvendelsen af disse har (betydeligt) reduceret både implementeringstiden og risici. Det mest bemærkelsesværdige er dog Sci-kit Learn-biblioteket, som har tilladt at importere de træbaseredeklassifikatorer. For dataklargøring er Numpy og Pandas anvendt.

3.2.2 Databeskrivelse

Det anvendte datasæt omfatter det skandinaviske aktiemarked (Danmark, Norge og Sverige) i tidsperioden 1994(1)-2022(12)[2]. Regnskabsdataene er fra Refinitiv Worldscope, og aktiemarkedsdataene er fra Refinitiv Datastream. En anden vigtig kilde er Institutional Brokers' Estimate System (I/B/E/S), som ligeledes er en database i Datastream (leveret af Thomson Reuters).

Datastream er en omfattende finansiel og makroøkonomisk database, men på trods af dens omfattende dække, er det også klart, at ukritisk brug af de downloadede nøgletal er problematisk. Som påpeget af Ince & Porter (2006), er Datastream-data ofte for rå til at indgå direkte i analysen. For at sikre, at analysens resultat ikke er præget af mangelfulde data (øge reliabiliteten af afhandlingen), er der foretaget omhyggelig screening og korrektion med inspiration fra bl.a. Ince & Porter (2006) og Schmidt et al. (2011), som har søgt at få fejlfrie data fra Datastream.

Ovenstående betyder, at adskillige statiske men også dynamiske skærme er anvendt for at sikre, at datasættet kun består af almindelige aktier og giver den højeste datakvalitet. Først er aktierne identificeret ved hjælp af aktielister i Refinitiv Datastream for såvel aktive som inaktive. Datasættet er dermed fri for overlevelsesbias og inkluderer alle selskaber, der var eller er børsnoteret i ét af de tre markeder i 2022(12). Efter Ince & Porter (2006) elimineres ikke-almindelige aktier gennem en række statiske skærme. Desuden er flere dynamiske skærme anvendt til aktieafkast og -kurser for at udelukke fejlagtige og illikvide observationer. Bilag A giver en mere detaljeret beskrivelse af de anvendte skærme. Denne afhandling følger endvidere Fama & French (2017) og tildeler aktierne i tre størrelsesgrupper (mikro, lille og stor) for de enkelte markeder og måneder. Store aktier er defineret som de største aktier, der tilsammen tegner sig for 90 pct. af markedets samlede markedsværdi. Små aktier er defineret som de aktier, der udgør de næste 7 pct. af den samlede markedsværdi. Mikroaktier omfatter de resterende 3 pct. Selvom mikroaktier kun repræsenterer 3 pct. af den samlede markedsværdi, udgør de omkring 50 pct. af antallet af aktier (i overensstemmelse med litteraturen)[3]. For at undgå, at resultaterne er drevet af mikroaktier,

[2]Dermed fokuserer afhandlingen på en kortere og nyere tidsperiode sammenlignet med litteraturen.
[3]Denne misbalance udgør specifikke vanskeligheder, fordi det generelle ML-mantra om, at mere data er bedre

følger denne afhandling Hou et al. (2020) ved at ekskludere dem (for også at følge afhandlingens fokus på implementerbare resultater ved at undgå illikviditet og lav omsætning). Resultatet er et datasæt, der spænder over 1.080 unikke aktier og mere end 89.000 månedsobservationer, hvilket er betydeligt mindre end de anvendte datasæt i litteraturen. Bilag C har en kort beskrivelse af markedet men også en sammenfattende statistik for den endelige stikprøve af selskaber.

Selskabsspecifikke karakteristika

Dataene i datasættet er indsamlet med varierende tidsfrekvens (nogle er på månedsbasis mens andre er årlige), og hvis de er valutarelateret, er de noteret i euro. Datasættet er bestående af 42 selskabsspecifikke variable, som er almindelig anvendt i litteraturen om kapitalaktivers prisfastsættelse og har vist sig at forklare det meste af tværsnitsvariationen i forventede afkast og er taget fra Kelly et al. (2019) og Windmuller (2022). Bilag B skitserer den detaljerede konstruktion af variablene og deres tidsfrekvens. Det er antaget, at markedsdataene er offentlige med det samme, mens regnskabsdataene er kendte seks måneder efter regnskabsårets afslutning[4], hvilket også er en anvendt tilgang i litteraturen. Det vil sige, at de årligt opdaterede variable, der inkorporerer data fra regnskabsåret, der slutter i kalenderåret $t-1$, er anvendt fra slutningen af juni i året t til og med maj i året $t+1$ for at undgå et look-ahead bias.

De selskabsspecifikke variable er: assets-to-market (A2ME), total assets (AT), sales-to-assets (ATO), book-to-market (BEME/BEME$_m$), market beta (Beta), cash-and-short-term-investment-to-assets (C), cash-based operating profits-to-asset (CbOPtA), capital turnover (CTO), capital intensity (D2A), leverage (Debt2P), ratio of change in property, plants, and equipment to change in total assets (DPI2A), earnings-to-price (E2P), fixed costs-to-sales (FC2Y), cash flow-to-book (FreeCF), idiosyncratic volatility (Idiovol), investment (INV), market capitalization (LME), turnover (LTurnover), net operating assets (NOA), operating accruals (OA), operating leverage (OL), price relative to its 52-week high (P2P52WH), price-to-cost margin (PCM), profit margin (PM), gross profitability (Prof), Tobin's Q (Q), momentum (r_{12-2}), intermediate momentum (r_{12-7}), short-term reversal (r_{2-1}), long-term reversal (r_{36-13}), return on net operating assets (RNA), return on assets (ROA), return on equity (ROE), sales-to-price (S2P), the ratio of sales and general administrative costs to sales (SGA2S), unexplained volume (SUV), og illiquidity (Illiqu). Dog er CbOPtA, FC2Y og DPI2A slettet grundet manglende data. Finansielle selskaber er ikke ekskluderet, men en række variable er sat til at være nul, fordi de ikke er defineret meningsfuldt for disse.

end mindre data, også er gældende i kapitalaktivers prisfastsættelse. Men på den anden side møder denne litteratur også ofte kritik fra praktikere, da mange resultater er stærkt afhængige af mikroaktier, og som mikroaktier generelt er mindre likvide og dyrere at handle, bliver resultaterne mindre implementerbare.

[4]Dette er endog meget konservativt men sikrer, at oplysningerne faktisk er i det offentlige domæne.

Dette vedrører: ATO, C, D2A, FreeCF, CF2P, GP2A, OA, PCM, PM, Prof, RNA, SGA2S, og NOA som i Hanauer & Kalsbach (2023)[5]. Foruden ovenstående variable har denne afhandling også inddraget 6 sentimentbaserede variable som analytikers gennemsnitlige køb/salg/hold-anbefaling, medianen af analytikers forecast af vækst i implicit indtjening per aktie (EPS) for FY1, FY2 og FY3 samt DY (Dividend Yield).

Afkastet er beregnet fra det samlede afkastindeks fra Refinitiv Datastream. Det skyldes, at dette er et mere retvisende billede af det faktiske afkast, som det ikke kun inkorporerer ændringerne i prisen, men også virkningerne af udbytte, aktieopdelinger og andre virksomhedshandlinger, der kan påvirke det samlede afkast. Til beregning af merafkastet anvendes tre-måneders EURIBOR-renten, skaleret til én måneds horisont, som den risikofrie rente, ligesom i Drobetz & Otto (2021). Det (gennemsnitlige) månedlige merafkast er 1,02 pct. med en standardafvigelse på 11,49 pct. og tenderer til den høje side, formodentligt grundet det anvendte afkastindeks. Det (gennemsnitlige) månedlige merafkast af en lang portefølje, baseret på alle aktive aktier inden for de givne måneder, er 1,35 pct. for en ligevægtet portefølje og 0,88 pct. for en værdivægtet portefølje, jf. tabel 3.1. Førstnævnte svarer til et annualiseret merafkast på 17,51 pct.

Tabel 3.1 Sammenfattende statistikker efter markedsportefølje

	Mean	Std	Min	25%	50%	75%	Max
Ligevægtet [%]	1,35	5,05	-18,44	-1,05	1,62	4,60	20,99
Værdivægtet [%]	0,88	4,62	-19,18	-1,17	1,30	3,95	15,82

Den empiriske litteratur om kapitalaktivers prisfastsættelse foreslår to tilgange til behandlingen af manglende værdier (NaN). En tilgang er at begrænse dataene til de aktier, der har fuldstændig information for de specifikke måneder, hvilket vil efterlade denne afhandling med omkring 50 pct. af de samlede observationer. Alternativt imputeres dataene ved hjælp variablenes månedlige tværsnitsmedian. Det skal bemærkes, at sidstnævnte metode introducerer kunstige tidsserieudsving i variablene og supplerer dataene med forkert information. Denne afhandling erkender dog, at fordelene ved at inkludere mere data opvejer den (relativt) lille stigning i fejl, da det muliggør en mere omfattende analyse. Forøgelse af datastørrelsen forbedrer ofte præcisionen og generaliserbarheden af statistiske modeller. Desuden er imputering med medianen i vid udstrækning anvendt i lignende undersøgelser. Denne tilgang er især velegnet, når dataene ser ud til at være **Missing at Random** (MAR), hvilket betyder, at sandsynligheden for, at en værdi mangler,

[5] Gu et al. (2020) inkluderer ligeledes finansielle selskaber i deres empiriske analyse.

ikke er relateret til den faktiske værdi af den manglende variabel men kun af andre observerbare variable. I sådanne tilfælde kan imputering med medianen hjælpe med at bevare dataenes statistiske egenskaber, såsom middelværdi og standardafvigelse, og forhindre biased estimater af modelkoefficienterne. I overensstemmelse med litteraturen, herunder henvisning til Gu et al. (2020), tilpasser imputering med medianen afhandlingens metodologi med etableret praksis og letter sammenligneligheden. Dernæst rangeres alle variable, efter Gu et al. (2020), måned for måned for hvert marked i $[-1, 1]$-intervallet for at begrænse effekten af outliers. Denne markedsbaserede gruppering har ligeledes til hensigt at kontrollere for forskellige regnskabsstandarder på tværs af markederne og tager derfor højde for forskelle i variablene.

Opsummering

Ovennævnte screening og korrektion er udført for at undgå forskellige biases, der påvirker resultaterne og dermed robustheden af de dragede konklusioner. På den anden side bør for meget manipulation med data ligeledes undgås, da dette også kan introducere uønskede biases. Sammenfattende er alle screeninger og korrektioner også at finde i relateret litteratur.

Kapitel 4

Algoritmer og metodologi

Kapitlet starter med en kort introduktion til ML og problemerne med over- og undertilpasning. Dernæst fortages problemopsætningen, hvor de anvendte algoritmer (OLS, RF samt GB) præsenteres, og den enkeltes faldgrubber/begrænsninger er belyst. Slutteligt præsenteres de anvendte evalueringsmatricer, både statistiske og økonomiske, der skal sammenfatte præstationen.

4.1 Machine Learning

ML er en gren af kunstig intelligens[1], der anvender algoritmer til at analysere (samt lære) mønstre i høj-dimensionelt og komplekst data. Dermed automatiseres bearbejdningen af komplekst data mere effektivt samtidig med, at en bestemt opgave er udført:

> "The field of machine learning is concerned with the question of how to construct computer programs that automatically improve with experience. [...] A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E" (Mitchell 1997).

Generelt set betyder ML, at et computerprograms ydeevne forbedres med erfaring med hensyn til en opgave. Som sådan sigter det mod at automatisere opgaven med analytisk modelbygning. Dette opnås gennem algoritmer, der iterativt lærer af problemspecifikke træningsdata, som gør

[1] Bredt defineret omfatter kunstig intelligens enhver teknik, der gør det muligt for computere at efterligne menneskelig adfærd og reproducere (eller udmærke sig over) menneskelig beslutningstagning for at løse komplekse opgaver uafhængigt af (eller med minimal) menneskelig indgriben (Nilsson 1996).

det muligt for computere at finde skjulte indsigter og komplekse mønstre uden eksplicit at blive programmeret. Det er dog ikke altid passende at bruge en generel definition af ML. Beskrivelsen bør skræddersyes til det specifikke område, som teknikkerne er anvendt i. Derfor har Gu et al. (2020) skitseret en kontekstspecifik definition af ML i empirisk aktivprissætning:

> "[...] (a) a diverse collection of high-dimensional models for statistical prediction, combined with (b) so-called "regularization" methods for model selection and mitigation of overfit, and (c) efficient algorithms for searching among a vast number of potential model specifications" (Gu et al. 2020).

De tre elementer i definitionen af Gu et al. (2020) fanger essensen af ML. ML'ens højdimensionelle karakter (1) giver stor fleksibilitet med hensyn til at tilnærme de underliggende afkast. Ifølge Mullainathan & Spiess (2017) går denne fleksibilitet dog i hånd med en større risiko for overtilpasning. Derfor skal ML-tilgange inkludere justeringer i dens implementering, der beskytter mod denne overtilpasning (2), som også nævnt af Cawley & Talbot (2010). Gu et al. (2020) og Hwang (2018) understreger yderligere vigtigheden af det sidste element (3), da ML-teknikker bør designes til at tilnærme en optimal specifikation med overskuelige beregningsomkostninger.

Som videnskabelig disciplin indeholder ML mange teknikker, men det er muligt at inddele indlæringsproblemet i tre kategorier: (1) overvåget, (2) ikke-overvåget og (3) reinforcement læring. Denne afhandling vil fokusere på førstnævnte, og i modsætning til ikke-overvåget, er det rigtige svar indledningsvist givet til den lærende, som den kan lære af.

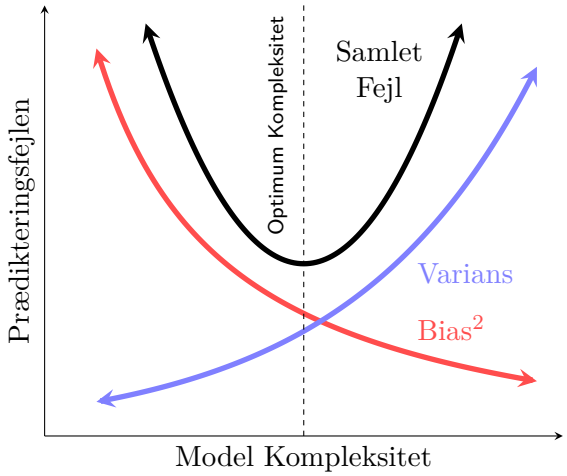
Bias-Varians afvejningen

En vigtig overvejelse ved valget af statistisk model er bias-varians afvejningen. Bias er defineret som den sande fejl i den anvendte statistiske model. En model siges at have et højt bias, hvis den er for enkel og ikke tilpasser sig til dataene (det vil sige, at den ikke finder relevante mønstre), hvilket fører til en undertilpasning. Varians er på den anden side defineret som variabiliteten af en model for et givent datapunkt. Høj varians fører til en overtilpasning, hvilket betyder, at modellen er for tæt tilpasset træningsdataene og dermed ikke generaliserer godt til nye, usete data. Dette fører til en stor forskel mellem fejl på de data, der bruges til at træne modellen, kontra de data, der bruges til at teste modellen. Det betyder at, hvor den simple model kan være biased grundet fejlagtige antagelser om den underliggende datagenereringsproces, kan den komplekse model have høj varians i fejlene grundet en modellering af støjen. Prædikterings-/generaliseringsfejlen for en regressionstilpasning kan udtrykkes med bias- og variansstermer:

$$E[(y^* - f^*(x))^2] = \text{støj}^2 + \text{varians} + \text{bias}^2 \tag{4.1}$$

, hvor den første term er den ikke-reducerebare fejl, som er fejlen, der er forbundet med dataenes naturlige variabilitet, så er de to sidste termer de reducerbare fejl, der kan øge nøjagtigheden af en model (Hastie et al. 2009). I en ideel verden bør modeller udvise både lavt bias og lav varians. Hvis en model har højt bias, vil både trænings- og testfejlene være høje, men hvis en model har høj varians, vil træningsfejlene være lave, mens testfejlene vil være høje. I forsøget på at gøre modellen mere kompliceret for at reducere dens bias, øges variansen, da modellen vil overtilpasse til træningsdataene. Dette fører til bias-varians afvejningen i figur 4.1, hvor den stiplede linje er den optimale bias-varians afvejning. Hver algoritme har sin egen måde at håndtere dette på.

Figur 4.1 Illustration af Bias-Varians afvejningen



4.2 Problemopsætning

Som i Gu et al. (2020) anvendes en additiv prædikteringsmodel til at beskrive det månedlige merafkast[2] for en aktie, $r_{i,t+1}$, hvilket kan skrives som:

$$r_{i,t+1} = E_t\left[r_{i,t+1} \mid x_{i,t}\right] + \varepsilon_{i,t+1} \tag{4.2}$$

, hvor $E_t\left[r_{i,t+1} \mid x_{i,t}\right]$ er det betingede forventede merafkast for aktie i over måned $t+1$ ved tidspunkt t givet vektoren af selskabsspecifikke karakteristika kendt ved t, $x_{i,t} \in \mathbb{R}^p$, mens $\varepsilon_{i,t+1}$ er prædikteringsfejlen. Objektivet er således at modellere det forventede merafkast ved den ukendte

[2]Defineret som aktiens afkast fratrukket den risikofrie markedsrente, tre-måneders EURIBOR-rente skaleret til én måneds horisont.

funktion f^*, $f^* : \mathbb{R}^p \to \mathbb{R}$, som estimerer det forventede afkast uafhængigt af anden information end vektoren af p selskabsspecifikke karakteristika, der er tilgængelige i t:

$$E_t\left[r_{i,t+1} \mid x_{i,t}\right] = f^*(x_{i,t}) \tag{4.3}$$

I tilfælde af overvåget ML bliver den ukendte funktion $f^*(x)$ tilnærmet af en eller anden funktion $f(x, \theta, \rho)$, som parametriseres af en koefficientvektor θ og et sæt af hyperparametre ρ. Mens θ er direkte afledt af de underliggende træningsdata med hensyn til ρ og en specifik tabsfunktion L, afhænger ρ selv af brugerinputtet, men er optimeret med hensyn til L baseret på tilgængelige data. Den nøjagtige funktionelle form af f er afhængig af den anvendte algoritme og kan enten være lineær eller ikke-lineær, såvel som parametrisk eller ikke-parametrisk. ML forsøger typisk at undgå stærke antagelser om $f^*(x)$, som f.eks. at antage, at f er lineær, og at x kun indeholder et sæt variable udvalgt a priori fra et større sæt af potentielle variable. I stedet er målet at lade dataene tale om $f^*(x)$ under kun svage antagelser om f og sættet af relevante variable[3].

Wolpert & Macready (1997) viser endog, at en universel læringsalgoritme ikke eksisterer (no-free-lunch teoremet i ML). Det betyder, at medmindre forskeren har en vis forudgående viden om prædikteringsproblemet, er der ikke grund til at foretrække, blot baseret på de første principper, én indlæringsalgoritme frem for en anden. Desuden, hvis forskeren har flere algoritmer, der passer lige godt til træningsdataene, kan denne ikke afgøre, uden forudgående viden, hvilke af disse der sandsynligvis vil give bedre prædikteringer på et endnu ikke set testdatasæt. At finde en algoritme, der producerer gode prædikteringer på testdataene, kræver derfor nødvendigvis (domænespecifik) forudgående viden om det prædikteringsproblem, som forsøges løst. Det kan være viden om den klasse af funktioner, som f tilhører, stabiliteten af denne funktion fra trænings- til testdata, og den statistiske fordeling af x og ε. Hermed kræver succesfuld prædiktering, at forskeren fordrejer valget af indlæringsalgoritme, herunder deltaljerne i dens implementering, mod algoritmer, der er passende for arten af det specifikke problem (Wilson & Martinez 1997).

På trods af, at ML forsøger at undgå stærke antagelser om $f^*(x)$, er algoritmerne baseret på en model, der stammer fra et sæt af antagelser. Dette induktive bias resulterer i fejl, hvis antagelserne ikke holder for fremtidige data. Dette øger motivationen for at udforske modeller, der

[3]Ifølge Gu et al. (2020) pålægger den anvendte ramme på trods af sin fleksibilitet også nogle vigtige begrænsninger på f^*. Funktionen f^* er ikke afhængig af hverken (i) eller (t). Ved at opretholde den samme form over i og t (en panel-datamodel), udnytter modellen information fra hele panelet, hvilket giver stabilitet til estimaterne for ethvert i. Dette står i kontrast til standardmetoderne ved kapitalaktivers prisfastsættelse, der re-estimerer en tværsnitsmodel for ethvert t, eller som uafhængigt estimerer en tidsseriemodel for ethvert i.

består af flere (basis) estimatorer[4], også kaldet ensembler. Teorien foreslår, at en sådan tilgang vil øge nøjagtigheden og har visse fordele, når den konfronteres med statistiske, beregnings- og repræsentationsmæssige hypoteser (Ballings et al. 2015). Den generelle idé med ensembler er at træne flere (basis) estimatorer ved additivt at kombinere de vægtede output til et samlet:

$$\hat{f}(x) = \sum_{b=1}^{B} c_b T_b(x) \tag{4.4}$$

, hvor $T_b(x)$ repræsenterer en (basis) estimator og c_b en konstant, der repræsenterer den vægt som hvert output tillægges. Problemet med ligning 4.4 bliver dermed til at finde kombinationen af (basis) estimatorer, som minimerer det empiriske træningstab. Dette er endog en udfordrende opgave, og det er ofte ikke muligt (uden en strategi) at finde en kombination, der perfekt balancerer styrkerne og svaghederne fra hver (basis) estimator. Årsagen er, at (basis) estimatorerne kan have forskelligt bias og varians, og at finde den optimale kombination vil kræve at balancere disse forskellige fejlkilder. Desuden kan interaktionerne mellem (basis) estimatorerne være komplekse, og der kan være ikke-lineære afhængigheder mellem prædikteringerne af disse. Derfor er der behov for en strategi til at udvælge (basis) estimatorerne men også til at kombinere prædikteringerne på en måde, der optimerer ensemblets præstation givet, at det beregningsmæssigt er muligt. Mulige strategier er: (1) Bagging, (2) Boosting samt (3) Stacking (Hastie et al. 2009). Denne afhandling vil som nævnt anvende RF og GB, der er specifikke typer af henholdsvis (1) og (2) og er gennemgået i afsnit 4.3.3 og 4.3.4, hvor også fordelene er præciseret yderligere.

4.3 Algoritmer

4.3.1 Simpel Lineær Regression

Den mindst komplekse metode i analysen og mest udbredte i forbindelse med empirisk prisfastsættelse er den simple lineære regressionsmodel estimeret ved hjælp af Ordinary Least Squares (OLS) metoden. Denne model er anvendt som benchmark til at sammenligne de mere komplekse ML-modeller med denne. I tilfælde af den simple lineære regression, er det muligt at modellere de betingede forventninger $f^*(x)$ ved hjælp af følgende model:

$$f(x_{i,t}, \alpha, \theta) = \alpha + \theta^T x_{i,t} \tag{4.5}$$

[4]I ensemblemetoder er (basis) estimatorer de indlæringsalgoritmer, der er anvendt til at prædiktere - de kan enten være af den samme type eller forskellige (homogene versus heterogene ensembler). Generelt set er en (basis) estimator, også kendt som en svag estimator, da den har begrænset prædikteringsevne alene, og dens fejlrate kun er lidt bedre end et tilfældig gæt. Beslutningstræer er velegnede, fordi de let kan simplificeres.

, hvor θ, $\theta^T = \theta_1, \theta_2, ..., \theta_p \in \mathbb{R}^p$, er kolonnevektoren af koefficienter, der kan estimeres med OLS ved at minimere tabsfunktionen:

$$L_{MSE}(\alpha, \theta) = \frac{1}{NT} \sum_{t=1}^{N} \sum_{t=1}^{T} (r_{i,t+1} - f(x_{i,t}, \alpha, \theta))^2 \tag{4.6}$$

Ligning 4.6 er ligeledes kendt som Mean Squared Error (MSE). OLS'en har den store fordel, at den ikke kræver et sæt af hyperparametre. Ved at minimere tabsfunktionen L_{MSE} kan der yderligere udvindes en unik analytisk løsning, der er let at fortolke, idet koefficienterne, θ, direkte beskriver, hvordan en ændring i variablene vil påvirke det forventede afkast. Ifølge Wooldridge (2019) vil OLS[5] give en effektiv og unbiased estimator, hvis antallet af observationer (N) er større end antallet af koefficienter (K), der skal estimeres. Hvis antallet af koefficienter nærmer sig antallet af observationer, har OLS svært ved at skelne mellem signalet og støjen. Mens in-sample R^2 i dette tilfælde vil være meget høj, så vil R^2 for out-of-sample prædiktering være meget lav eller måske endda negativ. Årsagen er, at at med mange koefficienter relativt til observationer, vil OLS overtilpasse dataet ved at justere koefficienterne til støjen frem for signalet. Mens signalet er den del, som kan forstås, modelleres og prædikteres, består støjen af det uforudsigelige komponent af selskabers kursbevægelser. Dette er af særlig betydning inden for kapitalaktivers prisfastsættelse, som empirisk er afhængig af en lav signal-til-støj ratio. Denne overtilpasning giver en højere præstation in-sample men en dårlig præstation out-of-sample. Disse tilsyneladende forudsigelige mønstre in-sample gentages derefter ikke i out-of-sample stikprøven, og forudsigelserne fungerer dårligt. I det ekstreme tilfælde med $K > N$, er OLS-estimaterne ikke unikke, fordi der findes et uendeligt antal løsninger, der alle passer perfekt til dataet. Yderligere kan multikollinearitet mellem koefficienterne føre til en fejlagtig fortolkning af teststatistikker såvel som vildledende koefficienter. Endelig modellerer eller evaluerer OLS ikke nogen ikke-linearitet af variablene eller potentielle interaktioner mellem dem. Enhver ikke-linearitet skal imputeres af brugeren (Wooldridge 2019, Nagel 2021).

4.3.2 Et beslutningstræ

For bedre at kunne forstå mekanismerne (og fundamentet) i både RF og GB vil det følgende afsnit introducere til beslutningstræer, der er de såkaldte (basis) estimatorer i disse ensembler.

Processen med at skabe et beslutningstræ involverer rekursiv opdeling af variabelrummet for at

[5] Se bilag G.1 for antagelserne bag dataene i en OLS-estimation.

prædiktere $r_{i,t+1}$. Opdelingen giver K distinkte og ikke-overlappende bladnoder $C_1, C_2, ..., C_K$[6]. For alle observationer, der falder ind under en bladnode C_k, gælder den samme prædiktering, som er middelværdien af $r_{i,t+1}$ for observationerne i C_k. Træet indleder med at vælge en variabel x_j og en tilsvarende tærskelværdi s således, at opdelingen af variabelrummet i områderne $x_i \mid x_{ij} < s$ og $x_i \mid x_{ij} \geq s$ realiserer den størst mulige reduktion af fejlene[7]. Hvor x_i repræsenterer det i'te datapunkt, og x_{ij} er værdien af den j'te variable for det i'te punkt. Det vil sige, at dette er to grene, der udgår fra en intern node i træet, og de repræsenterer forskellige beslutninger baseret på værdien af en variabel j og en tærskel s. En vigtig forskel mellem en intern node og en bladnode er, at en intern node (kaldet for en node i det efterfølgende) repræsenterer en beslutning baseret på værdierne af en bestemt variabel, mens en bladnode repræsenterer en endelig prædiktering (det vil sige, at der ikke sker yderligere opdelinger). Noder anvendes til at opdele variabelrummet, mens bladnoder anvendes til at tildele prædikteringer til observationerne.

I stedet for at evaluere global optimalitet (hvilket kan være beregningsmæssigt umuligt), er den grådige algoritme anvendt, som er baseret på en lokal evaluering af tilpasningen ved ethvert trin for at bestemme den bedste opdeling. Algoritmen er kaldet grådig, fordi den løser minimeringsproblemet med en brute-force tilgang ved at afprøve samtlige kombinationer af variable og tærskelværdier. Fordi opdelingen er valgt uden hensyntagen til de fremtidige (potentielle) forgreninger, er det muligt at omgå en ringere gren, der ville have ført til en fremtidig gren med en i sidste ende overlegen reduktion af fejlene (Gu et al. 2020). Mere præcist er samtlige variable tjekket for den størst mulige reduktion af fejlene, hvis opdelingen sker på baggrund af denne variabel med den tærskel (s), der maksimerer tilpasningen. Opdelingen sker således ud fra den variabel, der giver den bedste tilpasning. Denne proces med at opdele i noder, ved at vurdere alle tænkelige tærskler (s) på tværs af alle variable, som minimerer fejlen, fortsætter, indtil det punkt, hvor antallet af observationer inden for hver node når et minimumsantal (et stopkriterium sat inden start), eller hvis ingen yderligere opdeling meningsfuldt kan forbedre fejlene. Hermed opnås bladnoderne - den endelige prædiktering (Chan & Mátyás 2022). I tilfælde af et regressionstræ minimeres den såkaldte l_2-impurity (svarende til MSE) for den nye node (C):

$$H(\theta, C) = \frac{1}{\mid C \mid} \sum_{x_{i,t} \in C} (r_{i,t+1} - \theta)^2 \tag{4.7}$$

[6] I dette tilfælde repræsenterer K antallet af bladnoder i beslutningstræet, mens k refererer til indekset for en specifik bladnode.

[7] Hvor fejlene er et mål for, hvor godt prædikteringerne matcher de faktiske værdier. Det vil sige, at et regressionstræ grupperer observationer baseret på variable/karakteristika således, at observationer inden for en node er ens med hensyn til fremtidige afkast.

, hvor C er observationerne, der hører til den aktuelle node, $\mid C \mid$ er antallet af disse i (node/gren) C, og θ er prædikteringsværdien for dette C. Dermed er ligning 4.7 en minimering af forskellen mellem det faktiske afkast og det prædikterede for observationerne i C. Givet C er det tydeligt, at det optimale valg af θ er: $\theta = \frac{1}{C} \sum_{x_{i,t} \in C} r_{i,t+1}$. Denne MSE-værdi er således anvendt som et kriterium til at bestemme den anvendte variabel men også den anvendte tærskel til opdelingen. Hvor minimering af dette kriterium sikrer, at summen af kvadrerede forskelle minimeres for alle observationer inden for hver node, på tværs af alle noder i træet (Chan & Mátyás 2022), hvilket er afgørende for at forhindre en overtilpasning men også for at forbedre generaliseringsfejlen.

For et stort sæt af variable giver den ovenfor beskrevne proces endog højst sandsynligt en overtilpasset model[8]. Derfor er det vigtigt at tage højde for dybden af et træ, L, defineret som det maksimale antal opdelinger efter den længste gren fra rodnoden. Et større træ kan overtilpasse træningsdataene, mens et mindre træ (muligvis) ikke kan sikre en nøjagtig tilpasning. Grundet denne afvejning, reduceres dybden (L) af et træ efter følgende kriterium:

$$R_\alpha(L) = \frac{1}{\mid L \mid} \sum_{C \in L} \sum_{x_{i,t} \in C} (r_{i,t+1} - \theta_C)^2 + \alpha \mid L \mid \tag{4.8}$$

$R_\alpha(L)$ er omkostningskompleksitetskriteriet for et bestemt undertræ L med regulariseringsparameteren α og er anvendt til at bestemme det optimale niveau af beskæring for L i et beslutningstræ. Det første led beregner impurity-værdien af undertræet L. $\mid L \mid$ repræsenterer antallet af noder i L, C repræsenterer en specifik bladnode i L, $r_{i,t+1}$ er det sande afkast for det i'te datapunkt i bladnode C, og θ_C er den prædikterede værdi for bladnode C. Det andet led er en kompleksitetsstraf, der proportionalt med antallet af bladnoder i L modvirker overtilpasningen ved at straffe store undertræer. α er således en hyperparameter, der styrer styrken af straffen[9]. Målet med beskæring er at finde det undertræ, der minimerer omkostningskompleksitetskriteriet. Processen med at finde det optimale undertræ kan være beregningsmæssigt intenst, særligt for et større træ. En almindelig teknik til at fremskynde processen kaldes Weakest Link Pruning, som startende med det største undertræ (det vil sige hele træet), rekursivt fjerner de noder, der bidrager mindst til reduktionen af omkostningskompleksitetskriteriet, indtil et stopkriterium er

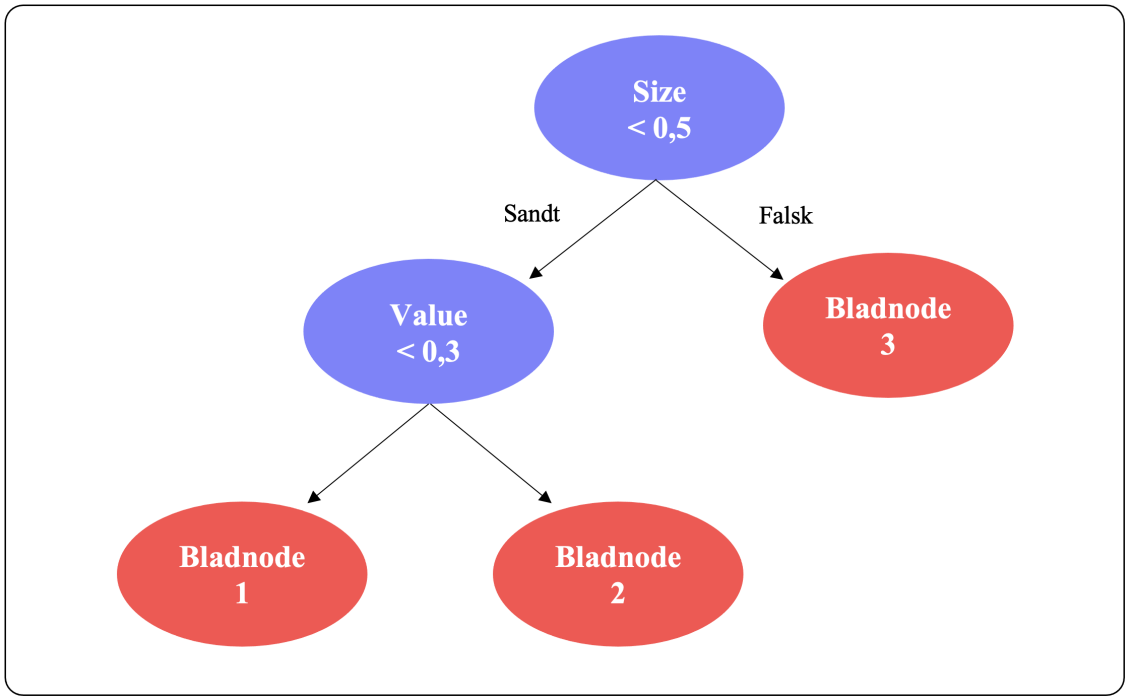
[8] Beslutningstræer er tilbøjelige til at overtilpasse, fordi de har en tendens til at skabe komplekse træer, der passer perfekt til træningsdataene men generaliserer dårligt til nye, usete data. Dette sker, fordi det er muligt for beslutningstræer at blive ved med at opdele dataene i mindre og mindre delmængder, indtil hver delmængde kun indeholder ét datapunkt, hvilket resulterer i et træ, der er alt for komplekst og specifikt for træningsdataene (fanger støjen) (Hastie et al. 2009, Chan & Mátyás 2022).

[9] Det vil sige afvejningen/balancen mellem den gode tilpasning (målt ved det første led i R_α) og træets kompleksitet (målt ved det andet led) i forhold til, hvor meget træet skal beskæres. En større værdi af α vil resultere i et simplere (det vil sige mindre) træ, mens en mindre værdi af α vil give et mere komplekst (det vil sige større) træ. For hvert α er det muligt at påvise, at der er et unikt mindste undertræ R_α, der minimerer $R_\alpha(L)$.

opfyldt. Dette fjerner de noder, der ikke bidrager til at forbedre nøjagtigheden af træets prædiktering, hvilket fører til en enklere og mere fortolkelig model, der opnår en god balance mellem at over- og undertilpasse. Samtidig bevarer træets dets evne til at prædiktere (Hastie et al. 2009).

Figur 4.2 viser et eksempel på et simpelt træ med to variable, *size* og *value*. Selve toppen af et træ, hvor den første opdeling finder sted, er kaldet rodnoden. Herfra vokser træet sekventielt. Ved ethvert trin sorterer en gren det tilbageværende data fra forgående trin baseret på én af de andre (tilbageværende) variable. Hvis en node ikke har forgreninger, er den kaldet en bladnode. Processen med at bygge et træ består i ligning 4.8, også kendt som beskæring (*pruning*).

Figur 4.2 Eksempel på et simpelt beslutningstræ med inspiration fra Gu et al. (2020)



Prædikteringen af et træ, $\mathcal{T}$, med K bladnoder og dybde L, kan skrives som:

$$f(x_{i,t};\theta,K,L) = \sum_{k=1}^{K} \theta_k \boldsymbol{I}_{x_{i,t}\in C_k(L)} \tag{4.9}$$

$$\theta_k = \frac{1}{N_k} \sum_{x_{i,t}\in C_k(L)} r_{i,t+1} \tag{4.10}$$

, $C_k(L)$ er den k'te bladnode i træet $\mathcal{T}$ med dybde L[10], θ_k er det gennemsnitlige afkast inden

[10]Med andre ord er $C_k(L)$ det sæt af datapunkter, der ender i den k'te bladnode, når beslutningstræet $\mathcal{T}$ er konstrueret til at have en maksimal dybde på L.

for noden (i den næste periode), og $\boldsymbol{I}_{x_{i,t} \in C_k(L)}$ angiver, hvorvidt $x_{i,t}$ er en del af $C_k(L)$. Ligning 4.9 illustrerer, hvordan den prædikterede værdi for et givet input $x_{i,t}$ betegnes som den vægtede sum af de faktiske afkast $r_{i,t+1}$ for bladnoderne, som $x_{i,t}$ tilhører. Ved at følge denne metode kan et træ med dybde L fange op til $L-1$ interaktioner. Eksemplet i figur 4.2 har følgende prædikteringsligning, hvor indikatorfunktionen $I(.) = 1$, hvis argumentet er sandt, og nul ellers:

$$f(x_{i,t}; \theta, 3, 2) = \theta_1 \mathbf{I}_{size_{i,t}<0,5} \mathbf{I}_{value_{i,t}<0,3} + \theta_2 \mathbf{I}_{size_{i,t}<0,5} \mathbf{I}_{value_{i,t}\geq 0,3} + \theta_3 \mathbf{I}_{size_{i,t}\geq 0,5} \tag{4.11}$$

Først sorteres observationer efter *size*, som giver den største forklaringskraft sammenlignet med den lokale opdeling baseret på *value*. Observationer over tærsklen på 0,5 tildeles bladnode 3. Tærsklen er er valgt således, at den giver den bedste tilpasning i form af MSE, som f er tilnærmet med en ligevægtet middelværdi af $r_{i,t+1}$ i de to noder. Dem med lille *size* sorteres derefter yderligere, hvor dem med *value* under 0,3 henføres til bladnode 1, mens dem med *value* over 0,3 henføres til bladnode 2. Sammenfattende handler det at konstruere et træ således om at minimere MSE (ligning 4.7) ved at finde de tærskelværdier, der bedst skelner mellem de potentielle resultater (Gu et al. 2020, Chan & Mátyás 2022). Det vil sige, at træer tilnærmer sig den ikke-lineære funktion $f(x_i)$ gennem en multidimensionel trinfunktion som et ligevægtet gennemsnit af $r_{i,t+1}$ i den specifikke bladnode. Deres fleksibilitet er dog også deres begrænsning. Træer bruges sjældent individuelt, da de er ustabile med hensyn til ændringer i de anvendte variable og er udsat for alvorlige overtilpasningsproblemer. I modsætning til OLS kan et træ kraftigt overtilpasse selv med få variable, og med tilstrækkelig dybde kan den perfekte tilpasning opnås (Moritz & Zimmermann 2016). For at afhjælpe ulemperne ved individuelle træer anvender denne afhandling to populære ensemblealgoritmer, bagging og boosting, beskrevet i de kommende to afsnit.

4.3.3 Random Forest

Som bemærket af Breiman (1996) er træer velkendte for at være ustabile, da en lille ændring i træningssættet kan få den resulterende struktur til at ændre sig betydeligt. Denne ændring indebærer, at prædikteringerne også ændrer sig, hvilket fører til høj varians i disse[11]. To populære metoder, der er mulige at anvende på træer for at forbedre deres prædikteringsevne, er Bagging og RF. De er begge baseret på aggregering, hvorved den overordnede prædiktering opnås ved at kombinere prædikteringer fra mange (basis) estimatorer, som er træer i RF (Chan & Mátyás 2022). RF er en ikke-parametrisk metode, hvilket betyder, at den i modsætning til OLS ikke gør

[11] Det skyldes, at et enkelt træ er tilbøjelig til at overtilpasse, mens beskæring ikke kan garantere stabilitet.

antagelser om den underliggende datafordeling udover, at den skal være uafhængig og identisk fordelt (i.i.d). Dette gør algoritmen særlig nyttig ved komplekse data, hvor relationerne mellem variablene ikke er godt forstået, da det ikke er nødvendigt at specificere den funktionelle form. I stedet er RF baseret på et stort antal træer, der trænes på bootstrappede datasæt, og hvert træ er konstrueret ved hjælp af en tilfældig delmængde af de tilgængelige variable. Denne randomisering hjælper med at reducere overtilpasningen (og øge generaliserbarheden). Det betyder, at RF er et populært alternativ til penalisering (beskæring, ligning 4.8) af størrelsen på et træ.

Bagging

Bagging (Bootstrap Aggregering/Bootstrapping) er en generel ensemblemetode, der er mulig at anvende med enhver type (basis) estimator, herunder beslutningstræer. Bootstrapping er hovedideen bag bagging og en statistisk metode til at generere en række nye datasæt fra det originale. Fra det originale datasæt, defineret som $\boldsymbol{Z} = \{Z_1, Z_2, ..., Z_n\}$, hvor $Z_i = (x_i, y_i)$ for $[n]$ givet ved $1, ..., n$, er et bootstrap datasæt med tilbagelægning udtrukket, der har samme størrelse (n) som det originale (de er i.i.d.). For dette bootstrap datasæt $\boldsymbol{Z^{*b}}$, $b = 1, 2, ..., B$, tilpasses en estimator, hvilket resulterer i en given prædiktering $f^{*b}(x)$. Denne proces er gentaget B gange[12] for slutteligt at tage middelværdien af prædikteringerne over samlingen af bootstrappede datasæt:

$$f_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B} f^{*b}(x) \tag{4.12}$$

Baggings største fordel er, at den betydeligt kan reducere prædikteringens varians og dermed afhjælpe overtilpasningsproblemet. Dette vil ske, idet middelværdien af ethvert konstrueret træ, som er trænet på forskellige bootstrap datasæt, reducerer variansen. Variansen af $f_{bag}(x)$ (4.12) er en funktion af antallet af baggede estimatorer (B), den gennemsnitlige varians af den enkelte estimators prædiktering (σ^2) og den gennemsnitlige korrelation mellem prædikteringerne (ρ):

$$Var(f_{bag}) = \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2 \tag{4.13}$$

Det første led er variansen grundet korrelationen mellem bootstrap-estimaterne. Det andet led er variansen grundet støjen i bootstrap-estimaterne. Det vil sige, at dette er variansen af gennemsnittet af bootstrap-estimaterne, som reduceres, efterhånden som B øges, og er den mængde af varians, der er mulig at reducere ved at øge B[13]. Ligningen (4.13) viser imidlertid, at bagging

[12] Det vil sige, at estimatorerne tilpasses uafhængigt af hinanden, hvorfor modellen kan tilpasses parallelt.

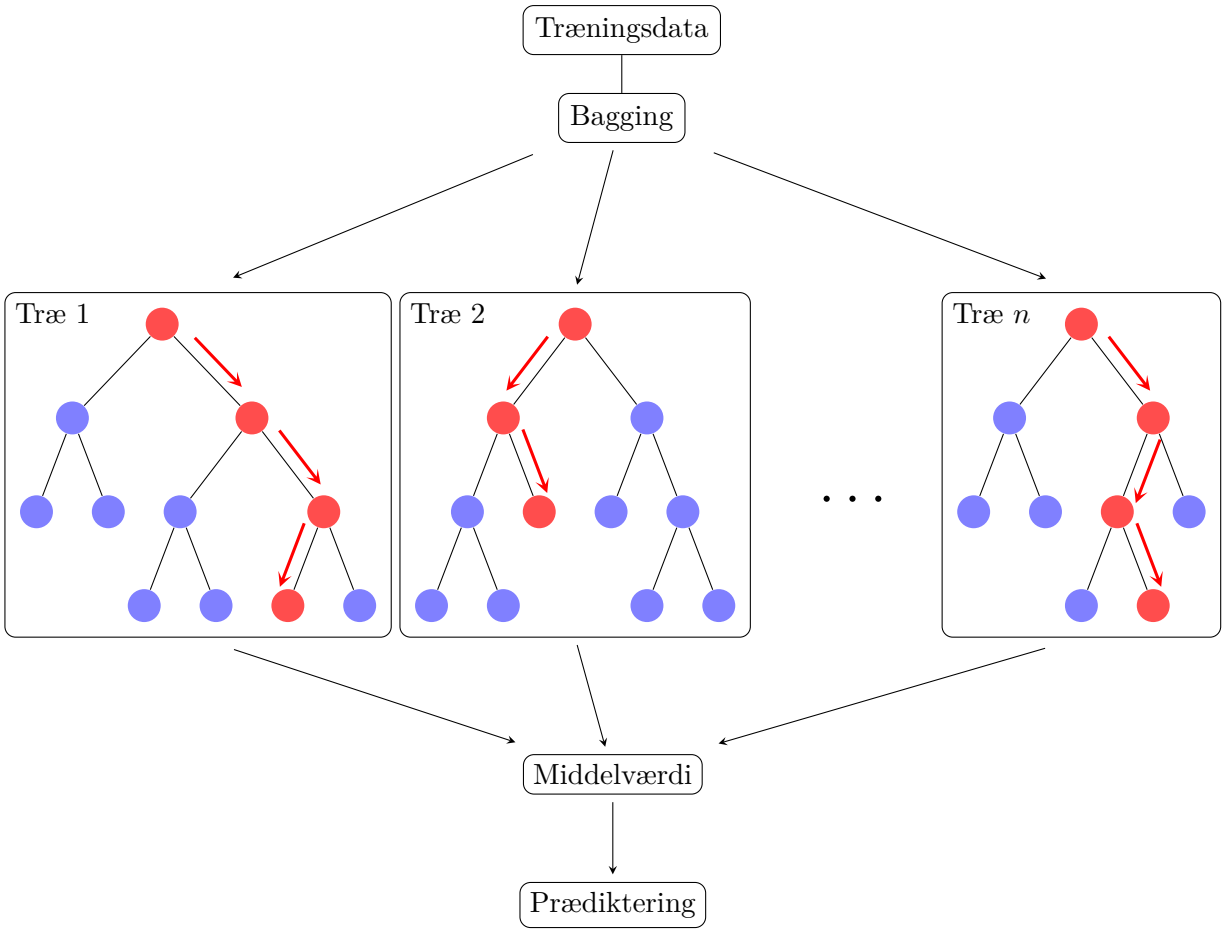
[13] Det vil give en nedre grænse for variansen af bagging-estimatet. Dette udtryk kan dog aldrig helt elimineres, selvom B nærmer sig uendeligt, fordi der altid vil være en vis mængde støj i de individuelle estimater.

kun er effektivt i den udstrækning, at $\rho < 1$; hvilket skyldes, at som $\rho \to 1 \Rightarrow Var(f_{bag}) \to \sigma^2$. Det betyder, at bagging-estimatet ikke er meget mere nøjagtigt end ethvert individuelt estimat, som variansen af bagging-estimatet vil nærme sig variansen af et enkelt estimat (σ^2). Et af målene med bootstrapping er dog at producere datasæt så uafhængige som muligt[14] for derved at reducere ρ, som vil reducere variansen af bagging-estimatet. Her vil bagging-estimatet være meget mere nøjagtigt end ethvert individuelt estimat. Som B øges, vil det andet udtryk forsvinde men det første forblive og derfor vil effektiviteten af bagging til at reducere variansen i sidste ende afhænge af graden af korrelation mellem bootstrap-estimaterne (ρ) (Hastie et al. 2009).

Random Forest som en udvidelse af bagging

RF deler mange ligheder med bagging, i betydningen af at træne uafhængige individuelle estimatorer over bootstrappede datasæt, men ideen er at forbedre variansen yderligere ved at reducere korrelationen mellem estimatorerne (ρ). Hvis x_i har p variable, udvider RF bagging-tilgangen ved tilfældigt kun at overveje $m < p$ variable ved enhver node fra det bootstrappede datasæt. Det betyder, at kun en delmængde af alle variable er anvendt som mulige i opdelingsprocessen med at konstruere ethvert træ i skoven. Dette yderligere element af tilfældighed, sammenlignet med bagging, reducerer korrelationen på tværs af træerne, hvilket ser ud til at reducere variansen men også forbedre den overordnede prædiktering (Breiman 1996, 2001). Intuitivt indeholder de bootstrapede datasæt til en vis grad forskelligt støj. Følgelig, hvis de konstruerede bootstrap træer overtilpasses til denne støj, vil de resulterende prædikteringer blive påvirket forskelligt af dette. Desuden vil det tilfældige udvalg af variable forstørre disse forskelle. Ved at konstruere flere (bootstrap) træer og aggregere prædikteringerne, er det mindre sandsynligt, at RF vil huske støjen fra træningsdataene og er mere tilbøjelig til at fange underliggende mønstre. Middelværdien vil således udligne mange af disse støj-inducerede effekter (og reducere overtilpasningen) (Breiman 2001). I lighed med bagging aggregeres det samlede til en endelige prædiktering, jf. figur 4.3 (der grafisk illustrerer en RF-algoritme). Den røde sti indikerer vejen til den endelige prædiktering for et sæt af observationer (tilhørende en aktie i på tidspunkt t) baseret på, hvilken side af tærskelværdierne, de tilhører. Middelværdien af prædikteringerne fra bladnoderne er således prædikteringen tilhørende dette sæt (for i ved tidspunkt t for $t + 1$).

[14] Det er dog ikke altid muligt at producere helt uafhængige bootstrap-datasæt, særligt ikke hvis det originale datasæt har stærke afhængigheder, eller hvis størrelsen af datasættet er lille. I disse tilfælde kan korrelationen mellem bootstrap-estimaterne begrænse fordelene ved bagging, selv når antallet af bootstrap-prøver B stiger.



Figur 4.3 Random Forest

Forbedringerne i prædikteringsevnen (opnået ved såvel bagging som RF) er imidlertid (primært) et resultat af variansreduktion. Det skyldes, at metodernes bias typisk er større (i absolutte tal) end for det ubeskårede træ som følge af randominiseringen men også det reducerede stikprøverum. Hvilket alt andet lige pålægger en række restriktioner, som metoderne ikke har al tilgængelig information, og som nogle af sammenhængene (mellem x og y) kan være sløret. I et simpelt træ vælges den bedste variabel ved samtlige noder, hvilket kan resultere i en høj varians grundet en overtilpasning, hvis træet er dybt og komplekst. I modsætning hertil konstruerer RF flere træer ved hjælp af tilfældigt udvalgte variable ved samtlige noder, hvilket kan hjælpe med at reducere overtilpasningen men også forbedre generaliseringsfejlen. Denne randominisering kan endog også føre til en lille stigning i bias, fordi nogle informative variable udelukkes fra nogle af træerne, hvilket resulterer i tab af information. Denne stigning i bias opvejes dog normalt af en større reduktion i varians, hvilket resulterer i forbedret prædikteringsevne (Hastie et al. 2009).

Sammenfattende er begrænsningerne pålagt af RF (og bagging) en afvejning mellem bias og varians, og fordelene i form af variansreduktion og forbedret prædikteringsevne opvejer typisk den potentielle stigning i bias, men balancen findes ved omhyggeligt at justere hyperparametrene.

Et eksempel på implementering af en forenklet version af Breimans RF-algoritme er at finde i nedenstående **Algoritme 2**.

Algorithm 1 Breimans originale RF-algoritme med inspiration fra Hastie et al. (2009)

1: **for** $b = 1$ til B: **do**
2: (a) Udtræk et bootstrap datasæt $\boldsymbol{Z}^*$ fra træningsdataene.
3: (b) Byg et RF-træ fra det bootstrappede data ved rekursivt at gentage følgende trin for noderne i træet, indtil den mindste nodestørrelse n_{min} er nået:
4: i. Vælg m variable tilfældigt fra de p variable.
5: ii. Vælg den bedste variabel (og det bedste splitpunkt, s) blandt de m variable.
6: iii. Opdel noden i to datternoder/to forgreninger.
7: **end for**
8: **Result:** Det endelige RF output er givet ved gennemsnittet af alle B-træer:

$$f(x_{i,t}, L, B) = \frac{1}{B} \sum_{b=1}^{B} f_b(x_{i,t}, \theta, L)$$

RF er en interessant ML-tilgang, da den kan tage højde for både korrelationen og interaktionen mellem variable. Gennemsnitsbetragtningen beskytter mod risikoen for at overtilpasse in-sample samtidig med, at den håndterer de hårde tærskler, som opdelingsprocessen inducerer. Omvendt kommer fleksibiliteten ikke uden omkostninger, for med gennemsnitsbetragtningen er den simple fortolkning, fra det enkelte beslutningstræ, tabt. Derudover er det ikke muligt at opsummere modellen med en simpel ligning i rummet af selskabsspecifikke karakteristika. En af årsagerne til populariteten af OLS er i dens gennemsigtighed (Moritz & Zimmermann 2016).

4.3.4 Gradient Boosting

Konceptet med GB blev først introduceret af Friedman (2001) og er siden blevet en **state-of-the-art** ML-teknik med en række konkurrencedygtige implementeringer, som f.eks. XGBoost (Chen & Guestrin 2016). Forskelligt fra RF, forbedrer GB den tidligere model i ethvert trin ved at korrigere dennes fejl, hvorfor træerne er udvalgt på en afhængig måde sammenlignet med den uafhængige måde i RF (de individuelle beslutningstræer). Men ligesom RF, er GB en ikke-parametrisk metode (en ensemblealgoritme), der kombinerer prædikteringer fra flere svage estimatorer med henblik på at producere en stærkere (overordnet) prædiktering. GB er indledt med at træne et svagt regressionstræ på det fulde træningssæt. Dette oversimplificerede træ er

en svag (og biased) estimator men initialiserer en prædiktering, der kan forbedres gennem en trinvis optimering af den anvendte tabsfunktion. I næste trin trænes et andet regressionstræ med samme dybde L på residualerne fra det første træ. Prædikteringerne fra disse to lægges sammen for at danne en ensembleprædiktering af resultaterne, mens bidraget fra det andet træ reduceres med faktoren υ (indlæringshastigheden), hvor $\upsilon \in (0, 1)$, for at undgå, at modellen overtilpasses residualerne. Det vil sige, at hvert træ lærer af det tidligere ved at gå et skridt i den korrekte retning af de observerede værdier. Ved ethvert trin b, indtil modellen har nået B-træer, konstrueres et nyt (lavt) træ på residualerne, som er baseret på den $b-1$'te model og lægges til det samlede med en vægt på υ. Den endelige prædiktering er således en additiv model af (lave) træer med tre tuningsparametre (L, υ, B), som adaptivt er valgt i valideringstrinnet.

Boosting

Boosting er en fremadrettet, trinvis additiv teknik, hvilket betyder, at den grundlæggende består i at skabe flere estimatorer i en fortløbende rækkefølge, ved at minimere en tabsfunktion, hvor hver estimator tager hensyn til både resultatet fra de tidligere estimatorer men også y (Hastie et al. 2009). På den måde bliver ensemblet af svage estimatorer boostet til en stærk estimator, som er hele ensemblet. Formålet med at booste er således at minimere generaliseringsfejlen for det endelige ensemble ved at minimere en tabsfunktion. For at løse dette optimalt bør samtlige svage estimatorer blive tilpasset under hensyntagen til alle andre estimatorer i ensemblet, både tidligere og fremtidige. Da dette problem generelt er vanskeligt at løse, udføres boosting på en sekventiel, grådig måde. Det betyder, at med beslutningstræer, $f^{(b-1)}$, hvor $b-1$ allerede er udvalgt, reduceres det teoretiske mål for f_b, det b'te og næste trin i boosting, til:

$$f_b(x) = \arg\min_{f_b} E_{x,y}[l(y, f^{(b-1)}(x) + \upsilon_b \cdot f_b(x))] \tag{4.14}$$

, hvor $f_b(x)$ er prædikteringen fra det b'te regressionstræ på inputvariablene x. $l(y, f^{(b-1)}(x) + \upsilon_b \cdot f_b(x))$ er tabsfunktionen, som er søgt minimeret. Denne måler forskellen mellem det sande y og det prædikterede, idet der tages højde for prædikteringerne fra de tidligere træer såvel som det aktuelle træ, der er ved at blive konstrueret. Det overordnede mål med denne funktion er at finde det bedst mulige regressionstræ f_b, der minimerer den forventede værdi af tabsfunktionen over træningsdataene (input-output parene (x, y)). Når det optimale f_b er fundet, føjes det til ensemblet af tidligere træer, og processen gentages for at konstruere det næste træ i sekvensen. Hermed tilpasser boosting en additiv model af formen:

$$f^B_{boosting}(x) = \sum_{b=1}^{B} \upsilon_b \cdot f_b(x) \tag{4.15}$$

, hvor $f^B_{boosting}(x)$ er ensemblet af B træer, hvilket skal ses som den aktuelle tilnærmelse af f i $y_i = f(x_i)+\varepsilon$ på trin B, som dermed er den endelige prædiktering i form af (basis) estimatorerne $f_b(x)$. Ved hvert trin tilpasses det b'te træ til tabsfunktionen givet tilnærmelsen af f bestående af de første $(b-1)$ træer, $f^{(b-1)}(x)$, således, at det bedste estimat er givet ved ligning 4.14.

Gradient descent

For at kunne løse 4.14 er det nødvendigt med en optimeringsalgoritme. **Gradient descent** er en algoritme, hvor ideen er at finde retningen af gradienten for den funktion f, som er ønsket minimeret, for dernæst at tage et skridt i denne retning. Gradienten bruges til at vise, hvilken retning parametrene bør ændres for at nærme den bestemte funktions minimum[15]. Det vil sige, at det er en førsteordens iterativ optimeringsalgoritme til at finde et lokalt minimum af en differentierbar funktion. Hvis $f = f(x, \theta)$ er objektivfunktionen, som er ønsket minimeret, er ideen at tage gradienten (den partielle afledede) af f med hensyn til θ, $\partial_\theta f(x, \theta)$, og dernæst opdatere parametrene θ i den negative retning af gradienten (da det vil være retningen for den største nedstigning). Dernæst genberegnes gradienten med hensyn til de opdaterede parametre, og den vil tage endnu et skridt i denne retning. Bevægelse i denne retning vil i sidste ende få gradienten i samme retning til at være ikke-negativ, hvilket betyder, at det ikke længere er muligt at forbedre tabet ved at bevæge sig på denne linje. På det tidspunkt er det nødvendigt igen at finde gradienten ved det nye punkt og lave en ny minimering i den nye største nedstigning. Normalt er ønsket ikke at bevæge sig helt til minimum ved samtlige nedstigningslinjer, idet bevægelse fra startpunktet kan ændre gradientens retning. Standardproceduren er således kun at bevæge sig i gradientens retning med en lille trinstørrelse υ, også kaldet indlæringshastigheden, sådan, at $\theta = \theta - \upsilon \cdot \partial_\theta f(x, \theta)$. Efter første beregning af gradienten, ændres parametrene marginalt for igen at udregne gradienten. Processen gentages indtil den bestemte funktion har konvergeret til et minimum, hvormed parametrene er ændret til det bestemte sæt, der medfører et minimum af den bestemte funktion (Hastie et al. 2009).

[15] Gradienten viser endog, hvilken retning parametrene skal ændres for at øge outputtet af den bestemte funktion, derfor skal parametrene ændres i den modsatte retning for at nå et minimum.

Kombinationen af boosting og gradient descent i GB

GB er baseret på gradient descent-optimeringsteknikken men i modsætning hertil, er den ikke foretaget i parameterrummet men snarere i funktionsrummet. Hver (basis) estimator er tilpasset den negative gradient af tabsfunktionen i forhold til den aktuelle models output snarere end dens parametre. Friedmans idé var at inducere et træ ved den b'te iteration, hvis prædikteringer er så tæt som muligt på den negative gradient[16], $-g_b(x_i)$, ved en mindste kvadraters optimering:

$$\theta_b = \arg\min_{\theta_b} \sum_{i=1}^{N} (-g_b(x) - \upsilon_b \cdot f(x, \theta_b))^2 \tag{4.16}$$

, hvor gradienten er den negative første afledede af tabsfunktionen i forhold til det prædikterede fra det foregående træ[17] $f^{(b-1)}(x)$:

$$-g_b(x) = -\frac{\partial l(f^{(b-1)}(x))}{\partial f^{(b-1)}(x)} \tag{4.17}$$

Ovenstående ligning 4.16 repræsenterer optimeringsmålet for GB-algoritmen, som vil søge efter det bedste parametresæt θ_b (det vil sige de optimale bladværdier) for det b'te træ ved at minimere summen af kvadrerede fejl mellem de negative gradientværdier, $-g_b(x_i)$, og prædikteringerne fra træet, $f(x_i, \theta_b)$ (med parametrene θ_b, der vil blive optimeret ved iteration b). Den negative gradient er anvendt for, at det b'te træ kan lære af de foregåendes fejl og tilsvarende tilpasse prædikteringerne hertil - et skridt i retningen af de observerede. Det vil sige, at ligning 4.16 iterativt viser processen med at tilpasse træerne til de negative gradientværdier ved at finde de bladværdier, der er i bedst overensstemmelse hermed. Slutteligt er gradienten (implicit) anvendt i linjesøgningen langs den stejleste nedstigning for at finde den optimale indlæringshastighed, υ_b, ved hver boostede iteration således, at nedenstående objektivfunktion er minimeret:

$$\upsilon_b = \arg\min_{\upsilon_b} \sum_{i=1}^{N} l(y, f^{(b-1)}(x) + \upsilon_b \cdot f(x, \theta_b)) \tag{4.18}$$

Det optimale υ_b er det, der minimerer tabsfunktionen l, når de forrige prædikteringer kombineres med de nye. Ved det optimale υ_b opdateres modellen, og processen gentages for den næste boosting iteration $b + 1$.

[16] Det vil med andre ord sige det $f(x, \theta)$ (træ), der er højest korreleret med $-g_b(x_i)$ over datafordelingen. Dette kan fortolkes som den stejleste nedstigningsretning i et n-dimensionelt datarum evalueret ved $f^{(b-1)}(x)$.

[17] Det vil sige retningen og størrelsen af den stejleste nedstigning i tabsfunktionen. Ved at tilpasse et nyt træ til den negative gradient, sigter algoritmen på at forbedre forudsigelserne ved at bevæge sig i retning af den stejleste nedstigning, der kan ses som et mål for residualerne eller fejlene fra den tidligere model. Dermed reduceres fejlene gradvist ved at justere prædikteringerne hertil.

Grundlæggende påføres begrænsningerne ved at tilpasse $f(x, \theta)$ (træet) til pseudoresidualerne, hvor y er substitueret med $-g_b(x)$ (gradienten). Dette tillader erstatning af det vanskelige funktionsminimeringsproblem i 4.14 med mindste kvadraters funktionsminimering (4.16), efterfulgt af kun en enkelt parameteroptimering baseret på det oprindelige kriterium (4.18). For ethvert $f(x, \theta)$, for hvilken der findes en mulig mindste kvadraters algoritme til løsning af 4.16, er det således muligt at anvende denne tilgang til at minimere ethvert differentierbart tab i forbindelse med en fremadrettet, trinvis additiv modellering (Hastie et al. 2009, Friedman 2001).

Mens prædikteringen i RF primært er forbedret via dens varians, er den i GB primært forbedret via dens bias. Det skyldes, at i samtlige iterationer af GB-algoritmen tilpasses et nyt træ til residualerne fra den tidligere iteration. Det nye træ trænes til at prædiktere residualerne (i stedet for det oprindelige y), og hermed fokuserer det nye træ på de områder, hvor de tidligere træer klarede sig dårligt, hvilket gradvist reducerer algoritmens bias (Hastie et al. 2009). GB har de samme fordele som RF, da den er i stand til at håndtere komplekse og ikke-lineære relationer, har en mekanisme til at forhindre overtilpasning men ligeledes en evne til at opnå høj nøjagtighed (Friedman 2001). En overtilpasning er dog mulig, hvis B er for stort, selvom denne overtilpasning har en tendens til at forekomme langsomt, hvis overhovedet (Hastie et al. 2009). Derudover inkluderer ulemperne algoritmens sensitivitet over for dens hyperparametre, og at processen er beregningsmæssigt meget tung, da træerne er konstrueret sekventielt. En anden ulempe inkluderer dens kompleksitet med hensyn til fortolkning sammenlignet med OLS.

4.3.5 Kombination af RF og GB

Mens både RF og GB er træbaserede algoritmer, er der nogle forskelle, herunder algoritmernes håndtering af bias-varians afvejningen. RF tenderer i højere grad til at have et højere bias men en lavere varians sammenlignet med GB, der typisk har et lavere bias og en højere varians[18]. Det vil sige, at som RF og GB repræsenterer forskellige bias-varians afvejninger og har deres respektive fordele, anvender flere artikler, med gode resultater, modelaggregering til at forbedre porteføljens præstation ved at udligne gennemsnittet af støj fra de to modeller. Det er på trods af, at de to modeller allerede udnytter fordelene ved at bygge ensembler som en integreret del af selve modeldefinitionen. Denne afhandling følger dermed fremgangsmetoden givet i litteraturen (Krauss et al. 2017, Gu et al. 2020, Drobetz & Otto 2021). Heri kombineres ML-modellerne ved

[18]RF har typisk et højere bias end GB, da træerne er konstrueret på en uafhængig måde, hvilket vil føre til en undertilpasning, hvis de er for få eller for simple. Omvendt har GB ofte et lavere bias end RF, fordi træerne er konstrueret sekventielt for at korrigere fejlene fra de tidligere. RF har typisk en lavere varians, da gennemsnittet over træerne, kan reducere virkningerne af støjende (eller irrelevante) variable eller tilfældig variation i dataene. Hvorimod GB kan føre til en overtilpasning til træningsdataene, hvis træerne er for komplekse eller for mange.

at anvende gennemsnittet af modellernes prædiktering for hvert selskab (i) på ethvert tidspunkt (t). Lad $\hat{r}_{i,t+1}^{ENS^{(1)}}$ angive ensembleprædikteringen af algoritmerne og denne være defineret som:

$$\hat{r}_{i,t+1}^{ENS^{(1)}} = \frac{1}{2}\left(\hat{r}_{i,t+1}^{RF} + \hat{r}_{i,t+1}^{GB}\right) \tag{4.19}$$

Det simple ligevægtede gennemsnit af prædikteringerne fra RF og GB. Dermed begrænser afhandlingen sig til et simpelt kombinationsskema, der ofte viser bedre resultater end mere sofistikerede metoder, der involverer yderligere parameterestimater (Elliott et al. 2006).

Generelt set fungerer disse prædikteringskombinationer (eller ensembler) som et værktøj til at risikosprede på tværs af individuelle modeller, mens de i mange tilfælde samtidig forbedrer præstationen (Elliott et al. 2006). Ifølge Dietterich (2000) er der tre grunde til, at ensembler med succes er anvendt i ML. For det første som følge af en statistisk fordel. Hver (basis) estimator søger i funktionsrummet F for at identificere den optimale funktion. Men set i lyset af begrænsede data sammenlignet med størrelsen af F, er det normalt at finde flere funktioner i F, der giver tilsvarende nøjagtighed på træningsdataene. Ved at tage et gennemsnit af disse funktioner reduceres risikoen for at vælge den forkerte funktion. For det andet som følge af en beregningsmæssig fordel. De anvendte ML-modeller udfører en lokal søgning i funktionsparameterrummet, der er tilbøjelige til at sidde fast i lokale optima. Et gennemsnit på tværs af flere af disse ML-modeller kan således resultere i en bedre tilnærmelse af den sande men ukendte funktion. For det tredje som følge af en repræsentativ fordel. Ofte er den sande men ukendte funktion ikke et element i F. Hvis der tillades kombinationer af flere funktioner fra F, øges løsningsrummet for repræsentative funktioner betydeligt. Sidstnævnte kan så omfatte den ukendte funktion. Dermed udgør et simpelt gennemsnit i mange tilfælde en relevant og effektiv tilgang.

I denne afhandling er kombinationen også anvendt som en proxy til at teste, om OLS indeholder information, der er komplementær til ML-modellernes. Hermed dannes også en kombination fra alle tre modeller:

$$\hat{r}_{i,t+1}^{ENS^{(2)}} = \frac{1}{3}\left(\hat{r}_{i,t+1}^{OLS} + \hat{r}_{i,t+1}^{RF} + \hat{r}_{i,t+1}^{GB}\right) \tag{4.20}$$

Hvis der er yderligere værdi, forventes ligning 4.20 at give overlegne prædikteringer.

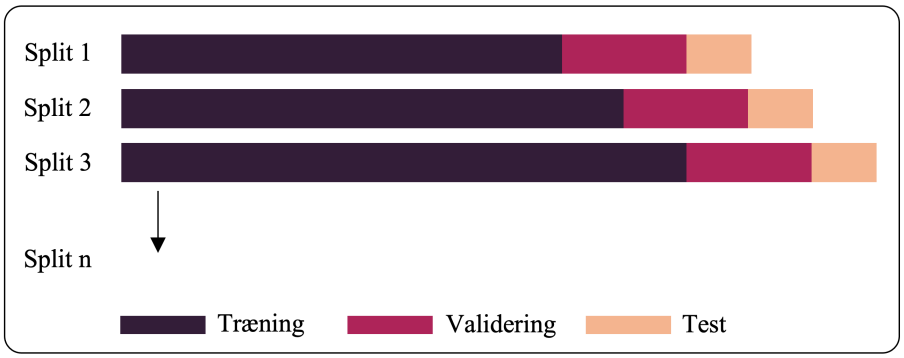
4.4 Hypertuning (Træning, Validering og Test)

Målet er at få algoritmerne til at præstere godt i out-of-sample prædikteringen. Derfor søges de hyperparametre, der minimerer prædikteringsfejlen. For at kunne minimere prædikteringsfejlen er det nødvendig med et estimat af denne for en given værdi af hyperparametrene. Det er ikke muligt at anvende in-sample fejlen fra træningsdatasættet, da dette er et (nedad) biased estimat af den faktiske out-of-sample fejl. En fuldstændig datadreven tilgang er krydsvalidering (KV). I KV bruges modelestimaterne opnået fra træningsdataene til at lave prædikteringer på et separat, uafhængigt valideringsdatasæt. Modellens fejl på dette valideringsdatasæt giver derefter et estimat af prædikteringsfejlen, som er minimeret, når hyperparametrene tunes (Nagel 2021). Det vil sige, at hyperparametertuningen iterativt reducerer in-sample tilpasningen ved at søge efter graden af modelkompleksitet, som vil producere pålidelig out-of-sample prædiktering. Til dette formål vil afhandlingen anvende en tidsseriekrydsvalideringstilgang, som vil inddele datasættet i de tre perioder; træning, validering og test, der altid vil opretholde denne tidsmæssige rækkefølge (Drobetz & Otto 2021). Først er algoritmerne estimeret for en række hyperparametre baseret på træningsdataene. Dernæst er det respektive tab af ethvert hyperparametersæt i valideringsdataene bestemt. Slutteligt er modellerne anvendt til at prædiktere de månedlige afkast for testdataene. Bemærk, at fordi hyperparametrene er valgt fra valideringssættet, er det ikke rigtigt out-of-sample. Testsættet bruges dog hverken til modelestimering eller parameterjustering. Derfor er det passende til at evaluere en models out-of-sample prædikteringsevne.

Denne afhandling følger standardtilgangen i litteraturen (Gu et al. 2020, Tobek & Hronec 2021, Drobetz & Otto 2021, Leippold et al. 2022) med hensyn til inddelingen i træning, validering og test. I en porteføljekontekst, hvor nye data kommer til med tiden, er det nødvendigt at anvende en opdelingsorden, der periodisk omfatter nyere data (se f.eks. West (2006) for en oversigt). Dette er grunden til, at metoderne **rullende** og **rekursivt vindue** gradvist forskyder trænings- og valideringssættene frem i tiden. Førstnævnte metode holder længden af trænings- og valideringssættene konstante; sidstnævnte øger dem gradvist. Derudover, fordi den rekursive tilgang altid inkorporerer hele datahistorikken, er den beregningsmæssigt mere intensiv end den rullende tilgang. På grund af dette, og fordi algoritmerne generelt er beregningsintensive, bruger Gu et al. (2020) en hybrid af de to tilgange, hvor de rekursivt undgår at skulle genoptræne dem for samtlige måneder. I stedet genoptrænes de en gang om året, som de fleste af de selskabsspecifikke karakteristika opdateres årligt. Hver gang øger de træningssættet med ét år, mens de holder længden af valideringssættet konstant, men ruller det ét år frem. Denne afhandling følger samme opdelingsorden som i Gu et al. (2020) (se figur 4.4). Først er algoritmerne estimeret baseret på

træningssættet fra 1995(1) til 1999(11). Dernæst bestemmes de bedste hyperparametre gennem valideringssættet fra 1999(12) til 2002(11). Endelig genoptrænes modellen med de optimale hyperparametre ved hjælp af data fra 1995(1) til 2002(11) og evalueres i testsættet fra 2002(12) til 2003(11). Denne tilgang sikrer, at der ikke lækkes fremtidige oplysninger fra en senere periode. Desuden tager den højde for den potentielle tidsvarierende betydning af forskellige variable og svarer på, om træbaserede porteføljesorteringer i princippet kan bruges til handelsformål.

Figur 4.4 Det anvendte opdelingsskema som i Gu et al. (2020)



Valget af hyperparametertuningsteknik er faldet på den populære grid search. Den starter med, at brugeren definerer et søgerumsgitter, der består af udvalgte hyperparameternavne og -værdier, hvor den udtømmende søger efter den bedste kombination af disse værdier. Grid search finder altid den bedst præsterende model med værdier nævnt i gitteret, men hvis gitteret er valgt dårligt, så vil selv den bedste kombination ikke fungere godt. Ulempen ved denne teknik er endog, at med et stigende antal værdier, der skal testes, kan den nemt blive meget beregningsintensiv. Hypertuning er som tidligere nævnt essentielt for at kunne opnå den bedst mulige bias-varians afvejning. Den optimale tuning er underlagt det specifikke data (Hastie et al. 2009), og afhængigt af algoritmen findes en vifte af mulige parameterkombinationer. Som følge af, at analytisk estimering af den bedste kombination ikke er muligt, er **trial-and-error** anvendt til at tilnærme sig de optimale parametre (eller søgerumsgitre), der er bestemt på modelscoren opnået fra valideringssættet (modellens ydeevne med det bedste sæt hyperparametre i form af den lavest opnåede MSE). Denne evalueringsmetrik vil guide hyperparameteroptimeringsprocessen.

De anvendte søgerumsgitre er desuden benchmarked mod standardparameterværdierne i Scikit Learn-biblioteket, hvilket giver en gunstig kombination med hensyn til ressourceforbrug og modelkompleksitet. Det skyldes, at RF er en ML-algoritme, som er kendt for at give gode resultater med standardparameterværdierne (Fernández-Delgado et al. 2014). Eksempelvis måler Probst et al. (2019) tunability af algoritmers hyperparametre og konkluderer, at RF er langt

mindre tunable end andre algoritmer, også sammenlignet med GB, der kan have høj variabilitet i dens nøjagtighed afhængigt af dens hyperparametre. Ikke desto mindre kan RF opnå en mindre præstationsgevinst via tuning sammenlignet med standardparameterværdierne, mens GB med optimerede hyperparametre ofte vil overgå RF, hvilket kan ses i dens meget stærke præstation i ML-konkurrencer (Chen & Guestrin 2016). Sammenfattende betyder det, at RF er en mere stabil algoritme med hensyn til dens hyperparametre og kræver mindre tuning end GB.

Ydermere har denne afhandling taget udgangspunkt i de anvendte hyperparametre hos Gu et al. (2020), Tobek & Hronec (2021), Drobetz & Otto (2021) og Leippold et al. (2022), men har også afsøgt både høje, lave og tilfældige værdier. De i litteraturen anvendte er forsøgt, da karakteren af dataene er ens, og ved at indsætte deres gittersøgning vil det give en intuition om, hvorvidt værdierne skal øges eller reduceres. Men da de har et større datasæt og andre variable, er det ikke sikkert, at deres parametervalg er de optimale i denne afhandling. Som tidligere nævnt er de alle sammenlignet med hinanden og til standardværdierne via modelscoren[19]. Tabel 4.1 opsummerer det anvendte hyperparmeterområde for ML-algoritmerne (OLS har ingen hyperparametre).

Tabel 4.1 Hyperparameterområdet

		Specifikation	Definition
OLS			
RF	L	$\{1, 2, 3, 4, 5, 6, 7\}$	Dybden af det enkelte regressionstræ
	M	$\{3, 5, 10, 20, 30\}$	Antal variable tilfældigt betragtet som splitvariable
	B	$\{100, 200, 300, 400\}$	Antal træer tilføjet til ensemble-prædikteringen
GB	L	$\{1, 2\}$	Dybden af det enkelte regressionstræ
	υ	$\{0,01, 0,05, 0, 1\}$	Vægt for indlæringshastigheden
	B	$\{100, 200, 300, 400, 500\}$	Antal træer tilføjet til ensemble-prædikteringen

4.5 Evaluering

Som i Gu et al. (2020) er out-of-sample R^2 anvendt til at vurdere prædikteringsevnen, men dette er imidlertid også en meget anvendt tilgang i finans- og forecastinglitteraturen:

$$R^2_{OOS} = 1 - \frac{\sum_{(i,t)\in\mathcal{T}_3}(r_{i,t+1} - \hat{r}_{i,t+1})^2}{\sum_{(i,t)\in\mathcal{T}_3}(r_{i,t+1} - 0)^2} \tag{4.21}$$

[19] Bilag D sætter lidt flere ord på hypertuningsprocessen.

, hvor $\mathcal{T}_3$ indikerer, at tilpasningen kun er vurderet på testdatasættet, hvis data aldrig er indgået i modelestimeringen eller -tuningen. Dermed måler R^2_{OOS}-statistikken den procentvise variation i merafkastet forklaret af de anvendte variable. Den samler prædikteringsfejl på tværs af selskaber (i) over tid (t) i en større panelvurdering af algoritmerne. R^2_{OOS}-statistikken er forholdet mellem MSE for den valgte algoritme (tælleren) og den for en naiv prædiktering på nul (nævneren). Tælleren kvantificerer den uforklarede variation i merafkastet og nævneren variationen heri. Som sådan er det muligt, at R^2_{OOS} er negativ, hvis den valgte algoritme klarer sig dårligere end denne naive prædiktering (hvilket typisk vil ske, hvis modellen overtilpasser støjen frem for for signalerne). Et vigtigt aspekt af ovenstående R^2_{OOS}-statistik er, at nævneren er summen af kvadrerede merafkast uden *de-meaning*. I mange out-of-sample prædikteringsapplikationer sammenlignes prædikteringerne med det historiske (gennemsnitlige) afkast. Selvom denne tilgang f.eks. er fornuftig for et samlet indeks (eller for porteføljer), er den særlig mangelfuld, når det kommer til at analysere individuelle aktieafkast. At prædiktere merafkast med det historiske gennemsnit underpræsterer typisk en naiv prædiktering på nul (og endda med en stor margin). Det vil sige, at det historiske afkast er så støjende, at det kunstigt sænker barren for god prædikteringsevne. Dermed undgås denne faldgrubbe ved at benchmarke statistikken mod en naiv prædiktering på nul, ligesom hos Gu et al. (2020).

Som i Drobetz & Otto (2021) beregnes også den prædiktive out-of-sample hældning. Denne giver mulighed for at vurdere nøjagtigheden af prædikteringerne og er afledt af en samlet regression af det realiserede merafkast på den tilsvarende prædiktering, hvor data på tværs af i og t samles for at opnå et enkelt estimat af forholdet mellem variablene:

$$PS_{OOS} = \frac{Cov(r_{i,t+1}, \hat{r}_{i,t+1})}{Var(\hat{r}_{i,t+1})} \tag{4.22}$$

PS_{OOS} måler graden (nøjagtigheden), hvormed de prædikterede værdier stemmer overens med de faktiske. Det vil sige, hvor godt det prædikterede merafkast fanger variationen og retningen af det realiserede afkast. I modsætning hertil kvantificerer R^2_{OOS} andelen af variansen i de observerede resultater, der kan forklares med prædikteringerne. Den måler algoritmens evne til at fange og redegøre for de underliggende faktorer, der driver de observerede data. Ved at fokusere på forskellige aspekter giver de to teststatistikker komplementære perspektiver, der forbedrer evalueringen af algoritmernes ydeevne. Statistikkerne er beregnet for hver algoritme og tre delprøver; hele stikprøven og to delprøver, der kun indeholder store eller små selskaber (baseret på markedsværdien).

For at kunne foretage parvise sammenligninger af algoritmerne med henblik på at teste, hvorvidt forskelle out-of-sample er statistisk signifikante, er Diebold & Mariano (1995) testen anvendt. Den statistiske tests nulhypotese er, at der; ikke er forskel i nøjagtigheden af to konkurrerende prædikteringer. Betingelsen for Diebold-Mariano testen er, at prædikteringsfejlene kun er svagt afhængige. Resonerende Gu et al. (2020), er det dog usandsynligt, at denne betingelse er opfyldt, grundet en potentiel stærk afhængighed i tværsnittet. Derfor følger denne afhandling Gu et al. (2020) og justerer testen til en tværsnitsindstilling ved at sammenligne tværsnitsgennemsnittet af prædikteringsfejlene snarere end de individuelle prædikteringsfejl på aktieniveau. Statistikken for algoritme (1) versus (2) vil være $DM_{12} = \bar{d}_{12}/\hat{\sigma}_{\bar{d}_{12}}$ med:

$$d_{12,t+1} = \frac{1}{n_{3,t+1}} \sum_{i=1}^{n_{3,t+1}} ((\hat{e}_{i,t+1}^{(1)})^2 - (\hat{e}_{i,t+1}^{(2)})^2) \tag{4.23}$$

, hvor $\hat{e}_{i,t+1}^{(1)}$ og $\hat{e}_{i,t+1}^{(2)}$ angiver prædikteringsfejlen for aktieafkast i på tidspunktet t ved brug af hver algoritme, og $n_{3,t+1}$ er antallet af aktier i testsættet (for $t+1$). Derefter angiver $\bar{d}_{12}$ og $\hat{\sigma}_{\bar{d}_{12}}$ middelværdien og Newey-West standardfejlen på $d_{12,t}$ over testsættet. Denne modificerede Diebold-Mariano-teststatistik, som nu er baseret på en enkelt tidsserie $d_{12,t+1}$ af fejlforskelle med ringe autokorrelation, er mere tilbøjelige til at opfylde de milde regularitetsbetingelser, der er nødvendige for asymptotisk normalitet og til gengæld give passende p-værdier til testen.

4.6 Porteføljesammensætning

Det endelige mål er at udnytte prædikteringerne omkring aktiernes merafkast til at konstruere profitable porteføljer. Det er vigtigt at bemærke, at testene beskrevet i afsnittet *Evaluering* (eller 4.5) er af statistisk karakter, men Leitch & Tanner (1991) antyder, at der muligvis kun er en svag sammenhæng mellem statistiske mål og økonomisk rentabilitet. Derfor er det nyttigt at forstå, hvorvidt forskelle i den statistiske prædikteringsevne omsættes til forskelle i prædikteringsevne fra det økonomiske perspektiv af en (realistisk) handelsstrategi. Ved at følge standardtilgangen i litteraturen, er aktierne sorteret i kvantilporteføljer baseret på de respektive merafkastestimater, sorteret i fem grupper. Kvantil 5 er aktierne estimeret til at have det største merafkast fra t til $t+1$, mens kvantil 1 er estimeret til have det mindste merafkast fra t til $t+1$. For samtlige modeller og kvantilporteføljer beregnes det lige- og værdivægtede gennemsnit af det prædikterede og realiserede merafkast. ML-signalerne bør ikke kun indeholde information om afkastprædikteringen i ligevægtede porteføljer, som mindre aktier måtte drive, men også i værdivægtede porteføljer, som på den anden side er domineret af større aktier. Slutteligt sammensættes en po-

rteføljе (H-L), der går lang i den højeste kvantilporteføljе og kort i den laveste. Ved at censurere den midterste del af kvantilerne, ekskluderes aktierne med den største (forventede) usikkerhed fra handel. For alle porteføljer sker omfordeling og rebalancering i slutningen af måneden.

4.6.1 Risikojusterede afkast (alfa)

Da højere gennemsnitlige afkast kan være forårsaget af højere systematisk risiko, er det nødvendigt at undersøge den risikojusterede præstation. For at benchmarke resultaterne af de forskellige ML-porteføljer, anvendes Fama & French (2015) femfaktormodel udvidet med Carhart (1997)'s momentumfaktor, som ligeledes er den mest anvendte benchmark i litteraturen, og dermed sikres sammenlignelighed. De tilsvarende seks faktorer er derfor marked (RMRF), størrelse (SMB), værdi (HML), rentabilitet (RMW), investering (CMA) og momentum (WML)[20]:

$$E[r_{p,t,ML} - r_{f,t}] = \alpha + \beta_1 \cdot E[RMRF_t] + \beta_2 \cdot E[SMB_t] + \beta_3 \cdot E[HML_t] + \beta_4 \cdot E[RWM_t] + \beta_5 \cdot E[CMA_t] + \beta_6 \cdot E[WML_t] + \varepsilon_t \quad (4.24)$$

α_i er skæringspunktet og parameteren af interesse, da denne vil angive, hvorvidt porteføljen fører til merafkast efter kontrol for risikofaktorerne. Hvis modellen fanger al variation i forventede afkast, vil α_i være nul. Hvis α_i er positiv efter kontrol for risikofaktorerne, er det en indikation af, at porteføljen kan generere (risikojusteret) merafkast (Fama & French 2015). Det vil sige merafkast ud over, hvad der forventes baseret på dens systematiske risiko, mens en negativ α_i tyder på en underpræstation. Dermed er regressionen anvendt til at evaluere investeringsporteføljens ydeevne, og algoritmernes evne til at identificere aktier, der er forkert prissat på markedet.

[20] Data på disse seks faktorer er indhentet fra databiblioteket på Kenneth R. Frenchs webside. Da data ikke er tilgængelige for individuelle lande, vil dataene *Fama French European 5 Factor* og *European Momentum Factor* blive anvendt. Desuden er faktorerne kort beskrevet i bilag G.2.

Kapitel 5

Empiriske resultater

Dette kapitel præsenterer de empiriske resultater af de anvendte algoritmer på det skandinaviske aktiemarked. Afsnittet in-sample test undersøger vigtigheden af de anvendte variable. Efterfølgende evaluerer afsnittet out-of-sample test den prædiktive præstation af algoritmerne fra både statistiske og økonomiske perspektiver med et hovedfokus på de forskelige algoritmers evne til at identificere afkast og allokering til en rentabel H-L porteføljе. Kapitlet afsluttes med et sæt robusthedstests. Resultaterne er løbende sammenlignet med den øvrige del af litteraturen.

5.1 In-Sample test

5.1.1 Variables betydning

I dette afsnit undersøges det, om specifikke karakteristika er vigtigere end andre for at prædiktere fremtidige merafkast. Variables betydning er et vigtigt aspekt af aktivprissætning, idet en betydelig del af litteraturen omkring kapitalaktivers prisfastsættelse er dedikeret til at identificere variable, der besidder forudsigelig kraft for aktivafkast (Feng et al. 2018, Cochrane 2011). Målet med analysen er imidlertid mere begrænset, som den kun vil identificere variable med en betydelig effekt, hvor interaktionseffekter, ikke-lineariter og overflødige variable ikke er diskuteret. Bilag E illustrerer vigtigheden af top-20 karakteristika for algoritmerne. Værdierne inden for algoritmerne er normaliseret til at summere til én, hvormed fortolkningen er af den relative betydning for den pågældende algoritme. Ifølge bilag E er det fundet, at algoritmerne i gennemsnit betegner lignende variable som de mest betydningsfulde. De kan grupperes til momentum (PR), handelsfriktioner (H) samt value (V), hvilket også er tilfældet hos Gu et al. (2020) og Hanauer & Kalsbach (2023), men også de inkluderede sentimentvariable er meget betydningsfulde, og

er det særligt hos ML-algoritmerne. Det skyldes, at særligt RF og GB trækker information fra meget lignende variable. Blandt top-20 er følgende karakteristika gennemgående i begge af algoritmerne: RECSELL, P2P52WH, DY, BEME, BEME_m, EPS1MD, EPS2MD, EPS3MD, LTurnover samt A2ME, hvor fire af disse er sentimentbaserede, tre er valuebaserede og to er handelsfriktionsbaserede. Vigtigheden af P2P52WH er også bekræftet af Hanauer & Kalsbach (2023).

OLS og GB er fokuseret på et mere koncentreret sæt af variable, da de fleste af deres relative vægte er på to variable (de to mest betydningsfulde variable har således væsentlig indflydelse på prædikteringerne). Hvorimod RF anser et meget større udvalg af variable for at være vigtige. Det vil sige, at RF trækker dens information mere ensartet fra de forskellige variable sammenlignet med OLS og GB. Men sammenlignet med litteraturen trækker algoritmerne på forholdsvis mange variable. De er ikke meget skævt mod enkelte variable men er mere demokratiske og trækker (prædiktiv) information fra et bredere sæt. Mange regnskabsvariable er ikke tilgængelige ved den månedlige frekvens, hvilket formodentligt kan forklare deres mere begrænsede betydning.

For RF er top-3 variable: LME, DY og RECSELL, men også særligt de momentumbaserede variable (PR) er relevante for denne, sammenlignet med de andre algoritmer. Den relative betydning af LME, som er den mest indflydelsesrige, er omkring 10 pct., hvilket er væsentligt lavere end de mest betydningsfulde i OLS og GB. Handelsfriktionsvariable er imidlertid den mest repræsenterede kategori (med fem variable) men hverken variable fra kategorierne investering (Inv) eller immaterielle aktiver (I) er blandt top-20 for RF. For GB er top-3 variable: RECHOLD, RECSELL og P2P52WH, hvor særligt analytikers (gennemsnitlige) anbefaling for henholdsvis at holde/sælge er udgørende for omkring 40 pct. af den relative betydning. Denne betydning er også afdækket i Singh & Khushi (2021), hvor de viste sig at være de bedste bidragsydere i deres model. GB har som den eneste algoritme variable fra samtlige af kategoriseringerne i top-20. For OLS-modellen er de to vigtigste variable BEME og BEME_m, og den er i høj grad afhængig af disse. Deres samlede relative betydning er udgørende for mere end 40 pct. Dertil er variable som P2P52WH, A2ME, LTurnover, EPS1MD og EPS2MD igen blandt top-20 karakteristika, men foruden disse har rentabilitetsvariable (såsom ROE, RNA, PCM, Prof samt ROA) en relativ større betydning i OLS-modellen sammenlignet med ML-algoritmerne. Hvorimod de sentimentbaserede variable er mindre betydningsfulde men stadig relevante. Denne sammenhæng er også bekræftet af Daul et al. (2022), der finder, at disse har større betydning hos ML-algoritmerne. OLS-modellen har variable fra alle kategoriseringer undtagen Inv. Dermed trækker den fortsat bred information, selvom den er højt koncentreret omkring BEME og BEME_m.

Generelt er billedet fra bilag E i fin overensstemmelse med litteraturen. Kelly et al. (2019), der

analyserer samme sæt af variable, foruden de sentimentbaserede, er forholdsvis enig med flere af resultaterne i denne afhandling, hvilket er bekræftende for afhandlingens validitet. På trods af dette, er det væsentligt at nævne, at litteraturen fortsat ikke har opnået konsensus om vigtigheden og forudsigelseskraften af de forskellige variable, den såkaldte faktor-zoo. Dertil kommer, at Drobetz & Otto (2021) viser, at den relative betydning ændres med tiden.

5.2 Out-of-sample test

5.2.1 Prædikteringsevne

Dette afsnit vil præsentere resultaterne af evalueringsmetrikkerne forklaret i kapitel 4.5: PS_{OOS}, R^2_{OOS} og Diebold-Mariano-statistikken.

Panel A i tabel 5.1 præsenterer tidsseriegennemsnittene af den prædiktive PS_{OOS} beregnet ud fra testsættet på reestimeringsdatoen. Hvor $PS_{OOS} > 1$ indikerer alt for snævre prædikteringer sammenlignet med den sande tværsnitsvariation, indikerer $PS_{OOS} < 1$ alt for brede prædikteringer. Det betyder, at des større denne er over 1, des større tegn er det på, at algoritmerne er for følsomme over for støj (og idiosynkrasier), der hindrer deres evne til at generalisere godt til nye, usete data[1]. PS_{OOS} for OLS (ved det fulde sæt) er tæt på 1 (0,96), hvormed prædikteringernes spredning primært er afspejlet i tværsnitsvariation parallelt med det sande merafkast. I tilfælde af ML-algoritmerne er PS_{OOS}-værdierne en smule større end 1, specifikt 1,17 og 1,09 for henholdsvis RF og GB. Selvom forskellen mellem ML-algoritmerne ikke er signifikant, er der tegn på, at de kan være en anelse mere følsomme over for støj (og idiosynkrasier) sammenlignet med OLS men bemærk, at det er en forholdsvis lille afvigelse fra den ideelle værdi på 1.

For det markedsværdi-sorterede sæt, har RF de største vanskeligheder med de største afvigelser i PS_{OOS} fra 1, mens GB har stort set identiske PS_{OOS}-værdier for de to sæt (begge en anelse større end 1). OLS har en tendens til at generere lidt for alt smalle prædikteringer for de mindste selskabers afkast og lidt for brede for de største. For at opsummere understreger disse resultater vigtigheden af at vurdere algoritmernes ydeevne på tværs af forskellige aktiestørrelser for at sikre nøjagtige prædikteringer. Overordnet er de prædikterede afkast positivt relateret til de faktisk realiserede afkast, og algoritmerne har dermed prædikteringsevne, omend ikke perfekt.

Panel B i tabel 5.1 præsenterer tidsseriegennemsnittene af den prædiktive R^2_{OOS} beregnet ud fra

[1] Mens $PS_{OOS} > 1$ ikke direkte indebærer en overtilpasning, kan det være en indikator for en potentiel overtilpasning, hvis forudsigelserne er overdrevent påvirket af de specifikke karakteristika (eller støj) i dataene. Hvor $PS_{OOS} < 1$ foreslår en begrænsning i modellens evne til at fange kompleksiteten og variationen. Det kan tyde på en undertilpasning, hvor modellen ikke er fleksibel nok til at fange de underliggende mønstre og dynamikker.

Tabel 5.1 Prædiktiv PS_{OOS} og prædiktiv R^2_{OOS}

	OLS	RF	GB
Panel A: Prædiktiv PS_{OOS}			
Fulde datasæt	0,9571	1,1731	1,0953
Store selskaber	1,1158	0,7875	1,1278
Mindre selskaber	0,8605	1,3172	1,0840
Panel B: Prædiktiv R^2_{OOS} [%]			
Fulde datasæt	3,060	2,115	2,357
Store selskaber	2,563	1,103	1,135
Mindre selskaber	3,275	1,843	2,881

Denne tabel præsenterer evalueringsmetrikkerne introduceret i kapitel 4.5. Panel A rapporterer out-of-sample PS_{OOS}, hvor Panel B rapporterer out-of-sample R^2_{OOS}. De er præsenteret for tre delprøver; den fulde og to delprøver, der kun indeholder store eller små selskaber (baseret på markedsværdien).

testsættet på reestimeringsdatoen. I betragtning af det fulde datasæt er resultaterne i overensstemmelse med ovenstående resultat for PS_{OOS}. OLS producerer en R^2_{OOS} på 3,06 pct., hvilket indikerer, at den i modsætning til litteraturen (se bl.a. Gu et al. (2020) samt Drobetz & Otto (2021)), som typisk har rapporteret en negativ R^2_{OOS} for OLS, ikke er domineret af at anvende en naiv prædiktering (nul-prædiktering). Det betyder, at OLS forklarer 3,06 pct. af variationen i (tværsnits-) afkastene, som OLS'en fanger visse underliggende mønstre og relationer i dataene. Det opnåede resultat til sammenligning med litteraturen er ikke overraskende, fordi fraværet af regularisering efterlader OLS meget modtagelig for in-sample overtilpasning, og som litteraturen ofte har anvendt et (meget) stort sæt af variable, er OLS ineffektiv. Mere bemærkelsesværdigt er det, at hverken RF eller GB klarer sig bedre end OLS i forhold til R^2_{OOS}-statistikken[2]. RF og GB opnår R^2_{OOS}-statistikker på henholdsvis 2,12 pct. og 2,36 pct., hvor GB opnår den næsthøjeste værdi efter OLS. For de markedsværdi-sorterede sæt, opnår samtlige tre algoritmer de højeste R^2_{OOS}-værdier for sættet bestående af de mindste selskaber. Resultatet indikerer, at algoritmerne opfanger unikke faktorer og dynamikker forbundet med mindre selskaber. Algoritmernes evne til at opfange disse faktorer øger deres forklaringsevne, hvilket resulterer i en mere præcis repræsentation af små selskabers afkast. Omvendt tyder den lidt lavere forklaringsevne, der er observeret for de større selskaber, på tilstedeværelsen af yderligere faktorer eller kompleksiteter, der ikke fuldt ud fanges af algoritmerne i forhold til denne gruppe. Ikke desto mindre, selv for større selskaber, forklarer algoritmerne stadig en andel af variationen.

[2]Bekræftende herfor finder De Nard et al. (2022) dog, at OLS ligeledes klarer sig rimeligt i deres opsætning med kun 94 karakteristika. De konkluderer, at så længe dimensionen ikke er for stor til OLS, giver den simple model en bemærkelsesværdig aggregeret out-of-sample forudsigelighed, der overgår RF.

R^2_{OOS}-statistikken for RF og GB er på trods af, at disse ikke formår at overgå OLS, højere end de rapporterede i litteraturen. Drobetz & Otto (2021) finder R^2_{OOS}-statistikker på 1,18 pct. og 1,14 pct. for henholdsvis RF og GB på det europæiske aktiemarked, hvor Gu et al. (2020) finder disse til henholdsvis 0,33 pct. og 0,34 pct. på det amerikanske aktiemarked. Sidstnævnte finder endvidere, at de træbaserede algoritmer er særligt succesfulde blandt store aktier (R^2_{OOS} mellem 0,52 pct. og 0,63 pct.). Sammenligning af størrelsen på R^2_{OOS} mellem undersøgelser kan give indsigt i den relative forklaringsevne af forskellige modeller, men det er vigtigt at bemærke, at fortolkningen og sammenligningen på tværs kan påvirkes af forskellige faktorer, herunder det specifikke datasæt, tidsperiode, metodologi (modelspecifikation) og anvendte variable. Men alle algoritmer genererer positive R^2_{OOS}-statistikker, der udkonkurrerer den naive prædiktering.

Samlet set stemmer forklaringerne for værdierne af PS_{OOS} og R^2_{OOS} overens med ideen om, at individuelle algoritmer har forskellige styrker og begrænsninger i forhold til at fange specifikke mønstre og faktorer i dataene. Denne heterogenitet bidrager til forskellene i deres prædikterings- og forklaringsevne og tildeler således yderligere støtte til konstruktionen af $\text{ENS}^{(1)}$ og $\text{ENS}^{(2)}$.

Hvor tabel 5.1 tilbyder en kvantitativ sammenligning af algoritmernes prædiktive præstation, vurderer tabel 5.2 den statistiske signifikans af forskelle mellem dem. Tabellen rapporterer den modificerede Diebold-Mariano-statistik, justeret med Newy-West standardfejlene (med fire lags for at tage højde for mulig heteroskedasticitet samt autokorrelation), for den parvise sammenligning af kolonnemodellen versus rækkemodellen med nulhypotesen om ingen (out-of-sample) forskel mellem dem. En positiv Diebold-Mariano-statistik indikerer, at kolonnemodellen klarer sig bedre end rækkemodellen - og omvendt. Det observeres, at resultaterne, jf. tabel 5.2, understreger tendenserne fundet indtil videre. Den første Diebold-Mariano statistik er -3,963, hvilket indikerer, at OLS har signifikant lavere prædikteringsfejl sammenlignet med RF, og dermed afvises nulhypotesen, og det konkluderes, at der er stærk evidens for, at OLS klarer sig bedre end RF. Dette er ligeledes tilfældet for sammenligningen mellem OLS og GB (-2,260), som endog er mindre signifikant og ikke signifikant ved 1 pct.-niveauet. Blandt ML-algoritmerne fremstår GB som den bedste præsterende, og forskellen er endda statistisk signifikant. Dette er i overensstemmelse med resultaterne i Gu et al. (2020), men i modstrid med resultaterne i Drobetz & Otto (2021), hvor RF er den bedst præsterende. I de to undersøgelser er forskellen mellem dem dog ikke signifikant. Sammenfattende bekræfter resultaterne fra tabel 5.2 resultaterne fra tidligere analyse, hvilket forstærker de tendenser, der er observeret i algoritmernes prædiktive ydeevne. OLS skiller sig ud med statistisk signifikant overlegenhed over ML-algoritmerne (GB, såvel som RF), i kontrast til litteraturen, men kun lige akkurat signifikant bedre end GB.

Tabel 5.2 Diebold-Mariano statistik for sammenligning

	OLS	RF	GB
OLS		**-3,9634** (0,0000)	**-2,2604** (0,0237)
RF			**2,6801** (0,0073)
GB			

Denne tabel præsenterer Diebold-Mariano testen introduceret i kapitel 4.5 for den fulde stikprøve. Parentesværdierne er p-værdierne, mens en **fremhævet** statistisk er signifikant ved 5 pct.-niveauet.

5.2.2 Porteføljepræstation

I betragtning af, at de evaluerede algoritmer viser signifikante forskelle i out-of-sample prædiktiv nøjagtighed, har dette afsnit til formål at undersøge, hvorvidt disse observerede statistiske forskelle, har praktiske implikationer med hensyn til økonomisk rentabilitet i porteføljepræstationen (via porteføljen præsenteret i kapitel 4.6). Først præsenteres kvantilporteføljerne for både det lige- og værdivægtede gennemsnit af prædikterede og realiserede merafkast. Baseret på det realiserede merafkast beregnes standardafvigelsen og t-statistikken, der tester nulhypotesen om, at det realiserede merafkast er nul, og alt dette er at finde i tabel 5.3. Tabellen viser et mærkbart mønster på tværs af kvantilerne, hvor kvantil 1 udviser et negativt merafkast, udviser kvantil 5 et positivt afkast, mens de mellemliggende kvantiler viser en gradvis stigning i afkastet svarende til deres respektive kvantiltal. Resultaterne viser algoritmernes evne til at skelne mellem aktier baseret på deres forventede afkast, hvilket er underbyggende for deres prædikteringsevne med hensyn til at adskille aktier med overlegne afkastudsigter fra aktier med ringere udsigter.

Generelt er mønstrene ens for de lige- og værdivægtede kvantilporteføljer, selvom de er væsentligt mere udtalte for de ligevægtede (både i form af højere afkast og t-statistik). Det gennemsnitlige realiserede merafkast er på linje med det gennemsnitlige forventede merafkast, hvilket resulterer i positive H-L-spænd, der er statistisk signifikante og økonomisk store (indikerer økonomisk rentabilitet). En tendens er for de realiserede afkast at være en anelse større end de prædikterede. H-L-afkastene er i gennemsnit 25 pct. lavere for de værdivægtede kvantilporteføljer, og dermed er en væsentlig del af merafkastet drevet af en (relativt) bedre præstation blandt aktierne med de lavere markedsværdier sammenlignet med de større. Resultaterne fremhæver endog, at alle præsenterede tilgange fanger tværsnitsvariationer i realiserede merafkast, men potentialet for at levere handelssignaler er (væsentligt) forskelligt på tværs af dem.

Ved en sammenligning af de ligevægtede H-L-afkast, som tenderer til at blive drevet mere af den lange side (kvantil 5) end den korte side (kvantil 1), indikerer værdierne, at forskelle i

Tabel 5.3 (Gennemsnitlige) merafkast af forventede kvantilsorterede porteføljer

	Ligevægtet				Værdivægtet			
	Pred [%]	Real [%]	Std [%]	t-stat	Pred [%]	Real [%]	Std [%]	t-stat
OLS								
Low(L)	-1,07	-0,54	5,62	-1,49	-1,05	-0,20	5,74	-0,54
2	0,10	0,49	5,14	1,5	0,098	0,36	5,25	1,07
3	0,78	1,25	5,13	3,79	0,77	0,88	5,19	2,63
4	1,49	2,08	5,02	6,44	1,47	1,96	5,03	6,04
High(H)	2,85	3,47	5,48	9,82	2,59	2,89	5,33	8,42
H-L	3,92	4,01	3,68	16,93	3,64	3,10	4,72	10,17
RF								
Low(L)	-0,005	-0,002	5,74	-0,005	0,001	0,25	5,62	0,67
2	0,40	0,86	5,26	2,53	0,39	0,39	5,16	1,18
3	0,60	1,28	5,02	3,97	0,60	1,02	4,71	3,33
4	0,81	1,81	5,08	5,29	0,81	1,51	4,61	5,07
High(H)	1,85	2,81	5,16	8,45	1,57	2,07	5,58	5,0
H-L	1,855	2,82	3,0	14,51	1,569	1,83	4,3	6,58
GB								
Low(L)	-0,62	-0,47	5,58	-1,30	-0,54	-0,34	5,39	-0,97
2	0,17	0,67	5,17	2,0	0,17	0,45	5,19	1,37
3	0,58	1,47	5,18	4,41	0,58	1,19	5,03	3,67
4	1,06	1,78	5,31	5,18	1,04	1,47	4,86	4,69
High(H)	2,43	3,32	5,03	10,25	2,14	2,55	5,37	7,34
H-L	3,05	3,79	3,34	17,61	2,68	2,89	4,28	10,46
ENS$^{(1)}$								
Low(L)	-0,24	-0,45	5,58	-1,27	-0,18	-0,23	5,53	-0,64
2	0,31	0,79	5,39	2,02	0,31	0,39	5,16	1,16
3	0,59	1,31	5,11	3,99	0,56	1,17	4,85	3,74
4	0,91	1,99	5,03	6,11	0,91	1,59	4,65	5,30
High(H)	2,07	3,23	5,15	9,71	1,79	2,53	5,33	7,36
H-L	2,31	3,68	3,18	17,95	1,97	2,76	4,28	10,02

	Ligevægtet				Værdivægtet			
	Pred [%]	Real [%]	Std [%]	t-stat	Pred [%]	Real [%]	Std [%]	t-stat
ENS$^{(2)}$								
Low(L)	-0,42	-0,68	5,66	-1,89	-0,37	-0,37	5,65	-1,02
2	0,26	0,53	5,23	1,56	0,26	0,36	5,36	1,04
3	0,67	1,34	5,17	4,02	0,67	1,16	5,03	3,56
4	1,1	2,07	5,04	6,36	1,10	1,87	4,72	6,14
High(H)	2,2	3,53	5,22	10,47	1,95	2,85	5,46	8,10
H-L	2,62	4,21	3,44	18,99	2,32	3,23	4,64	10,76

Note: T-statistikken, der tester nulhypotesen om, at realiserede merafkast er nul, er beregnet ved hjælp af de månedlige estimater, der inkorporerer en Newy-West korrektion med fire lags for at tage højde for mulig heteroskedasticitet og autokorrelation i de realiserede merafkast.

statistisk prædiktiv præstation omsættes til forskelle i økonomisk rentabilitet. De realiserede merafkast af kvantilporteføljen baseret på OLS-modellen er de mest lovende, da de spænder fra -0,54 pct. til 3,47 pct. per måned, hvilket resulterer i et månedligt H-L-afkast på 4,01 pct. (før transaktionsomkostninger), hvor den tilhørende t-statistik er signifikant med en værdi på 16,93. Afkastforholdet, det vil sige forholdet mellem realisererede og prædikterede (gennemsnitlige) H-L-afkast, er således på 1,02 (= 4,01 pct. / 3,92 pct.) og derfor stort set identisk med tidsseriegennemsnittet af den prædiktive PS_{OOS} (se tabel 5.1). Det realiserede afkast er lavest for RF med et H-L-afkast på 2,82 pct., mens GB kan formå at skabe et gennemsnitligt afkast på 3,79 pct. Begge H-L-afkast er statistisk forskellige fra nul og har dermed økonomisk relevans. RF formår i praksis ikke at fange aktierne i den første kvantil særlig godt, som den har et minimalt negativt afkast (statistisk ikke forskelligt fra nul[3]). Dette er forklarende for dens underpræstation sammenlignet med GB, da det realiserede merafkast for kvantil 5 er mere sammenligneligt i størrelsesorden. Resultaterne fremhæver dog OLS-modellens overlegne prædiktive nøjagtighed, som er forbundet med højere gennemsnitlige merafkast sammenlignet med ML-algoritmerne. Derudover overgår det merafkast, der genereres af OLS-modellen, markant dem, der er dokumenteret i litteraturen. For eksempel rapporterer Drobetz & Otto (2021) et gennemsnitligt OLS-afkast på 2,73 pct. baseret på H-L-spændet, mens Gu et al. (2020) finder et endnu lavere OLS-afkast på kun 0,94 pct. I modsætning hertil er de realiserede merafkast for de to ML-algoritmer sammenlignelige i størrelsesorden med litteraturen og overgår endda dem, der er rapporteret i undersøgelser som Gu et al. (2020), Tobek & Hronec (2021) og Hanauer & Kalsbach (2023). Det er dog vigtigt at bemærke, at sammenligning af disse merafkast på tværs

[3]Dette er imidlertid ikke tilfældet for nogle af porteføljerne i kvantil 1.

af forskellige undersøgelser er udfordret grundet variation i anvendte markeder og tidsperioder (hvor denne afhandling ikke har en lige så lang historisk periode med). Denne afhandling har desuden anvendt afkastindekset fra Refinitiv Datastream, hvilket formodentligt er årsagen til det højere (gennemsnitlige) merafkast. For ved sammenligning af gennemsnittet af de anvendte selskabers månedlig afkast er det tydeligt, at disse er højere i denne afhandling, end de i litteraturen rapporterede. På trods af den overlegne præstation af OLS observeret i denne afhandling, bør ML-algoritmernes præstation ikke nødvendigvis anses for at være ringere end den, der er rapporteret i litteraturen.

$ENS^{(1)}$ klarer sig ikke bedre end den bedst præsterende ML-algoritme (GB) med hensyn til H-L-afkast, hvilket tyder på, at ensemblestrategien ikke giver tilsyneladende fordele, bortset fra bekvemmeligheden ved at undgå algoritmevalg og samtidig opretholde et gunstigt H-L-afkast. Fraværet af en forbedret ydeevne kan tilskrives, at GB allerede demonstrerer en mere effektiv indkapsling af de underliggende mønstre, der er iboende i dataene. Ensemblets overordnede præstation er i sidste ende begrænset af kvaliteten og effektiviteten af algoritmerne. Desuden kan det stamme fra fraværet af komplementær information, som algoritmerne fanger de samme facetter (i afsnit 5.1.1 blev det illustreret, at de er afhængige af lignende variable). Slutteligt kan det være kombinationsmetoden, der ikke er ideelt egnet til algoritmernes og dataenes karakteristika. Ikke desto mindre afslører $ENS^{(2)}$, at ML-algoritmerne besidder komplementær information til den OLS-genererede portefølje således, at den faktisk kan overgå OLS-modellen i form af merafkast. Det vil sige, at inkorporeringen af ML-algoritmerne med den lineære i en ensemblestrategi kan give øget værdi.

Det bemærkes også, at H-L-afkastene har betydelig lavere spredning, særligt for den ligevægtede tilgang, sammenlignet med kvantilerne selv. Det skyldes, at porteføljekonstruktionen, selvom det ikke er dens primære fokus, introducerer en vis grad af variansreduktion. H-L-elementet muliggør potentielle gevinster i både opadgående og nedadgående markedsbevægelser. Dermed er strategien potentielt mindre følsom over for markedsudsving. Selvom fokus er på at udvælge aktier inden for specifikke kvantiler baseret på prædikteringer, kan opbygningen af en H-L-portefølje stadig inkorporere diversificering inden for kvantilerne. Følgelig kan porteføljen omfatte flere aktier med adskilte risikoprofiler, hvilket yderligere reducerer den samlede porteføljevarians. Det er dog vigtigt at erkende, at det primære formål med en sådan strategi er at fange relative præstationsforskelle og generere positive afkast, snarere end eksplicit at målrette variansreduktion.

Nøgletal på porteføljepræstationen

Panel A og C i tabel 5.4 rapporterer det (gennemsnitlige) merafkast for H-L-spændet (Gns.), annualiseret merafkast (Gns. Ann), standardafvigelse (Std. Ann) samt Sharpe-Ratio (SR Ann), men også maksimale drawdown (Max DD), det mest negative månedlige afkast (Max 1M Tab) samt omsætning (TO) for det værdivægtede H-L-afkast[4], men også for det lange element (kvantil 5), på grund af udbredte short-selling restriktioner, der kan forhindre udførelsen men også rentabiliteten af det korte element (kvantil 1). Max DD er defineret som det største procentvise fald observeret i en portefølje fra dets top til dets bund:

$$\text{Max DD} = \max_{0 \leq t_1 \leq t_2 \leq T} (Y_{t_1} - Y_{t_2}) \tag{5.1}$$

, hvor Y_t er det kumulative log afkast fra dato 0 til og med t.

Omsætningshastigheden er et mål til at kvantificere niveauet af handelsaktivitet i en portefølje, som afspejler i hvilket omfang porteføljeforvalteren har været nødt til at købe og sælge for at opretholde den ønskede porteføljesammensætning:

$$\text{TO} = \frac{1}{T} \sum_{t=1}^{T} \Big(\sum_{i} \mid w_{i,t+1} - \frac{w_{i,t}(1 + r_{i,t+1})}{1 + \sum_j w_{j,t} r_{j,t+1}} \mid \Big) \tag{5.2}$$

, hvor $w_{i,t}$ er vægten af aktie i på tidspunkt t i porteføljen. Formlen er en vægtet sum af de absolutte forskelle mellem de aktuelle vægte af aktierne i porteføljen ($w_{i,t}$) og de ønskede vægte af aktierne i porteføljen, givet de nuværende og forventede fremtidige afkast ($\frac{w_{i,t}(1+r_{i,t+1})}{1+\sum_j w_{j,t} r_{j,t+1}}$). Vægtene er beregnet ud fra aktiernes forventede afkast i den næste periode. Omsætningen beregnes dernæst som gennemsnittet af disse vægtede forskelle på tværs af tidsperioderne og er dermed den gennemsnitlige procentvise ændring i beholdningerne fra porteføljerne. Denne er medtaget, fordi højere omsætningshastigheder forbundet med investeringsstrategier let kan opveje fordelene i form af de rå afkast. Derfor undersøger afhandlingen, om OLS-modellens overlegne præstation sammenlignet med ML-algoritmernes kan tilskrives væsentligt højere omsætningshastigheder. Det er dog vigtigt at erkende, at opnåelse af realistiske transaktionsomkostningsmodeller giver udfordringer, fordi sådanne data ikke er let tilgængelige eller praktiske at anvende (Lesmond et al. 1999). Desuden er sande transaktionsomkostninger i sagens natur uobserverbare. Som et resultat heraf vil ethvert forsøg på at estimere disse nøjagtigt være fejlbehæftet eller forkert.

[4]Dette er valgt, da denne er mest repræsentativ for et realistisk opnået afkast på strategien, da det således er drevet mindre af de mindste aktiers afkast, som kan være sværere at hente. Samme tabel for den ligevægtede er dog at finde i bilag F tabel F.1.

Tabel 5.4 Værdivægtet drawdowns, omsætning og risikojusteret præstation af porteføljerne

	OLS	RF	GB	ENS$^{(1)}$	ENS$^{(2)}$
Panel A: Kvantilsorterede porteføljeperformance					
Gns. [%]	3,1	1,83	2,89	2,77	3,23
Gns. Ann. [%]	44,24	24,29	40,70	38,73	46,37
Std. Ann. [%]	16,34	14,89	14,81	14,81	16,08
SR Ann.	2,27	1,47	2,34	2,24	2,41
Max DD [%]	-20,26	-29,26	-11,81	-18,51	-16,37
Max 1M Tab [%]	-13,06	-17,61	-10,07	-11,79	-10,69
TO	158,6	123,96	172,81	158,99	156,51
Panel B: Risikojusteret performance					
α_{FF6} [%]	2,73	1,28	2,65	2,43	2,77
t-stat$_\alpha$	8,72	4,39	8,85	8,30	8,92
R^2 [%]	15,53	12,21	7,59	11,0	14,42
Panel C: Kvantil 5 porteføljeperformance					
Gns. [%]	2,89	2,07	2,55	2,53	2,85
Gns. Ann. [%]	40,88	27,92	35,20	35,02	40,16
Std. Ann. [%]	18,47	20,23	18,60	18,46	18,91
SR Ann.	1,88	1,22	1,64	1,65	1,81
Max DD [%]	-33,93	-50,76	-45,65	-45,67	-37,99
Max 1M Tab [%]	-16,89	-26,07	-18,82	-19,15	-18,24
TO	96,19	78,89	97,22	92,79	89,0
Panel D: Risikojusteret performance kvantil 5					
α_{FF6} [%]	2,1	1,13	1,76	1,71	2,03
t-stat$_\alpha$	7,89	4,08	6,53	6,60	7,29
R^2 [%]	51,88	57,63	51,22	54,9	50,35

Note: Tabellen viser den værdivægtede out-of-sample præstation af kvantilporteføljen (H-L) men også kvantil 5. Hvor Panel A og C opsummerer det gennemsnitlige månedlige merafkast (Gns.), annualiseret merafkast, standardafvigelse og Sharpe-Ratio, men også maksimale drawdown (Max DD), det mest negative månedlige tab (Max 1M tab) og gennemsnitlige omsætning (TO). Panel B og D rapporterer det gennemsnitlige alfa (α_{FF6}) fra seksfaktormodellen, tilsvarende Newey-West justeret t-statistik med 4 lags (t-stat$_\alpha$) og den tilsvarende R^2.

Den bedst præsterende er den OLS-genererede porteføljе med et annualiseret merafkast på 44,24 pct. (før transaktionsomkostninger) og en annualiseret volatilitet på 16,34 pct. svarende til en annualiseret Sharpe-Ratio på 2,27, gældende for den værdivægtede H-L-porteføljе. En Sharpe-Ratio på 2,27 anses generelt for at være tilfredsstillende, hvor en højere værdi er en indikation på en bedre risikojusteret præstation. Typisk anses en Sharpe-Ratio > 1 for acceptabelt, mens en ratio over to anses for godt (Hillier et al. 2012). Det betyder, at strategien har genereret et stærkt afkast i forhold til dens risiko, hvilket set fra et investorperspektiv er yderst ønskværdigt, da højere afkast dermed er relativt billigere i forhold til risiko end lave afkast. På trods af et lavere (gennemsnitligt) månedligt merafkast, har GB-algoritmen en marginalt højere Sharpe-Ratio end OLS-modellen for den værdivægtede porteføljе (2,27 vs. 2,34). Den GB-genererede porteføljе har, for den værdivægtede porteføljе, imidlertid det laveste drawdown på 11,81 pct., hvilket næsten er det halve af den OLS-genererede. Dette er også observeret i litteraturen, at velfungerende ML-modeller reducerer nedadgående risiko (Gu et al. 2020). På den anden side udviser den RF-genererede porteføljе den laveste præstation i forhold til annualiseret afkast (24,29 pct.), men genererer stadig en Sharpe-Ratio på 1,47, hvilket er acceptabelt. RF har tilmed også det mest ekstreme 1-måneds tab (-17,61 pct.), næsten det dobbelte af de andre, samt det mest ekstreme drawdown. De opnåede Sharpe-Ratioer er ikke unikke for denne afhandling men er også at finde i Gu et al. (2020), Drobetz & Otto (2021) og Hanauer & Kalsbach (2023).

Resultaterne (og mønstrene) for kvantil 5 porteføljen (Panel C) stemmer (kvalitativt) fint med den værdivægtede, men afkastene og Sharpe-Ratioerne er lavere, dog kun marginalt, og er endda højere for RF, mens drawdown og maksimale én-månedes tab er større. Dette giver intuitivt mening, da H-L-porteføljen er markedsneutral - som kvantil 1 beskytter ved markedsnedgang.

Den overlegne præstation af OLS og GB, sammenlignet med RF, er på bekostning af en højere omsætning. Denne omsætningsstigning er fra 123,96 pct. for RF til værdier mellem 158,6 pct. og 172,81 pct. for OLS og GB. Dette resultat er ikke (kvalitativt) mere forhøjet end i litteraturen (Gu et al. 2020, Hanauer & Kalsbach 2023, Drobetz & Otto 2021) og stemmer med, at en bedre præstation ofte opnås med forholdsvis høje porteføljeomsætninger. Dette er dog ikke overraskende i lyset af, at det er dokumenteret, at algoritmerne primært er afhængige af de månedlige variable, og derfor kræver handelssignalerne typisk dynamisk porteføljerotation. Som referenceramme er den månedlige omsætning af den kortsigtede vendingseffekt - en anomali der kræver aktiv porteføljeopbygning - 172,6 pct. for dens H-L-spænd (Novy-Marx & Velikov 2016). Omsætningen er mindre for kvantil 5 porteføljen, dermed er det en mindre aktiv handelsstrategi.

Samtlige af de genererede porteføljer er altså meget profitable. Sammenfattende er den stærkeste

evidens (igen) pegende på OLS-modellen, hvis der ses bort fra ENS$^{(2)}$. Panel B og D i tabel 5.4 illustrerer den risikojusterede præstation ved hjælp af faktormodellerne. I den eksisterende litteratur har porteføljerne opnåede meget signifikante alfaer ved anvendelse af femfaktormodellen forstærket med momentum. Selvsamme faktormodel forklarer i denne afhandling omkring 15,53 pct. af det gennemsnitlige merafkast for den OLS-generede H-L-portefølje og mellem 7-12 pct. af variationen for de ML-generede. Dette resulterer i signifikante t-statistikker på tværs af algoritmerne. Hvor det positive alfa indikerer, at porteføljen har klaret sig bedre end det forventede afkast baseret på seksfaktormodellen (des højere alfa, des bedre ydeevne). Resultatet for kvantil 5 i panel D afslører igen en kvalitativ identisk konklusion, men t-statistikkerne er lavere og særligt R^2 er væsentligt forhøjet. Det skyldes primært (signifikante) positive korrelationer til faktorerne RMRF og WML og negative korrelationer til CMA. Det betyder, at dens afkast generelt bevæger sig i samme retning som markedet men med lavere volatilitet eller systematisk risiko (da koefficienten er < 1), og drager fordel af perioder med positivt momentum. Hvor den negative korrelation til CMA indikerer, at porteføljen trækker på selskaber med aggressive investeringsstrategier. Sammenfattende betyder ovenstående, at de opnåede merafkast (H-L-afkast) ikke er et resultat af højere systematisk risiko, som de riskojusterede værdier er meget tæt på deres rå modparter. Seksfaktormodellen har således ringe forklaringsevne (R^2) i forhold til de gennemsnitlige merafkast, særligt for H-L porteføljen. Det er muligt at argumentere for, at afhandlingen burde have konstrueret faktorerne selv ud fra datasættet, men dette har været uden for denne afhandling.

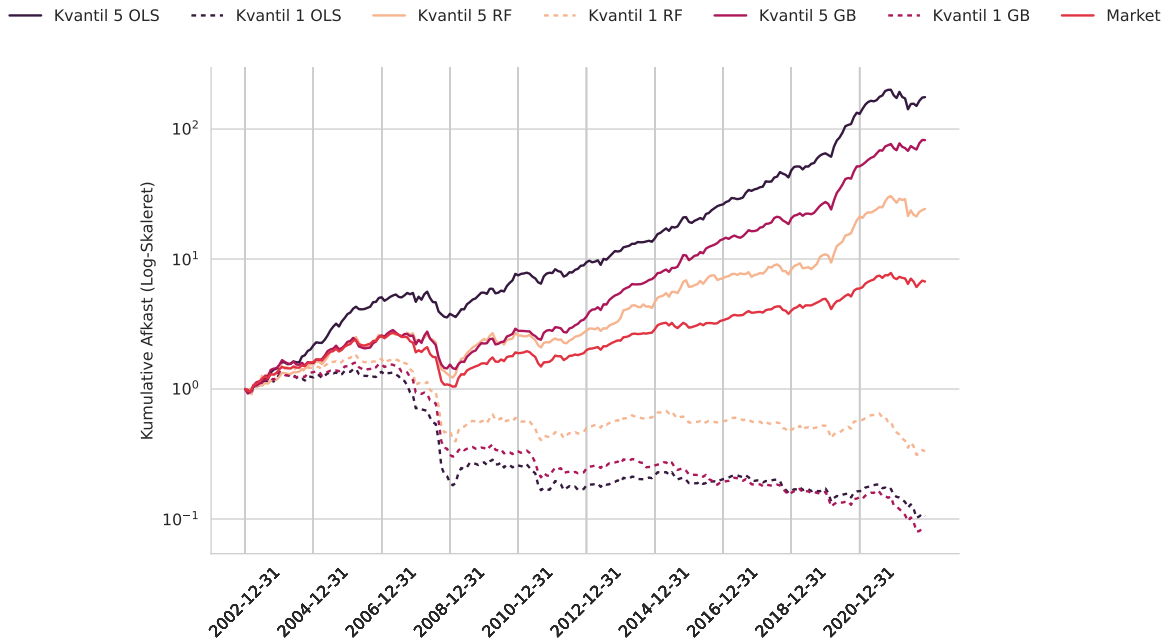
Investeringsporteføljens præstation

På baggrund af beviserne fra de kvantilporteføljer, der er diskuteret i de forrige to afsnit, er det væsentligt at evaluere den faktiske værditilvækst af disse algoritmer under realistiske handelsbetingelser. For at opnå dette analyseres investeringsstrategiens kumulative præstation i forhold til det kumulative markedsmerafkast[5], der fungerer som et benchmark. Kvantil 5 og 1 er endog præsenteret individuelt således, at det er muligt at vurdere det mest værditilvækstende element[6] (over t). Den kumulative præstation er præsenteret for den værdivægtede orden med en initial investering på 1 euro i 2002(12). Det er vigtigt at bemærke, at investeringsstrategiens præstation, jf. figur 5.1, tager højde for transaktionsomkostninger. I den henseende følges den forsigtige tilgang, anvendt i Drobetz & Otto (2021), hvor 57 basispoint (bps) fratrækkes porteføljernes må-

[5] En portefølje, der er lang i samtlige (tilgængelige) aktier i datasættet.

[6] Bemærk, at det jf. figur 5.1 er plottet som den realiserede kursændring, dermed er kort-elementet (kvantil 1) negativt illustreret, men som positionen kortsælges er det en gevinst.

Figur 5.1 Kumulative (værdivægtede) afkast af genererede porteføljers præstation



Note: Figuren viser det kumulative (værdivægtede) logafkast for porteføljer sorteret efter afkastprædikteringerne (kvantil 1 og 5). De ubrudte og stiplede linjer repræsenterer henholdsvis lange (kvantil 5) og korte (kvantil 1) positioners realiserede afkast. Det vil sige, at fald i realiserede afkast for kort positionerne er en gevinst for strategien, da de kortsælges. Dertil er markedsporteføljen illustreret.

nedlige merafkast[7]. De valgte transaktionsomkostninger repræsenterer et konservativt skøn og er derfor en nedre grænse for udførelsen af en realistisk handelsstrategi.

Resultatet stemmer med de tidligere resultater og viser kvalitativ sammenlignelighed. Investeringsstrategien identificerer effektivt højt (og til dels lavt) præsterende aktier, selvom algoritmernes evne hertil varierer, hvilket resulterer i betydelige forskelle i præstation. OLS-modellen demonstrerer den stærkeste præstation, mens RF-algoritmen endnu en gang er den mindst gunstige. Præstationen er primært drevet af det lange element (kvantil 5), hvilket er fordelagtigt set fra et investorperspektiv. Historisk set er afkastene genereret af det lange element fra RF og GB tæt på linje indtil omkring 2010, hvor GB siden har udvist en stærkere præstation og hentet ind på det OLS-genererede afkast. Værditilvæksten af elementet er (for alle algoritmer) yderst tilfredsstillende. Det lange element har klaret sig bedre end markedet og med undtagelse af en kraftig nedgang i 2009, oplever dette element ikke andre væsentlige nedture. Denne nedgang kan tilskrives algoritmernes eksponering for momentum, da denne periode var vidne til et momentumkrak, som diskuteret af Daniel & Moskowitz (2016). Desuden førte det nylige

[7] Drobetz & Otto (2021) argumenterer, at de 57 bps er en mere forsigtig tilgang end gængse transaktionsomkostningsmodeller, som de også estimerer, da de finder, at de faktiske omkostninger er lavere end de 57 bps.

globale chok forårsaget af COVID-19-pandemien i begyndelsen af 2020 ikke til en betydelig reduktion i elementets præstation, i overensstemmelse med litteraturen (Gu et al. 2020, Hanauer & Kalsbach 2023). Det korte element (kvantil 1) har derimod bidraget med mere begrænset værditilvækst[8] til de førnævnte H-L-afkast, med undtagelse af under finanskrisen i 2008/09 og for nylig (siden 2020). Førstnævnte tilskrives den generelle markedsnedgang med finanskrisens fremkomst, omend det korte element har et betydeligt større afkast end tabet på det lange element (og på den måde synes aktierne korrekt sorteret). Det korte element har således sværere ved at skabe værdi. Omend det korte element genereret af GB synes at vise en mere negativ hældning sammenlignet med OLS, er forholdet svagt. Siden finanskrisen har RF stort set ikke skabt afkast på dette element før 2020. Det observerede kan skyldes to grunde. For det første er det muligt, at algoritmerne forkert klassificerer nogle af aktierne i kvantil 1. For det andet er det muligt, at nogle af aktierne i kvantil 1 har en positiv afkastprædiktering. Bemærk, at mens den første årsag er en prædikteringsfejl i algoritmen, der kun kan vurderes out-of-sample, kan den anden årsag forudses in-sample og korrigeres (ved kun at kortsælge dem, der prædikteres til at falde).

I undersøgelsen af Gu et al. (2020) observeres det, at det kumulative lange element afledt af den OLS-genererede portefølje overgår GB og er bemærkelsesværdigt lig det lange element genereret af RF. Denne observation vedrører specifikt deres simple OLS-model (med kun tre variable) og giver understøttende dokumentation for resultaterne præsenteret i denne afhandling. Ved at se nærmere på, hvad der præcist driver deres outperformance i forhold til den lineære model, findes det, at næsten alle Sharpe-Ratio gevinster kommer fra det korte element, mens ydeevnen af det lange element på tværs af metoderne er ens. I deres analyse kan kun RF og NN generere gevinster og dermed positive Sharpe-Ratioer ved at gå kort. Denne afhandling kan i kontrast hertil ikke på samme måde skabe afkast på det korte element. Det kan (muligvis) skyldes afhandlingens fokus på at ekskludere microcaps, der er mere ustabile og usikre, og derfor kan signalet være for svagt til at kunne skabe en anstændig kort portefølje. Desuden er gevinsterne på kort elementet, hos Gu et al. (2020), primært genereret før 2003, hvilket er misvisende i forhold til faktisk at skulle implementere strategien fremadrettet. Hermed er det dog vigtigt at bemærke, at rentabiliteten ikke er fuldstændig ikkeeksisterende fra kort-strategien i denne afhandling.

Derudover bekræfter Drobetz et al. (2019), der kun undersøger OLS-modellen (for det europæiske aktiemarked), at dens afkastprædikteringer udmønter sig i høj værditilvækst for investorer. For en aktiv handelsstrategi finder de en stærk markedsoutperformance fratrukket transaktions-

[8] Endda med negativ værditilvækst i nogle perioder.

omkostninger baseret på en række forskellige præstationsmål.

Alt i alt er særligt den værdivægtede OLS præstation overlegen de to andre i værditilvækst.

5.3 Robusthedsanalyse

For at tjekke robustheden af de præsenterede resultater, undersøges virkningen af forskellige metodologiske ændringer, opsummeret i tabel 5.5. Fokus er på at sammenligne benchmark OLS med ENS$^{(1)}$. ENS$^{(1)}$ er valgt for ikke at være drevet af en fremadrettet bias med hensyn til modelvalget og dens porteføljeydelse. I Panel A inkluderes resultatet af følgende to regressioner:

$$\begin{aligned} r_{LS,t,ENS^{(1)}} &= \alpha + \beta_{OLS} \cdot r_{LS,t,OLS} + \varepsilon_t \\ r_{LS,t,OLS} &= \alpha + \beta_{ENS^{(1)}} \cdot r_{LS,t,ENS^{(1)}} + \varepsilon_t \end{aligned} \tag{5.3}$$

, hvor et positivt og signifikant α er en indikation på, at afkastet på højre siden ikke fuldt ud kan forklare porteføljeafkastet på venstre siden. I den første regression repræsenterer $r_{LS,t,ENS^{(1)}}$ H-L-afkastet af porteføljen genereret af ENS$^{(1)}$-modellen, og $r_{LS,t,OLS}$ repræsenterer H-L-afkastet af porteføljen genereret af OLS-modellen. Dermed er regressionen en estimering af forholdet mellem disse to afkast og en evaluering af forklaringskraften og betydningen af afkastene mellem.

Det positive og signifikante alfa for regressionen med det værdivægtede afkast, jf. tabel 5.5 Panel A, er en indikation på, at der er bevis for unik information i porteføljernes afkast. Det vil sige, at $r_{LS,t,ENS^{(1)}}$ og $r_{LS,t,OLS}$ har genereret merafkast, der er højere end det, der kan forklares af den anden portefølje. Det vil sige, at begge porteføljer indeholder unik information, der ikke er indeholdt i den anden, hvilket også viste sig i ENS$^{(2)}$, der gav den bedste præstation.

I Panel B af tabel 5.5 konstrueres H-L-spændet ved anvendelse af en decilsortering[9] frem for en kvantilsortering. Ved at fokusere på de mere ekstreme prædikterede merafkast øges det værdivægtede H-L-afkast til 4,35 pct. (t-statistik på 11,73) for OLS, hvorimod seksfaktormodellens alfa øges tilsvarende. ENS$^{(1)}$ viser en stigning i H-L-afkastet til 3,4 pct. (t-statistik på 8,33). Derfor viser både OLS og ENS$^{(1)}$, i overensstemmelse med forventningen, et stærkere resultat ved anvendelsen af en decilsortering. Stigningen i merafkastet af OLS er højere end af ENS$^{(1)}$.

Desuden er kvantilopdelingen også forsøgt grupperet inden for det enkelte marked således, at det undgås, at eksponeringen eksempelvis kun er i ét af disse, men de realiserede merafkast er

[9] Det vil sige en inddeling af prædikteringerne i ti grupper frem for fem.

Tabel 5.5 Robusthedstjek

	Værdivægtet			
	Gns. [%]	t-stat	α	t-stat$_\alpha$
Panel A: Regression				
ENS$^{(1)}$ $\sim$ OLS			0,0103	3,96
OLS $\sim$ ENS$^{(1)}$			0,0122	4,25
Panel B: Decilsortering				
OLS	4,35	11,73		
ENS$^{(1)}$	3,40	8,33		
Panel C: CAPM				
α_{CAPM} - OLS			3,14	10,19
α_{CAPM} - ENS$^{(1)}$			2,70	9,54

stort set uændrede, på tværs af algoritmerne. Ingen statistisk signifikante forskelle er at finde.

Som sidste robusthedstest er α fra en CAPM-model også estimeret med markedsporteføljen givet markedet i denne afhandling. Hvis α-estimaterne herfra udviser lignende mønstre og betydning, øger det generaliserbarheden af resultaterne og styrker validiteten. På trods af dens simplicitet, giver estimering af markedsporteføljen ud fra de anvendte data større kontrol, konsistens og relevans for analysen, hvilket sikrer, at markedsporteføljen stemmer overens med de aktier, der er undersøgt. Resultatet, jf. Panel C i tabel 5.5, stemmer med de tidligere, som α'erne stadig er økonomisk relevante og meget signifikante med lave forklaringsgrader (omkring 3 pct.)

5.3.1 Porteføljepræstation høj-risiko versus lav-risiko måneder

Som i Hanauer & Kalsbach (2023) samt Avramov et al. (2022) undersøges præstationen i høj- og lav-risiko måneder med det formål at forstå rentabiliteten af de genererede afkastprædikteringer. Analysen er motiveret af den teoretiske påstand om, at øget priseffektivitet forventes i nærvær af reducerede handelsfriktioner og øget arbitrageaktivitet. Derfor forventes handelsstrategier, der udnytter markedsanomalier at generere større rentabilitet under bindende grænser for arbitrage. Eksisterende empiriske undersøgelser har i vid udstrækning bekræftet denne forventning på tværs af forskellige anomalier. Det antages derfor, at de potentielle gevinster fra anomalibaserede strategier kan blive forstærket i perioder præget af uro på de finansielle markeder. Derfor undersøger denne analyse præstationen af de genererede porteføljer på tværs af forskellige delperioder ved at tage højde for risici ud over de traditionelle faktorer, der har været kontrolleret for.

Som risicimål betragtes tre proxyer: (i) retningen (up/down) af det skandinaviske aktiemarked som helhed (Mkt), (ii) et tidsvarierende risikoaversionsindeks foreslået af Bekaert et al. (2022) (RAbex)[10] og (iii) et investorstemningsindeks fra Baker & Wurgler (2007) (SENT)[11].

Tabel F.2 i bilag F præsenterer en oversigt over H-L-afkastene (både for den lige- og værdivægtede orden) for de anvendte algoritmer på tværs af forskellige delprøver. Resultaterne indikerer, at præstationerne tenderer til at være (en anelse) højere i måneder præget af markedsnedgang, høj risikoaversion og investorsentiment. Sidstnævnte er dog ikke ligeså udtalt for den ligevægtede portefølje, hvor H-L-afkastene mellem det høje og lave niveau ikke er så varierende. Observationen stemmer overens med begrebet **grænser for arbitrage**, da det antyder, at strategiens evne til at kapitalisere på aktier, der er forkert prissat, kan være påvirket af visse markedsforhold eller begrænsninger. Strategien kan muligvis fange nogle af disse fejlpriser og generere positive afkast. I **Panel B**, som repræsenterer de værdivægtede porteføljer, genererer OLS-modellen H-L-afkast fra 2,42 pct. til 3,76 pct. på tværs af markeds- og sentimentniveau. Tilsvarende viser RF-algoritmen afkast, der spænder fra 1,35 pct. til 2,24 pct., mens GB-algoritmen udviser afkast mellem 2,1 pct. til 3,68 pct. $ENS^{(1)}$- og $ENS^{(2)}$-modellerne producerer også positive og signifikante H-L-afkast i de forskellige delprøver. Det vil sige, at på trods af det observerede, er der betydelige afkast selv uden for perioderne, og dermed udelukkes strategiens praktiske udførelse ikke.

Delkonklusion

Opsummerende finder den komparative analyse den stærkeste evidens for OLS, tæt forfulgt af GB[12]. Dette er i kontrast til størstedelen af litteraturen, omend De Nard et al. (2022) også finder højeste Sharpe-Ratio for OLS i deres mere simple set-up. Mens OLS klarer sig overraskende godt for et mere sparsomt set-up, leverer den en dårligere, men stadig plausibel portefølje for et større set-up, svarende til deres RF. Alle modeller fanger dog tværsnitsvariationer og prædikterer med en vis præcision. Hovedparten af rentabiliteten er fra det lange element, og generelt leverer strategierne pæne afkast med moderat volatilitet, der fører til yderst pæne Sharpe-Ratioer. Strategierne har dog høje omsætningshastigheder, der kan begrænse den praktiske rentabilitet, og desuden er afkastene højere i perioder, hvor grænserne for arbitrage er højest.

[10] Indekset RAbex er tilgængeligt på: `https://www.nancyxu.net/risk-aversion-index`.
[11] Indekset SENT er tilgængeligt på: `http://apps.olin.wustl.edu/faculty/zhou/zpublications.html`.
[12] For økonomisk har GB dog den bedste Sharpe-Ratio og ser ud til at præstere bedre i den seneste periode, hvilket kan give støtte til påstanden om, at velfungerende ML-modeller kan reducere den nedadgående risiko (Gu et al. 2020).

Kapitel 6

Diskussion

For effektivt at kunne anvende ML i finans sammenhænge, er det afgørende at forstå det aktuelle problem og de særlige udfordringer forbundet med brugen på dette domæne. Dette afsnit skitserer nøgleudfordringer præsenteret af finansielle markeder, som er relevante for anvendelsen af ML, sammen med andre relevante aspekter, som potentielt forklarer resultatet i den empiriske analyse.

6.1 Kritisk diskussion af litteraturen

I modsætning til litteraturen præsenterer afhandlingen resultater, der fastslår, at OLS-modellen er den bedst præsterende både statistisk og økonomisk - dog med GB-algoritmen tæt forfulgt. Imidlertid understøtter denne påstand en række interessante observationer (og forbehold) fremhævet i litteraturen, hvor nogle allerede er diskuteret i den empiriske analyse. Generelt viser OLS gunstige og robuste resultater, hvis den er estimeret med færre variable (De Nard et al. 2022) end de +900 i Gu et al. (2020), og kan føre til Sharpe-Ratioer på niveau med ML. Desuden har forskellige artikler bekræftet, at OLS med penalisering (f.eks. LASSO) er en konkurrencedygtig implementering sammenlignet med ML-algoritmerne. Navnlig opstår et konsistent mønster, hvor ML-algoritmerne tenderer til at give optimale resultater med et bredt og forskelligartet sæt af variable (Blitz et al. 2023)[1]. Derfor er det vigtigt at forstå resultaterne fra denne afhandling i lyset af de metodologiske valg men også gennem en kritisk vurdering af litteraturen. Som OLS-benchmarket i Gu et al. (2020) kun er baseret på tre variable: størrelse, værdi og momentum, kan ML-metodernes forbløffende overpræstation også være en artefakt af det store sæt af vari-

[1] Selvom dette ikke er universelt gældende, som Fieberg et al. (2023) med færre variable finder ML overlegen sammenlignet med OLS.

able. Eksempelvis finder Fieberg et al. (2023) i et ikkerapporteret resultat, at OLS-benchmarket opnår en præstation svarende til den bedst ydende i Gu et al. (2020), hvis der er anvendt et sammenligneligt sæt af variable på de originale data herfra, hvilket stemmer med denne afhandling. Et generelt problem er, at sammenligningerne, mellem de lineære og ikke-lineære metoder, ikke er én-til-én. Alt i alt, finder Gu et al. (2020) porteføljegevinster ved at anvende ML-metoder i forhold til et simpelt benchmark. De er økonomisk og statistisk signifikante, men gevinsterne er på ingen måde revolutionære. Bemærk først, at deres OLS benchmark sandsynligvis er et alt for forenklet benchmark, da selv usofistikerede kapitalforvaltere har en tendens til at bruge en mere avanceret model end dette. For det andet, mens præstationen af strategierne måles brutto af transaktionsomkostninger, betyder det ikke, at sammenligningen nøjagtigt er til. Des mere sofistikeret en model er, des mere sandsynligt er det at opdage øget forudsigelighed i regioner af de data, der er de dyreste at handle. Begge overvejelser indsnævrer kløften mellem ML og OLS.

En række undersøgelser er også mere forsigtige med hensyn til merværdien af ML. Novy-Marx (2015) har f.eks. udtrykt en bekymring over, at "strategies selected by combining multiple signals suffer severe overfitting biases, because underlying signals are typically signed such that each predicts positive in-sample returns". Han viser, at selv med tilfældigt genererede signaler er det nemt at generere en meget signifikant præstation i backtests. Selvom denne bekymring også gælder for traditionelle multifaktormodeller, forværrer de yderligere frihedsgrader, der følger med ML-tilgange, problemet. Blitz et al. (2023) er også mere tilbageholdende i sin konklusion:

> "Being mindful of a positive publication bias, we caution that ML is not a panacea, as users need to make important methodological choices, the models can overfit the data, and they are based on the premise that past relations will continue to hold in the future" (Blitz et al. 2023).

I samme stil er konklusionen hos Kelly et al. (2020):

> "Two key points are crucial for understanding the current state of machine learning in the practice of asset management. The first is that research is just taking off and many important questions are yet to be answered. The second is that early research evidence indicates economically and statistically significant improvements in the performance of portfolios that leverage machine learning tools. However, the gains are evolutionary, not revolutionary" (Kelly et al. 2020).

Samlet set giver disse undersøgelser et forsigtigt perspektiv, der fremhæver behovet for omhyggelig overvejelse af metodiske valg, bevidsthed om potentielle overtilpasningsbiases og erkendelse af, at ML-tilgange ikke er en universel (one-size-fits-all) løsning. De understreger vigtigheden af løbende forskning, såvel som behovet for at fortolke præstationsgevinsten i sammenhæng med eksisterende viden og praksis inden for aktivallokering. Dertil er en af de mest udbredte fejl, der observeres i finansiel forskning, tendensen til at anvende ML-algoritmer på datasæt, udføre backtesting af de genererede prædikteringer og gentage denne proces, indtil et gunstigt backtest-resultat viser sig. Dette giver anledning til pseudo-opdagelser. Selvom backtesten inkorporerer en walk-forward out-of-sample metodologi, ligger det grundlæggende problem i den gentagne test af det samme datasæt, hvilket markant øger sandsynligheden for at komme frem til fejlagtige opdagelser (Lopez de Prado 2018). Dermed kompromitteres integriteten af forskningsresultaterne, hvilket bidrager til en spredning af vildledende resultater, der (potentielt) viser en oppustet præstation.

6.2 Machine Learning i finans

6.2.1 Lav signal-til-støj ratio

Implementering af ML-teknikker på det finansielle område giver visse udfordringer. Teknikkerne udmærker sig i datamiljøer karakteriseret ved høje signal-til-støj ratioer. Men prædiktering af afkast på aktiemarkedet udgør en udfordring grundet den lave signal-til-støj ratio, der er iboende i markedsdata (Kelly et al. 2020). Det er en egenskab, der sikres, og konstant forstærkes, af simple økonomiske kræfter som profitmaksimering og konkurrence. Hvis der findes pålidelig information, der kan prædiktere markedsudviklingen, vil markedsdeltagere udnytte denne, indtil dens værdi er faldet. Som et resultat justeres det relevante aktivs pris for at inkorporere det forudsigelige signal fuldstændigt, hvilket medfører, at det ikke længere er muligt at udnytte[2].

ML-modellernes udfordring som følge af lave signal-til-støj ratioer forvirres yderligere af markedets adaptive og dynamiske karakter, som opdagede anomalier (potentielt) er arbitrageret væk med tiden. Kelly et al. (2020) har en analogi til billede genkendelse (et forskningsfelt, hvor ML-metoderne er særlig velegnet), der siger, at på de finansielle markeder kan alle katte forvandle sig til hunde, når algoritmen har lært at bestemme en kat, og algoritmen skal således begynde at lære forfra. Analogien advarer om, at relevansen af tidligere datapunkter ikke er konstant, da den datagenererende proces kan ændre sig. Derfor udgør modellering af de finansielle markeder

[2]Dette er også i overensstemmelse med den efficiente markedshypotese.

et (meget) mere indviklet problem sammenlignet med andre domæner, hvor ML trives. Perioder med pålidelig modelydeevne kan hurtigt skifte på grund af ændrede forhold, hvilket gør andre eller nye signaler mere relevante for den datagenererende proces.

Slutteligt forværres problemet med den lave signal-til-støj ratio yderligere af begrænsningerne af de data, der er tilgængelige, for i praksis er finans et small data problem. Ved tværsnitsprædiktering kan ethvert t have tusinde aktier med alt fra et par år til et par årtiers data, i alt et par hundrede tusinde observationer, hvilket ikke er meget sammenlignet med ML-standarder. Særligt ikke ved overvejelse af, at det effektive antal observationer er mindre grundet signifikant tværsnitskorrelation i afkastene. En models værdi er ikke begrænset af antallet af forklarende variable, der er mulige at frembringe, men af antallet af observationer på venstre siden (y), som den kan lære af (Kelly et al. 2020). På den måde står aktivallokering igen i kontrast til de miljøer, hvor ML trives. I aktivallokering har frekvensen (daglig, månedlig, årlig, etc.) en indvirkning på transaktionsomkostningerne men også den tilgængelige datamængde. Dermed er modellering med ML altså ikke nødvendigvis oplagt til sådan en problemstilling, medmindre det kan påvises, at de konsistent og i høj nok grad udkonkurrerer andre metoder, selv med få dataobservationer.

Med udgangspunkt i denne afhandling kan der således argumenteres for, at én af grundene til, at ML-algoritmerne er udkonkurreret af den simple lineære OLS er, at det anvendte datasæt er for småt. Dette er også tilfældet, hvis det sammenlignes med de i litteraturen anvendte[3]. Det er således sandsynligt, at det ikke er muligt at fange de potentielle ikke-lineære forhold, som måtte eksistere. Omend de anvendte algoritmer (RF og GB) godt kan fungere med mindre datasæt, vil det at have et større datasæt generelt forbedre modelpræstationen og generaliserbarheden.

6.2.2 Inkorporering af økonomisk teori

Flere argumenterer endvidere for, at inkorporering af økonomisk teori (og faktorstrukturen) kan være en mere effektiv tilgang i et lav-dimensionelt datamiljø med lave signal-til-støj ratioer. På den måde udnytter ML-algoritmen forudgående viden og maksimerer brugen af tilgængelige data (Kelly et al. 2019, 2020, Avramov et al. 2022). Et af grundprincipperne i statistisk analyse er, at teori- og modelparametre er substitutter. Ved at pålægge en model struktur skal der estimeres færre parametre, hvilket gør det muligt for modellen effektivt at udnytte tilgængelige observationer og filtrere støj fra. En forenklet model kan dog filtrere vigtige signaler fra, så i et datarigt miljø med høje signal-til-støj ratioer er en mere kompleks model at foretrække. På den

[3]Hos Gu et al. (2020) er det gennemsnitlige antal månedlig aktier 6.200, og hos Drobetz & Otto (2021) varierer det fra 96 aktier i 1990 til 1.012 aktier i 2020. Hvor den mindste stikprøve, der er fundet i litteraturen, er omkring 500 månedlige aktier (Li et al. 2022). Denne afhandling har i gennemsnit 267 aktier hver måned.

anden side kan enkelhed være fordelagtig, når signal-til-støj rationen er lav, da frafiltrering af støj kan opveje risikoen for at gå glip af et signal. Der kan således argumenteres for, at dette (potentielt) kunne forbedre ML-algoritmerne i denne afhandling, som afkastmodellering med en ustruktureret tilgang kan vise sig ineffektiv. Det skyldes, at der en stræk tro i finansprofessionen på, at afkast følger en faktorstruktur. Økonomiske teorier tyder konsekvent på, at afkast er drevet af nogle få almindelige risikokilder, mens den resterende variation er idiosynkratisk støj. Afhandlingen har haft troen på, at ML-algoritmerne kunne opdage den ud fra data, men måske der ikke har været tilstrækkeligt med data sammenlignet med den del af litteraturen, der også estimerer uden økonomisk struktur, men stadig finder, at ML-algoritmerne er overlegne.

6.3 Mulige forbedringer

De træbaserede algoritmer, RF og GB, viser ikke en konkurrencefordel i forhold til OLS, men GB viser imidlertid en stærk afkast-/risikopræstation. Hvorimod prædikteringsstatistikkerne ikke påviser en signifikant statistisk forbedring, hvilket tyder på, at de kan være modtagelige for overtilpasning på trods af de regulariseringsteknikker, der er indlejret i disse, såsom bagging (Breiman 2001) og boosting (Friedman 2001). En overtilpasning opstår, når algoritmen lægger for meget vægt på støj som følge af kompleksiteten af de underliggende data. Hovedsymptomet på overtilpasning er, når algoritmen klarer sig godt på træningsdataene, men markant underpræsterer på testdataene. Dette synes ikke tilfældet, da fejlene på test- og træningsdataene ikke er væsentligt forskellige, dog højere for testdataene, hvilket er forventeligt. Problemet kan derimod skyldes en undertilpasning, at modellerne er for simple til at fange de underliggende mønstre. Dette kan resultere i højt bias og dårlig præstation på både trænings- og testsættene[4].

Det er sandsynligt, at modellerne ikke inkluderer alle relevante variable, der er nødvendige for at forklare variabiliteten i målvariablen. Derfor kunne et næste skridt således være at inkludere anden-/tredjegradspolynomier og interaktioner. I litteraturen er det fastslået, at ikke-linearitet men også interaktion mellem variable er vigtige (Gu et al. 2020, Daul et al. 2022, Li et al. 2022), selvom ikke alle er enige. Anvendelsen af feature-engineering ville ligeledes være oplagt med henblik på at skabe yderligere meningsfulde variable eller transformere eksisterende variable for at give mere information til modellen (kan hjælpe med at fange nuancerne i dataene og forbedre modellens ydeevne (Hastie et al. 2009)). Datasættet i sig selv kan have høj variabilitet eller støj, hvilket gør det udfordrende for modellen at fange de underliggende mønstre nøjagtigt. Hermed er det vigtigt at skabe stærke signaler, som modellerne kan lære af. En populær metode er

[4]Tabel 5.1 gav også indikation på visse vanskeligheder for særligt RF.

Pincipal Component Analysis (PCA), som er et feature-extraction værktøj, der transformerer en højdimensionel variabelvektor til en lavdimensionel variabelvektor (Hastie et al. 2009).

Resultaternes reliabilitet er desuden påvirket negativt af datasættets kvalitet. På trods af den grundige screening af både statiske data og dynamiske data, er det endelige datasæt stadig ikke fejlfrit, da det er tydeligt, at der har været en relativ mængde af manglende data, som er imputeret med medianen men håndteret som i litteraturen. Det er klart, at dataets repræsentativitet er vigtigt for præstationen. Det bemærkes også, at hvis der igen sammenlignes til Gu et al. (2020), er mængden af historiske data, der tages i betragtning, meget forskellig, da de har data fra 1957 til 2016, mens denne afhandling har data fra 1994 til 2022. Det må også betyde, at resultaterne i Gu et al. (2020) kan være drevet af imputeringen. Fieberg et al. (2023) rapporterer f.eks. at i 10 pct. af de 94 karakteristika, som Gu et al. (2020) inkluderer, er mere end 30 pct. af disse manglende i 1985. Det kunne således være interessant at begrænse stikprøven til selskaber med kun den fulde information for at se, hvorledes dette har en indflydelse på resultaterne.

Antallet af hyperparametre giver også en udfordring i forhold til at definere, om den bedst mulige model er opnået. Der eksisterer nogle begrænsninger, både i form af tid og beregningskraft, som begrænser antallet af mulige kombinationer, der kan testes. Desværre er disse begrænsninger svære at løse og udgør mere en afvejning mellem kompleksitet og gennemførlighed end den sande (optimale) løsning, og her har afhandlingen været begrænset på beregningskraften. Derfor kan suboptimal tuning også være en eventuel forklaring på afhandlingens resultat. Afhandlingen skulle måske i større grad også have afsøgt andre parametre end de i litteraturen anvendte, f.eks. min_sample_split eller en mindre beregningstung teknik som random search, hvor tilfældige kombinationer udtages fra søgeområdet (Chan & Mátyás 2022).

Kapitel 7

Konklusion

Undersøgelsen af Gu et al. (2020), der statuerer ML-modellers overlegenhed til at prædiktere tværsnitsforskelle i aktieafkast, har igangsat en spirende litteratur, der undersøger almenheden af denne konklusion. Derfor har denne afhandling foretaget en sammenligning af den almindelig anvendte lineære regressionstilgang (OLS) med de træbaserede ML-tilgange, RF og GB. Dette er sket ved at afvige fra litteraturen på tre væsentlige punkter: (1) at undersøge det uudforskede skandinaviske aktiemarked, (2) at bruge et mere begrænset sæt af variable, der tillader en sammenligning i øjenhøjde, og (3) at ekskludere microcaps. Herpå analyserer afhandlingen den statistiske evne til at prædiktere merafkastet (hypotese **1**) men også den økonomiske præstation i forhold til, at den omsættes til økonomisk rentabilitet (hypotese **2**). Det er dog afgørende at træne og tune ML-modellerne effektivt for at mindske risikoen for en over-/undertilpasning.

In-sample identificerer modellerne lignende vigtige variable såsom momentum, men også værdi-, handelsfriktions- og sentimentbaserede variable er i top-20. Selvom OLS giver de mest pålidelige prædikteringer, tæt forfulgt af GB, fanger alle modeller tværsnitsvariationer og demonstrerer rimelig præcision. Forskelle i den statistiske præstation fører til forskelle i den økonomiske rentabilitet. De anvendte afkastmål indikerer, at OLS klarer sig bedre end både RF og GB, selvom den risikojusterede præstation er en anelse bedre for GB, hvilket kan give støtte til påstanden om, at velfungerende ML-modeller kan reducere den nedadgående risiko. Alle udmønter sig dog i yderst tilfredsstillende Sharpe-Ratioer. Størstedelen af rentabiliteten er fra det lange element (kvantil 5), hvilket er fordelagtigt fra et investorperspektiv. Implementeringen i praksis kan dog støde på begrænsninger i form af høje omsætningshastigheder og tendensen til, at afkastet er højere, når arbitragebegrænsningerne er mest fremtrædende. Dermed afvises den opstillede hypotese **3**, der statuerede, at ML-modellerne ville udkonkurrere OLS. Den bemærkelsesværdige

underpræstation af ML-modellerne, til sammenligning med litteraturen, kan potentielt tilskrives forskellige faktorer, herunder begrænsningerne i datasættets størrelse sammen med iboende begrænsninger i finansielle applikationer og/eller suboptimal tuning (der kan føre til en over- eller undertilpasning af ML-modellerne). Der findes dog moderat støtte til afhandlingens resultat, hvis der ses på de metodologiske valg, og hvis der ses mere kritisk på de eksisterende resultater i litteraturen. Fremtidig forskning bør forsøge sig med feature-engineering, Deep-Learning, interaktioner/polynomier samt en klassifikationstilgang som i Drobetz & Otto (2021).

Litteratur

Andrew Karolyi, G. (2016), 'Home Bias, an Academic Puzzle', *Review of Finance* **20**(6), 2049–2078.

Annaert, J., De Ceuster, M. & Verstegen, K. (2013), 'Are extreme returns priced in the stock market? European evidence', *Journal of Banking & Finance* **37**(9), 3401–3411.

Avramov, D., Cheng, S. & Metzker, L. (2022), 'Machine Learning vs. Economic Restrictions: Evidence from Stock Return Predictability', *Management Science* **69**(5), 2587–2619.

Baker, M. & Wurgler, J. (2007), 'Investor Sentiment in the Stock Market', *The Journal of Economic Perspectives* **21**(2), 129–151.

Ballings, M., Van den Poel, D., Hespeels, N. & Gryp, R. (2015), 'Evaluating multiple classifiers for stock price direction prediction', *Expert Systems with Applications* **42**(20), 7046–7056.

Banz, R. W. (1981), 'The Relationship Between Return and Market Value of Common Stocks', *Journal of Financial Economics* **9**(1), 3–18.

Basu, S. (1977), 'Investment Performance of Common Stocks in Relation to Their Price-Earnings Ratios: A Test of the Efficient Market Hypothesis', *Journal of Finance* **32**(3), 663–82.

Bekaert, G., Engstrom, E. C. & Xu, N. R. (2022), 'The Time Variation in Risk Appetite and Uncertainty', *Management Science* **68**(6), 3975–4004.

Bianchi, D., Büchner, M. & Tamoni, A. (2020), 'Bond Risk Premiums with Machine Learning', *The Review of Financial Studies* **34**(2), 1046–1089.

Blitz, D., Hoogteijling, T., Lohre, H. & Messow, P. (2023), 'How can machine learning advance quantitative asset management?', *Journal of Portfolio Management* .

Breiman, L. (1996), 'Bagging predictors', *Machine Learning* **24**, 123–140.

Breiman, L. (2001), 'Random forests', *Machine Learning* **45**, 5–32.

Bryzgalova, S., Pelger, M. & Zhu, J. (2020), 'Forest through the trees: Building cross-sections of stock returns', *SSRN Working Paper* (3493458).

Cakici, N., Fieberg, C., Metko, D. & Zaremba, A. (2022), 'Machine learning goes global: Cross-sectional return predictability in international stock markets', *SSRN Working Paper* (4141663).

Carhart, M. M. (1997), 'On Persistence in Mutual Fund Performance', *The Journal of Finance* **52**(1), 57–82.

Cawley, G. C. & Talbot, N. L. C. (2010), 'On over-fitting in model selection and subsequent selection bias in performance evaluation', *Journal of Machine Learning Research* **11**, 2079–2107.

Chan, F. & Mátyás, L. (2022), *Econometrics with machine learning*, Vol. 53 of *Advanced Studies in Theoretical and Applied Econometrics*, Springer.

Chen, L., Pelger, M. & Zhu, J. (2023), 'Deep Learning in Asset Pricing', *arXiv.org* .

Chen, T. & Guestrin, C. (2016), XGBoost: A Scalable Tree Boosting System, *in* 'Proceedings of the 22nd ACM SIGKDD International Conference on knowledge discovery and data mining', Vol. 13-17- of *KDD '16*, ACM, pp. 785–794.

Cochrane, J. H. (2011), 'Presidential Address: Discount Rates', *The Journal of Finance* **66**(4), 1047–1108.

Cong, L., Tang, K., Wang, J. & Zhang, Y. (2021), 'AlphaPortfolio: Direct Construction Through Deep Reinforcement Learning and Interpretable AI', *SSRN Working Paper* (3554486).

Daniel, K. & Moskowitz, T. J. (2016), 'Momentum crashes', *Journal of Financial Economics* **122**(2), 221–247.

Darmer, P., Jordansen, B., Madsen, J. A. & Thomsen, J. (2010), *Paradigmer i praksis*, Handelshøjskolens Forlag.

Daul, S., Jaisson, T. & Nagy, A. (2022), 'Performance attribution of machine learning methods for stock returns prediction', *The Journal of Finance and Data Science* **8**, 86–104.

De Nard, G., Hediger, S. & Leippold, M. (2022), 'Subsampled factor models for asset pricing: The rise of Vasa', *Journal of Forecasting* **41**(6), 1217–1247.

Diebold, F. & Mariano, R. (1995), 'Comparing Predictive Accuracy', *Journal of Business & Economic Statistics* **13**(3), 253–63.

Dietterich, T. G. (2000), Ensemble Methods in Machine Learning, *in* 'Multiple Classifier Systems', Springer Berlin Heidelberg, pp. 1–15.

Drobetz, W., Haller, R., Jasperneite, C. & Otto, T. (2019), 'Predictability and the Cross Section of Expected Returns: Evidence from the European Stock Market', *Journal of Asset Management* **20**(7), 508–533.

Drobetz, W. & Otto, T. (2021), 'Empirical asset pricing via machine learning: evidence from the European stock market', *Journal of Asset Management* **22**(7), 507–538.

Elliott, G., Granger, C. W. J. & Timmermann, A. G. (2006), Handbooks in Economics, *in* 'Handbook of Economic Forecasting', Vol. 1, Elsevier Science & Technology.

Fama, E. F. & French, K. R. (1993), 'Common risk factors in the returns on stocks and bonds', *Journal of Financial Economics* **33**(1), 3–56.

Fama, E. F. & French, K. R. (2015), 'A Five-Factor Asset Pricing Model', *Journal of Financial Economics* **116**(1), 1–22.

Fama, E. F. & French, K. R. (2017), 'International tests of a five-factor asset pricing model', *Journal of Financial Economics* **123**(3), 441–463.

Fama, E. F. & MacBeth, J. D. (1973), 'Risk, Return, and Equilibrium: Empirical Tests', *Journal of Political Economy* **81**(3), 607–636.

Feng, G., Giglio, S. & Xiu, D. (2020), 'Taming the Factor Zoo: A Test of New Factors', *The Journal of Finance* **75**(3), 1327–1370.

Feng, G., He, J. & Polson, N. G. (2018), 'Deep Learning for Predicting Asset Returns', *arXiv.org* .

Fernández-Delgado, M., Cernadas, E., Barro, S. & Amorim, D. (2014), 'Do we need hundreds of classifiers to solve real world classification problems?', *Journal of Machine Learning Research* **15**, 3133–3181.

Fieberg, C., Metko, D., Poddig, T. & Loy, T. (2023), 'Machine learning techniques for cross-sectional equity returns' prediction', *OR Spectrum: Quantitative Approaches in Management* **45**(1), 289–323.

Fong, K. Y. L., Holden, C. W. & Trzcinka, C. A. (2017), 'What Are the Best Liquidity Proxies for Global Research?', *Review of Finance* **21**(4), 1355–1401.

Freyberger, J., Neuhierl, A. & Weber, M. (2020), 'Dissecting Characteristics Nonparametrically', *The Review of Financial Studies* **33**(5), 2326–2377.

Friedman, J. H. (2001), 'Greedy Function Approximation: A Gradient Boosting Machine', *The Annals of Statistics* **29**(5), 1189–1232.

Fuglsang, L. & Bitsch-Olsen, P. (2013), *Videnskabsteori i samfundsvidenskaberne - på tværs af fagkulturer og paradigmer*, Samfundslitteratur.

Green, J., Hand, J. R. M. & Zhang, X. F. (2017), 'The Characteristics that Provide Independent Information about Average U.S. Monthly Stock Returns', *The Review of Financial Studies* **30**(12), 4389–4436.

Griffin, J. M., Kelly, P. J. & Nardari, F. (2010), 'Do Market Efficiency Measures Yield Correct Inferences? A Comparison of Developed and Emerging Markets', *The Review of Financial Studies* **23**(8), 3225–3277.

Grobys, K. & Huhta-Halkola, T. (2019), 'Combining value and momentum: evidence from the Nordic equity market', *Applied Economics* **51**(26), 2872–2884.

Gu, S., Kelly, B. & Xiu, D. (2020), 'Empirical Asset Pricing via Machine Learning', *The Review of Financial Studies* **33**(5), 2223–2273.

Gu, S., Kelly, B. & Xiu, D. (2021), 'Autoencoder asset pricing models', *Journal of Econometrics* **222**(1), 429–450.

Hanauer, M. X. & Kalsbach, T. (2023), 'Machine learning and the cross-section of emerging market stock returns', *Emerging Markets Review* **55**, 101022.

Hanauer, M. X., Kononova, M. & Rapp, M. S. (2022), 'Boosting agnostic fundamental analysis: Using machine learning to identify mispricing in European stock markets', *Finance Research Letters* **48**, 102856.

Hastie, T., Tibshirani, R. & Friedman, J. (2009), *Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics, Springer.

Haugen, R. A. & Baker, N. L. (1996), 'Commonality in the Determinants of Expected Stock Returns', *Journal of Financial Economics* **41**(3), 401–439.

Hillier, D., Grinblatt, M. & Titman, S. (2012), *Financial Markets and Corporate Strategy*, McGraw-Hill European Editions, McGraw-Hill.

Hou, K., Xue, C. & Zhang, L. (2020), 'Replicating Anomalies', *The Review of Financial Studies* **33**(5), 2019–2133.

Hwang, T. (2018), 'Computational Power and the Social Impact of Artificial Intelligence', *arXiv.org* .

Ince, O. S. & Porter, R. B. (2006), 'Individual Equity Return Data from Thomson Datastream: Handle with Care', *The Journal of Financial Research* **29**(4), 463–479.

Jegadeesh, N. & Titman, S. (1993), 'Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency', *The Journal of Finance* **48**(1), 65–91.

Kelly, B., Israel, R. & Moskowitz, T. (2020), 'Can Machines "Learn" Finance?', *Journal of Investment Management* **18**(2).

Kelly, B. T., Pruitt, S. & Su, Y. (2019), 'Characteristics are covariances: A unified model of risk and return', *Journal of Financial Economics* **134**(3), 501–524.

Krauss, C., Do, X. A. & Huck, N. (2017), 'Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500', *European Journal of Operational Research* **259**(2), 689–702.

Leippold, M., Wang, Q. & Zhou, W. (2022), 'Machine learning in the Chinese stock market', *Journal of Financial Economics* **145**(2), 64–82.

Leitch, G. & Tanner, J. E. (1991), 'Economic Forecast Evaluation: Profits Versus the Conventional Error Measures', *The American Economic Review* **81**(3), 580–590.

Lesmond, D. A., Ogden, J. P. & Trzcinka, C. A. (1999), 'A New Estimate of Transaction Costs', *The Review of Financial Studies* **12**(5), 1113–1141.

Leung, E., Lohre, H., Mischlich, D., Shea, Y. & Stroh, M. (2021), 'The Promises and Pitfalls of Machine Learning for Predicting Stock Returns', *The Journal of Financial Data Science* **3**(2), 21–50.

Lewellen, J. (2015), 'The Cross-section of Expected Stock Returns', *Critical Finance Review* **4**(1), 1–44.

Li, Y., Simon, Z. & Turkington, D. (2022), 'Investable and Interpretable Machine Learning for Equities', *The Journal of Financial Data Science* **4**(1), 54–74.

Lopez de Prado, M. (2018), *Advances in Financial Machine Learning*, John Wiley & Sons.

Messmer, M. (2017), 'Deep Learning and the Cross-Section of Expected Returns', *SSRN Electronic Journal* .

Mitchell, T. M. (1997), *Machine Learning*, McGraw-Hill International Editions, McGraw-Hill.

Moritz, B. & Zimmermann, T. (2016), 'Tree-Based Conditional Portfolio Sorts: The Relation between Past and Future Stock Returns', *Capital Markets: Market Efficiency eJournal* .

Mullainathan, S. & Spiess, J. (2017), 'Machine Learning: An Applied Econometric Approach', *The Journal of Economic Perspectives* **31**(2), 87–106.

Nagel, S. (2021), *Machine Learning in Asset Pricing*, Vol. 8, Princeton University Press.

Nasdaq-Nordic (2023), 'Shares - share prices for all companies listed on NASDAQ Nordic', https://www.nasdaqomxnordic.com/aktier. Tilgået den: 26-05-2023.

Nilsson, N. J. (1996), 'Artificial intelligence: A modern approach', *Artificial intelligence* **82**(1-2), 369–380.

Novy-Marx, R. (2015), 'Backtesting Strategies Based on Multiple Signals', *NBER Working Paper Series* pp. 21329–.

Novy-Marx, R. & Velikov, M. (2016), 'A Taxonomy of Anomalies and Their Trading Costs', *The Review of Financial Studies* **29**(1), 104–147.

Oslo-Børs (2023), 'Oslo Børs', https://www.euronext.com/en/markets/oslo. Tilgået den: 26-05-2023.

Probst, P., Boulesteix, A.-L. & Bischl, B. (2019), 'Tunability: Importance of hyperparameters of machine learning algorithms', *Journal of Machine Learning Research* **20**.

Rubesam, A. (2022), 'Machine learning portfolios with equal risk contributions: Evidence from the Brazilian market', *Emerging Markets Review* **51**(Part B), 100891–.

Schmidt, P. S., Schrimpf, A., von Arx, U., Wagner, A. F. & Ziegler, A. (2011), 'On the Construction of Common Size, Value and Momentum Factors in International Stock Markets: A Guide with Applications', *IDEAS Working Paper Series from RePEc* .

Singh, J. & Khushi, M. (2021), 'Feature Learning for Stock Price Prediction Shows a Significant Role of Analyst Rating', *Applied System Innovation* **4**(1), 17–.

Tobek, O. & Hronec, M. (2021), 'Does it pay to follow anomalies research? Machine learning approach with international evidence', *Journal of Financial Markets* **56**, 100588–.

Welch, I. & Goyal, A. (2008), 'A Comprehensive Look at the Empirical Performance of Equity Premium Prediction', *The Review of Financial Studies* **21**(4), 1455–1508.

West, K. D. (2006), Chapter 3 Forecast Evaluation, *in* 'Handbook of Economic Forecasting', Vol. 1, Elsevier B.V, pp. 99–134.

Wilson, D. & Martinez, T. (1997), Bias and the probability of generalization, *in* 'Proceedings Intelligent Information Systems. IIS'97', pp. 108–114.

Windmuller, S. (2022), 'Firm Characteristics and Global Stock Returns: A Conditional Asset Pricing Model', *Review of Asset Pricing Studies* **12**(2), 447–499.

Wolpert, D. & Macready, W. (1997), 'No free lunch theorems for optimization', *IEEE Transactions on Evolutionary Computation* **1**(1), 67–82.

Wooldridge, M. J. (2019), *Introductory Econnometrics - A Modern Approach*, Vol. 7th edition, Cengage Learning.

Appendiks A

Datastream

Listerne for samtlige aktive og døde selskaber (for hvert af de skandinaviske aktiemarkeder) er indhentet fra Refinitiv Datastream og anvendt i den statiske filteringsproces. Stikprøven begrænses til almindelige aktier ved hjælp af flere statiske skærme, som illustreret i tabel A.1. Skærme (1) til (7) er ligetil at anvende og almindelige i litteraturen. Skærm (8) er baseret på arbejde fra blandt andre Ince & Porter (2006). Disse forfattere giver generiske filteringsregler (søgeord) for at kunne ekskludere ikke-almindelige aktier fra Refinitiv Datastream. Derfor er de identificerede søgeord anvendt og matchet med navnene fra Datastream. En aktie er ekskluderet fra stikprøven, hvis søgeordet er sammenfaldende med dens navn. De følgende tre Datastream-elementer gemmer navne og er anvendt på søgeordsfiltrene: NAME, ENAME og ECNAME. Tabel A.2 giver en oversigt over de anvendte søgeord, hvoraf nogle naturligt er oversat til henholdsvis dansk, norsk og svensk.

For de tilbageværende aktier (fra de ovenfor beskrevne statiske skærme) blev data om afkast og markedsværdi indhentet fra Datastream, mens regnskabsdata blev hentet fra Worldscope. Adskillige dynamiske skærme, der er almindelige i litteraturen, blev implementeret for at tage højde for datafejl, hovedsageligt inden for afkast-variablen. De dynamiske skærme er vist i tabel A.3.

Tabel A.1 Statiske skærme

Tabellen viser de statiske skærme anvendt i afhandlingen, hovedsageligt efter Ince & Porter (2006), Schmidt et al. (2011) og Griffin et al. (2010). Kolonne 3 viser de involverede Datastream-elementer (til venstre for lighedstegnet) og de værdier, som de indstilles til i processen (til højre for lighedstegnet).

Nr.	Beskrivelse	Datastream	Kilde
(1)	For selskaber med mere end ét værdipapir, er kun den med den største markedsværdi og likviditet anvendt	MAYOR = Y	Schmidt et al. (2011)
(2)	Værdipapirstypen skal være af aktie	TYPE = EQ	Ince & Porter (2006)
(3)	Kun de primære noteringer af et værdipapir analyseres	PRIMARY = Y	Fong et al. (2017)
(4)	Selskaber er beliggende i det respektive hjemland	GEOGN = Country	Ince & Porter (2006)
(5)	Værdipapirer er noteret i det respektive hjemland	GEOLN = Country	Griffin et al. (2010)
(6)	Værdipapirer, hvis noterede valuta er forskellig fra den i det associerede land, ses bort fra	PCUR = Currency	Griffin et al. (2010)
(7)	Værdipapirer, hvis ISIN-kode er forskellig fra den af det associerede land, ses bort fra	ISIN = Country	Annaert et al. (2013)
(8)	Værdipapirer, hvis navnefelter angiver tilhørsforhold til ikke-almindelige aktier, ses bort fra	NAME, ENAME, ECNAME	Ince & Porter (2006)

Tabel A.2 Generiske søgeordssletninger

Tabellen rapporterer søgeordene, der er søgt efter i NAME, ENAME og ECNAME. Hvis et skadeligt søgeord detekteres som en del af navnet på en aktie, ekskluderes den fra stikprøven.

Ikke-almindelige aktier	Nølgleord
Dubletter	1000DUPL, DULP, DUP, DUPE, DUPL, DUPLI, DUPLICATE, XSQ, XETa
Depotbeviser	ADR, GDR
Foretrukne aktier	PF, 'PF', PFD, PREF, PREFERRED, PRF
Warrants	WARR, WARRANT, WARRANTS, WARRT, WTS, WTS2
Gæld	%, DB, DCB, DEB, DEBENTURE, DEBENTURES, DEBT
Fonde	.IT, .ITb, TST, INVESTMENT TRUST, RLST IT, TRUST, TRUST UNIT, TRUST UNITS, TST, TST UNIT, TST UNITS, UNIT, UNIT TRUST, UNITS, UNT, UNT TST, UT
ETF'er	AMUNDI, ETF, INAV, ISHARES, JUNGE, LYXOR, X-TR
Udløbne værdipapirer	EXPD, EXPIRED, EXPIRY, EXPY
Diverse*	ADS, BOND, CAP.SHS, CONV, DEFER, DEP, DEPY, ELKS, FD, FUND, GW.FD, HI.YIELD, HIGH INCOME, IDX, INC.&GROWTH, INC.&GW, INDEX, LP, MIPS, MITS, MITT, MPS, NIKKEI, NOTE, OPCVM, ORTF, PARTNER, PERQS, PFC, PFCL, PINES, PRTF, PTNS, PTSHP, QUIBS, QUIDS, RATE, RCPTS, REAL EST, RECEIPTS, REIT, RESPT, RETUR, RIGHTS, RST, RTN.INC, RTS, SBVTG, SCORE, SPDR, STRYPES, TOPRS, UTS, VCT, VTG.SAS, XXXXX, YIELD, YLD

* Hovedsageligt taget fra Ince & Porter (2006).

Tabel A.3 Dynamiske skærme

Tabellen viser de dynamiske skærme anvendt på dataene i afhandlingen, efter Ince & Porter (2006), Griffin et al. (2010) og Schmidt et al. (2011). Kolonne 3 viser de respektive Datastream-elementer.

Nr.	Beskrivelse	Datastream	Kilde
(1)	Alle nul-afkast i slutningen af den afkasttidsserie, der eksisterer, slettes, fordi i tilfælde af en afnotering, viser Datastream forældede priser fra datoen for afnoteringen til slutningen af den respektive tidsserie. De tilhørende markedsværdier slettes også	RI, MV	Ince & Porter (2006)
(2)	De tilhørende afkast og markedsværdier i tilfælde af unormale priser (ujusterede priser > 1000000) slettes	RI, MV, UP	Schmidt et al. (2011)
(3)	Månedlige (daglige) afkast og tilhørende markedsværdier, hvis afkastet overstiger 990 pct. (200 pct.) slettes	RI, MV	Ince & Porter (2006)
(4)	Månedlige afkast og tilhørende markedsværdier i tilfælde af kraftige returvendinger, defineret som $(1 + r_{t1})(1 + r_t)1 < 0,5$ givet, at enten r_{t1} eller $r_t \geq 3,0$, slettes	RI, MV	Ince & Porter (2006)
(5)	Daglige afkast og tilhørende markedsværdier i tilfælde af kraftige returvendinger, defineret som $(1 + r_{t1})(1 + r_t)1 < 0,2$ med r_{t1} eller $r_t \geq 1,0$, slettes	RI, MV	Griffin et al. (2010)
(6)	Observationer af aktier, der viser ikke-nul kursændringer i mindre end 50 pct. af de handlede måneder i de foregående 12 måneder, slettes	RI, MV	Griffin et al. (2010)
(7)	Observationer af aktier i de laveste 3 pct. af et lands aggregerede markedsværdi slettes	MV	Hou et al. (2020)

Appendiks B

Definition af variable

Dette afsnit skitserer konstruktionen af de anvendte variable. For hver variabel angives respektive Datastream- og Worldscope-elementer, kategorien[1] og hyppigheden[2]. For relevante referencer til konstruktionerne (og hvor de er fundet relevante) henvises til Hanauer & Kalsbach (2023).

Tabel B.1 Variable

Variabel	Type	Frekvens	Beskrivelse
A2ME	V	Y	**Assets-to-Market:** Forholdet mellem samlede aktiver (WC02999) og markedsværdien i december $t-1$.
AT	H	Y	**Total Assets:** Samlede aktiver målt i euro (WC02999).
ATO	R	Y	**Sales-to-Assets:** Beregnet som nettoomsætningen (WC01001) over laggede nettodriftsaktiver, hvor nettodriftsaktiver er forklaret i konstruktionen af NOA.
BEME	V	Y	**Book-to-Market:** Forholdet mellem den bogførte værdi af egenkapitalen og markedsværdien af egenkapitalen. Den bogførte værdi af egenkapitalen er defineret som fælles egenkapital (WC03501) plus udskudt skat (WC03263). Hvis der ikke er givet udskudt skat, er den bogførte værdi af egenkapitalen lig med fælles egenkapital (WC03501). Markedsværdien af egenkapitalen er per december $t-1$.

[1] Hvor V = Value, H = Handelsfriktioner, R = Rentabilitet, I = Imatrielle aktiver, Inv = Investering og PR = Past Return (momentum)

[2] Hvor Y = Årlig, og M = Månedlig.

Tabel B.2 Variable

Variabel	Type	Frekvens	Beskrivelse
BEME_m	V	M	**Book-to-Market:** Forholdet mellem den bogførte værdi af egenkapitalen og den seneste markedsværdi af egenkapitalen. Den bogførte værdi af egenkapitalen er defineret som for BEME. Den seneste markedsværdi af egenkapitalen er den for slutningen af måned t til prædiktering af afkastet for $t+1$.
Beta	H	M	**Markedsbeta:** Beregnet som summen af regressionskoefficienterne for daglige merafkast på det lokale markeds merafkast og et lag af det lokale markeds merafkast for de seneste 12 måneder. Kravet er mindst 126 observationer for gyldige estimater.
C	V	Y	**Cash-and-Short-Term-Investment-to-Assets:** Forholdet mellem kontanter og kortsigtede investeringer (WC02001) over samlede aktiver (WC02999).
CF2P	V	Y	**Cash Flow-to-Price:** Forholdet mellem nettopengestrømme fra driftsaktiviteter (WC04860) over markedsværdien per december $t-1$.
CTO	R	Y	**Capital Turnover:** Forholdet mellem nettoomsætning (WC01001) og laggede samlede aktiver (WC02999).
D2A	I	Y	**Capital Intensity:** Forholdet mellem afskrivning og amortisering (WC01151) over samlede aktiver (WC02999).
Debt2P	V	Y	**Leverage:** Forholdet mellem de samlede aktiver (WC02999) minus fælles egenkapital (WC03501) over markedsværdien per december $t-1$.
E2P	V	Y	**Earnings-to-Price:** Forholdet mellem indkomst før ekstraordinære poster (WC01551) og markedsværdien per december $t-1$.
FreeCF	V	Y	**Cash Flow-to-Book:** Defineret som frie pengestrømme over den bogførte værdi af egenkapitalen. Frie pengestrømme er beregnet som nettoindkomst (WC01551) plus afskrivninger og amortiseringer (WC01151) minus ændringer i driftskapital minus kapitaludgifter (WC04601). Den bogførte værdi af egenkapitalen er defineret i konstruktionen af BEME.

Tabel B.3 Variable

Variabel	Type	Frekvens	Beskrivelse
GP2A	R	Y	**Gross Profits-to-Assets:** Nettosalg (WC01001) minus omkostninger ved solgte varer (WC01051) divideret med samlede aktiver (WC02999).
Idiovol	H	M	**Idiosynkratisk volatilitet:** Standardafvigelsen af residualerne fra en regression af merafkast på en Fama & French (1993) trefaktormodel. Hertil anvendes daglige data for måneden (med mindst 15 ikke-manglende observationer).
INV	Inv	Y	**Investering:** Den procentvise år-til-år vækstrate af samlede aktiver (WC02999).
LME	H	M	**Markedsværdi:** Markedsværdien ved udgangen af den foregående måned målt i euro.
LTurnover	H	M	**Turnover:** Handelsvolumen (VO) divideret med udestående aktier (NOSH) i løbet af den foregående måned.
NOA	Inv	Y	**Net Operating Assets:** Defineret som forskellen mellem driftsaktiver og driftsforpligtelser, skaleret efter laggede samlede aktiver. Driftsaktiver er samlede aktiver (WC02999) minus kontanter og kortsigtede investeringer (WC02001). Driftsforpligtelser er samlede aktiver (WC02999), minus samlet gæld (WC03255), minus minoritetsinteresser (WC03426), minus præferenceaktier og fælles egenkapital (WC03995).
OA	I	Y	**Operating Accruals:** Beregnet som ændringer i driftskapital minus afskrivninger (WC01151) skaleret efter laggede samlede aktiver (WC02999). Ændringer i driftskapital er ændringer i omsætningsaktiver (WC02201) minus ændringer i likvide midler og kortfristede investeringer (WC02001) minus ændringer i kortfristede forpligtelser (WC03101), plus ændringer i gæld i kortfristede forpligtelser (WC03051) plus ændringer i skyldige indkomstskatter (WC03063).
OL	I	Y	**Operating Leverage:** Defineret som summen af omkostningerne ved solgte varer (WC01051) og salgs-, generelle og administrative omkostninger (WC01101) over samlede aktiver (WC02999).

Tabel B.4 Variable

Variabel	Type	Frekvens	Beskrivelse
P2P52WH	H	M	**Price Relative to its 52-week High:** Forholdet mellem den ujusterede aktiekurs (UP) ved udgangen af den foregående måned og de seneste 52 ugers højeste kurs.
PCM	R	Y	**Price-to-Cost Margin:** Nettosalg (WC01001) minus omkostninger ved solgte varer (WC01051), divideret med nettosalg (WC01001).
PM	R	Y	**Profit Margin:** Beregnet som driftsindtægt efter afskrivninger eller EBIT (WC18191) over salg (WC01001).
Prof	R	Y	**Gross Profitability:** Nettosalg (WC01001) minus omkostninger ved solgte varer (WC01051) divideret med den bogførte værdi af egenkapitalen. Den bogførte værdi af egenkapitalen er defineret i konstruktionen af BEME.
Q	V	Y	**Tobins Q:** Defineret som samlede aktiver (WC02999) plus markedsværdien per december $t-1$ minus kontanter og kortsigtede investeringer (WC02001) og minus udskudte skatter (WC03263), skaleret efter samlede aktiver (WC02999).
r_{12-2}	PR	M	**Momentum:** Afkast fra måned $t-12$ til $t-2$.
r_{12-7}	PR	M	**Intermediate Momentum:** Afkast fra $t-12$ til $t-7$.
r_{2-1}	PR	M	**Short-Term Reversal:** Det laggede en-måneds afkast.
r_{36-13}	PR	M	**Long-Term Reversal:** Afkast fra $t-36$ til $t-13$.
RNA	R	Y	**Return on Net Operating Assets:** Forholdet mellem driftsindtægter efter afskrivninger eller EBIT (WC18191) og laggede nettodriftsaktiver. Nettodriftsaktiver er forklaret i konstruktionen af NOA.
ROA	R	Y	**Return on Assets:** Forholdet mellem indtjening før ekstraordinære poster (WC01551) og laggede samlede aktiver (WC02999).

Tabel B.5 Variable

Variabel	Type	Frekvens	Beskrivelse
ROE	R	Y	**Return on Equity:** Indtjening før ekstraordinære poster (WC01551) til den laggede bogførte egenkapital. Den bogførte værdi af egenkapitalen er defineret i konstruktionen af BEME.
S2P	V	Y	**Sales-to-Price:** Forholdet mellem nettosalg (WC01001) og markedsværdien per december $t-1$.
SGA2S	I	Y	**Sales and General Administrative Costs to Sales:** Forholdet mellem salgs-, generelle og administrative omkostninger (WC01101) og nettoomsætning (WC01001).
Illiqu	H	M	**Illiquidity:** Beregnet som det aritmetiske gennemsnit af følgende forhold for den forgående måned: det daglige absolutte afkast divideret med produktet af lukkekursen (UP) og den daglige handelsvolumen (VO).
SUV	H	M	**Unexplained Volume:** Defineret som forskellen mellem den faktiske volumen og forudsagte volumen i den foregående måned. Den forudsagte volumen kommer fra en regression af daglig volumen på en konstant og de absolutte værdier af positive og negative afkast. Hertil anvendes to måneders data til at estimere modelparametrene (data fra $t-2$ og $t-1$) og estimere den forudsagte volumen ved hjælp af data fra den foregående måned. Der er brug for mindst 15 daglige observationer i den foregående måned. Den er standardiseret ved standardafvigelsen af residualerne fra regressionen.

Dertil kommer DY (Dividend Yield) samt RECSELL, RECHOLD og RECBUY, der er procentdelen af salgs-, hold- og købsanbefalinger. Og til sidst er EPS1MD, EPS2MD og EPS3MD inkluderet, der er medianen af Earnings-per-Share (implicit indtjening per akite) forecasts fra analytikere for FY1, FY2 og FY3. De sidste seks variable er sentimentbaserede, og er taget med inspiration fra Daul et al. (2022), hvor DY generelt har viste gode resultater i litteraturen.

Appendiks C

Det skandinaviske aktiemarked

Det skandinaviske aktiemarked er et relativt nyt marked sammenlignet med f.eks. det amerikanske. Det begyndte at vokse hurtigt i 1990'erne, da økonomierne begyndte at vokse (Grobys & Huhta-Halkola 2019). Norske aktier handles på Oslo Børs med 196 børsnoterede selskaber. Olie, shipping og skaldyr er de største industrier i Norge, og kendte virksomheder omfatter Lerøy Seafood Group, Norwegian Air Shuttle og Equinor (Oslo-Børs 2023). Børsnoterede aktier i Danmark handles på Nasdaq. Per maj 2023 er der 133 børsnoterede selskaber på Københavns Fondsbørs (Nasdaq-Nordic 2023). Nøglebrancherne i Danmark omfatter cleantech, såsom vindkraft, og pharma og biotek, som bl.a. er repræsenteret af Vestas og Novo Nordisk. Novo Nordisk er klart det største individuelle selskab i datasættet, som udgør meget store procenter af den samlede markedsværdi i Danmark. Endelig repræsenterer Sverige det største marked af de tre lande. Svenske aktier er også noteret på Nasdaq, og der er i alt 383 børsnoterede selskaber per maj 2023 (Nasdaq-Nordic 2023). Store virksomheder i Sverige er Ericsson, Volvo og Hennes & Mauritz, der også repræsenterer nogle af Sveriges nøgleindustrier: bilindustrien og materialevidenskabsindustrien. De præsenterede nøgleindustrier stemmer overens med datasættet, dertil er også særligt banksektoren meget repræsenteret sammen med software/computer services.

Tabel C.1 viser en sammenfattende statistik for det endelige datasæt, hvor det tydeligt fremkommer, at Sverige er det største af de tre markeder, mens markedsværdien på det enkelte selskab (i gennemsnit) er størst på det danske marked - formodentligt stærkt påvirket af Novo Nordisk. Baseret på de analyserede markedsværdidata blev det desuden konstateret, at de mindre selskaber i datasættet er repræsenteret over færre måneder, mens større selskaber er repræsenteret i flere måneder. Dette resultat er i overensstemmelse med forventningerne, da større selskaber (typisk) har en længere historik med at være børsnoterede men også er mere modstandsdygtige

over for økonomiske vanskeligheder sammenlignet med mindre. Hver måned er et minimum af 161 selskaber aktive og inkluderet i modellerne for at identificere tværsnitsforskelle.

Tabel C.1 Sammenfattende statistikker efter de anvendte markeder

Marked	Totale antal	Min antal	Mean antal	Max antal	Mean size	Median size	Totale size	% af totale
Danmark	220	36	66	122	3.781	1.310	174.557	26,48
Norge	394	56	82	139	1.819	481	150.739	22,87
Sverige	466	69	119	228	2.665	803	333.885	50,65

Note: Size er opgjort i euro.

Appendiks D

Hypertuning

Det primære fokus er at minimere fejlen (MSE) fra valideringssættet (modelscoren), der fungerer som en indikator for modellens ydeevne på nye, usete data. Ved at vælge hyperparametre, der fører til lavere valideringsfejl, sigtes det mod at forbedre modellens generaliseringsevne. Men også validerings- og træningsfejlene er blevet undersøgt. Ved at overveje både modelscoren og trænings-/valideringsfejlene sigter valget af de bedste hyperparameterværdier på at finde en balance mellem modelkompleksitet og generaliseringsevne. Målet er at finde det sæt af hyperparametre, der maksimerer ydeevnen på usete data samtidig med, at modellen undgår over- eller undertilpasning. Generelt blev det fundet, at det var relevant med brede gitterrum, fordi kompleksiteten var varierende over t. Hyperparameterjusteringen har været er en iterativ proces med henblik på at forfine søgerummet.

I tilfældet af RF viste hyperparametrene M og L at have den største indflydelse på nøjagtigheden/modelscoren. I litteraturen er valget af M fremlagt som en klassisk bias-varians afvejning. Hvor den generelle tendens, at hvis M reduceres, øges bias. Det vil sige, at en forøgelse af M kan øge modellens fleksibilitet, reducere bias, men også øge risikoen for en overtilpasning (Hastie et al. 2009). Derfor er der blevet eksperimenteret med forskellige værdier, herunder standardindstillingen ($\sqrt{p}$) men også høje og lave værdier. De høje værdier blev forsøgt, fordi de ifølge litteraturen har en tendens til at performe godt ved mindre datasættet, da det øger chancen for netop at fange det stærkeste signal. Forskellige værdier blev udforsket med inspiration fra litteraturen (Gu et al. 2020, Drobetz & Otto 2021, Hanauer & Kalsbach 2023, Leippold et al. 2022). Det viste sig her særlig relevant at have et bredere spektrum af værdier, hvilket også er afspejlet i den endelige kombination, og det blev fundet med særlig inspiration fra værdierne hos Gu et al. (2020). Med hensyn til hyperparameteren L blev forskellige værdier også testet. Litteraturen foreslår gode

resultater med standardindstillingen[1], men efterhånden som M øges, kan det være nødvendigt at reducere L (træets kompleksitet). Hvor høje værdier for L førte til overtilpasninger, var det vigtigt at finde den rigtige balance. At øge B over 400 reducerede ikke modelscoren væsentligt, hvis overhovedet, hvilket også er et veldokumenteret faktum (Hastie et al. 2009). Af processen blev det bekræftet at, RF er en mere robust algoritme, der er sværere at tune, som dens MSE ikke ændrede sig væsentligt afhængigt af værdierne[2]. Der var dog væsentlige forbedringer at hente sammenlignet med standardparameterværdierne. Det resulterede i parametrene præsenteret i tabel 4.1. GB har de tre hyperparametre (L, υ, B), der styrer bias-varians afvejningen. Ofte fungerer meget lave værdier godt til L, som f.eks. $L = 1$ (og i så fald er ethvert træ en stub, der består af kun én opdeling). Et lavt L kan medføre højt bias, mens et højt L kan medføre høj varians. I afhandlingen blev det bekræftet, at de lave værdier for L resulterede i de laveste modelscores. υ er også ofte et lille positivt tal, der styrer hastigheden, hvormed boosting lærer. Typiske værdier er mellem 0,001-0,2, men det rigtige valg er naturligvis kontekstafhængigt, som det at øge υ kan medføre, at modellen bliver for sensitiv over for træningssættet, mens reduceret υ omvendt er forbundet med øget beregningsmæssig kompleksitet. Et meget lille υ kan endog kræve brug af et stort B for at opnå en god prædikteringsevne. I modsætning til bagging og RF, kan boosting overtilpasse, hvis B er for stort, selvom denne overtilpasning har en tendens til at forekomme langsomt, hvis overhovedet (Hastie et al. 2009). For GB viste υ sig som den med størst indflydelse. De helt lave υ viste sige ikke lige så gode som de højere værdier, hvorfor der blev arbejdet videre med de høje værdier med henblik på at finde et godt interval og et passende B. Forskellige værdier blev afsøgt, og alle de forsøgte kombinationer gav også her forbedringer over standardparameterværdierne[3] og resulterede i værdierne i tabel 4.1.

Målet med hyperparametertuningen er at minimere valideringsfejlen og sigte mod, at den kun er marginalt højere end træningsfejlen. Dette sikrer, at algoritmen kan generalisere godt til nye, usete data. En overtilpasning er tilfældet, hvis valideringsfejlene er høje, mens træningsfejlene er lave. Omvendt sker en undertilpasning, hvis både validerings- og træningsfejlene er høje. Det betyder, at modellen er for enkel og har højt bias, resulterende i en dårlig ydeevne, som følge af manglende evne til at fange de underliggende mønstre i dataene (Hastie et al. 2009).

[1] For RF er disse: $B = 100$, $M = \sqrt{p}$ og $L =$ None.

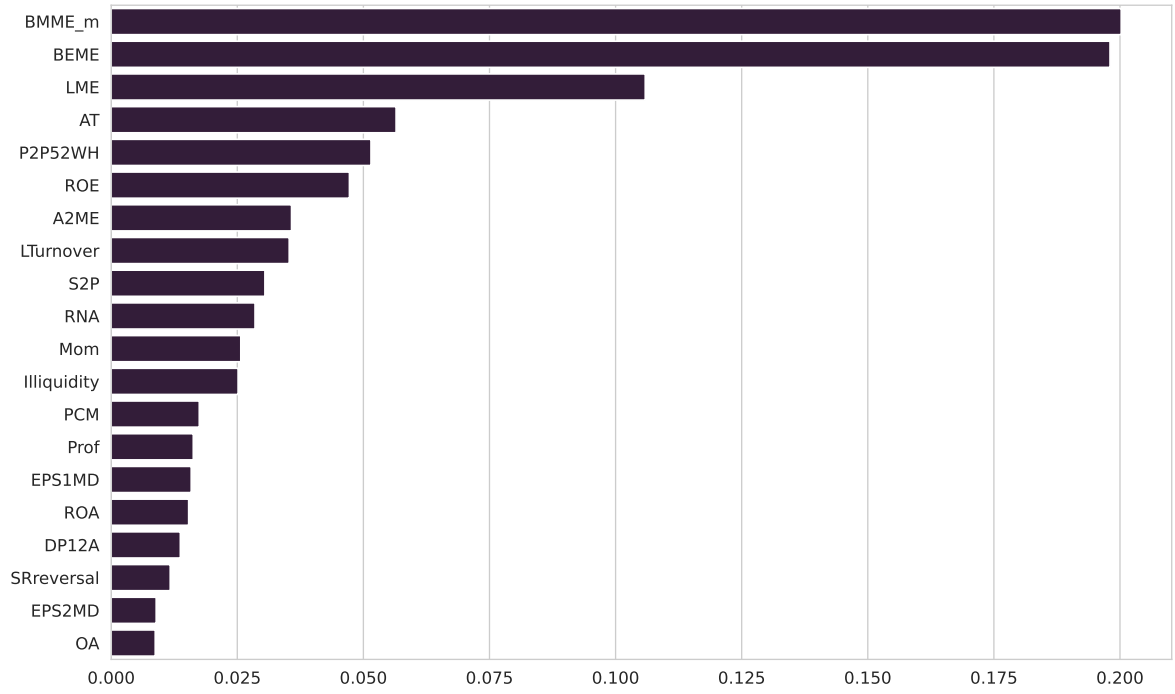
[2] Eventuelle årsager hertil kan være utilstrækkeligt hyperparameterområde, iboende begrænsninger i modellen, data-relaterede problemer, over-/undertilpasning eller utilstrækkelige data.

[3] For GB er disse: $B = 100$, $\upsilon = 0,1$ og $L = 3$.

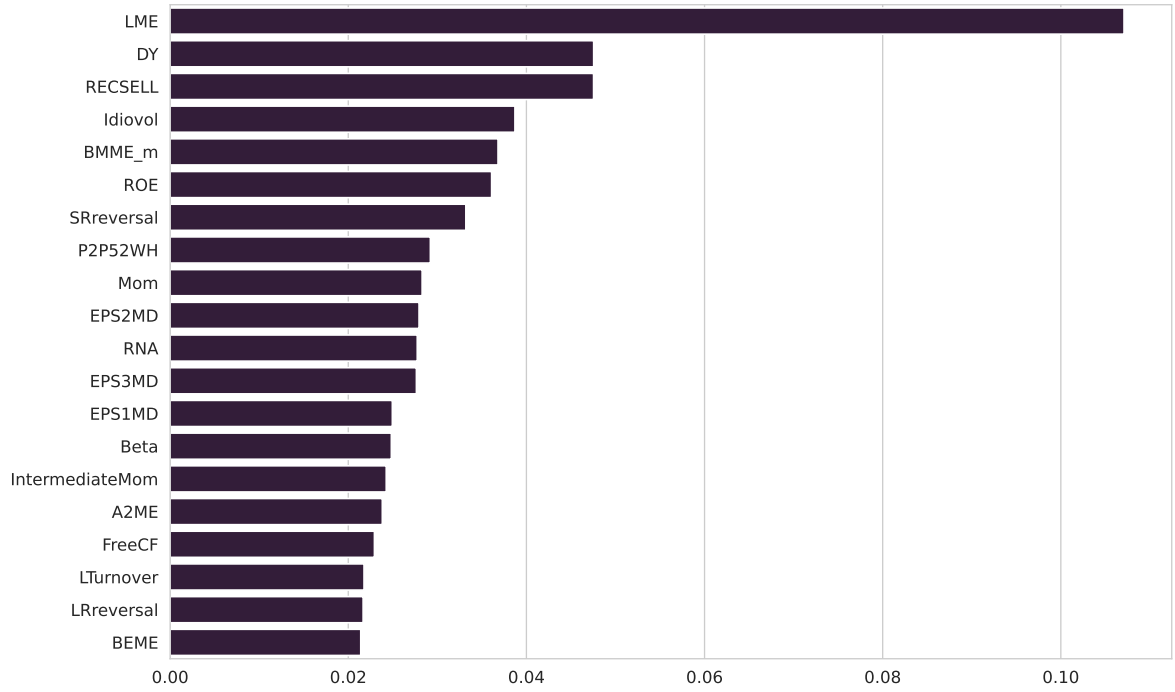
Appendiks E

Variables betydning

Figur E.1 Variables (relative) betydning for OLS-algoritmen

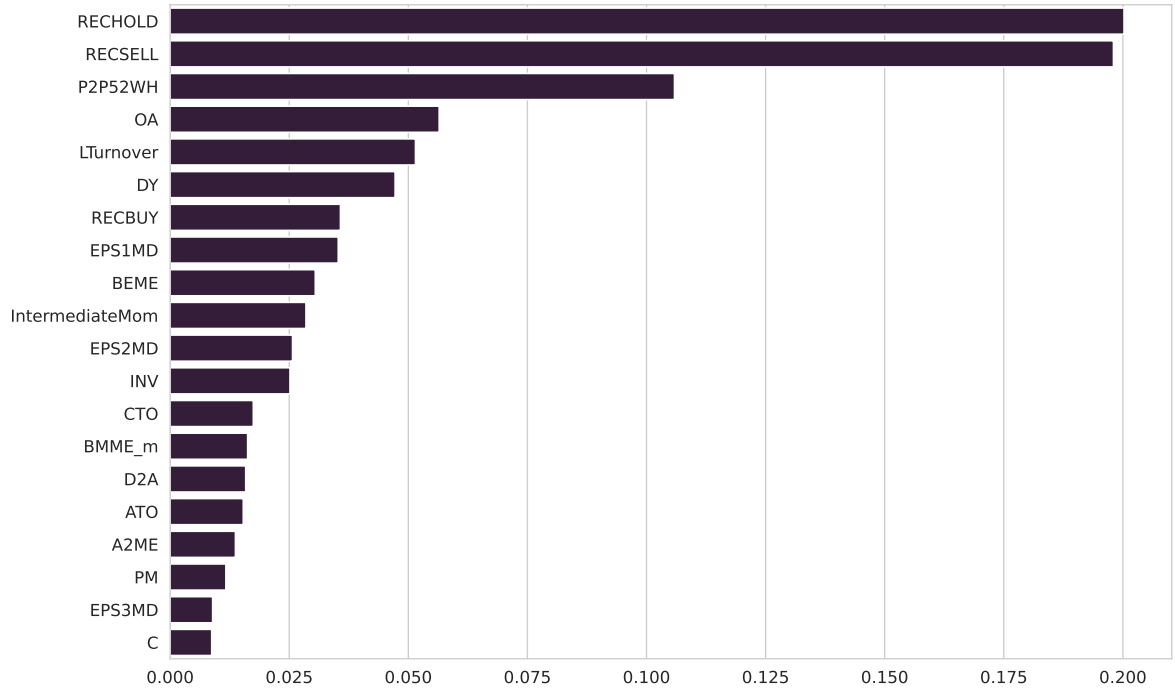


Figur E.2 Variables (relative) betydning for RF-algoritmen



Note: For RF beregnes variablenes betydning ved at aggregere de individuelle betydninger. Den kvantificerer den totale reduktion i objektiv funktionen opnået ved at opdele på en bestemt variabel på tværs af alle træer. Variable, der konsekvent giver de største forbedringer, anses for at være vigtigere.

Figur E.3 Variables (relative) betydning for GB-algoritmen



Note: I GB bestemmes betydningen af det kumulative bidrag fra den enkelte variabel gennem de boostende iterationer. Ved hver iteration retter et nyt træ fejlene fra de tidligere træer. Betydningen af en variabel måles ved at summere forbedringerne i objektive funktionen opnået af variablen ved hver iteration. Variable, der konsekvent giver de største forbedringer, anses for at være vigtigere.

Appendiks F

Supplerende til: Empiriske resultater

Tabel F.1 Ligevægtede drawdowns, omsætning og risikojusteret præstation af porteføljerne

	OLS	RF	GB	ENS$^{(1)}$	ENS$^{(2)}$
Panel A: Kvantilsorterede porteføljeperformance (ligevægtet)					
Gns. [%]	4,01	2,81	3,79	3,68	4,22
Gns. Ann. [%]	60,42	39,58	56,44	54,38	64,19
Std. Ann. [%]	12,73	10,42	11,57	11,02	11,92
SR Ann.	3,78	3,24	3,94	4,01	4,24
Max DD [%]	-5,64	-5,19	-3,06	-3,28	-2,82
Max 1M Tab [%]	-5,05	-5,18	-3,05	-2,74	-2,82
TO	114,47	90,68	126,87	113,02	110,43
Panel B: Risikojusteret performance (ligevægtet)					
α_{FF6} [%]	3,85	2,77	3,82	3,69	4,09
t-stat$_\alpha$	14,68	13,29	16,42	16,47	16,88
R^2 [%]	3,12	9,5	8,39	6,04	5,6

Note: Tabellen viser den ligevægtede out-of-sample præstation af H-L-spændet. Panel A opsummerer det gennemsnitlige månedlige merafkast (Gns.), annualiseret merafkast, standardafvigelse og Sharpe-Ratio, men også maksimale drawdown (Max DD), det mest negative månedlige tab (Max 1M tab) og gennemsnitlige omsætning (TO). Panel B rapporterer det gennemsnitlige alfa (α_{FF6}) fra seksfaktormodellen, tilsvarende Newey-West justeret t-statistik med 4 lags (t-stat$_\alpha$) og den tilsvarende R^2.

F.1 Robusthedstest

RAbex er designet til at fange det fremherskende niveau af risikoaversion på de finansielle markeder, hvilket afspejler investorernes villighed til at påtage sig risiko. Dens konstruktion involverer

brugen af en bred vifte af finansielle markedsdata, der inkorporerer information fra flere aktivklasser og markedsindikatorer. Ved at inkorporere RAbex i evalueringsprocessen kan ydeevnen af algoritmer vurderes under forskellige risikoaversionsregimer, hvilket fører til en mere omfattende forståelse af deres robusthed. Desuden muliggør denne integration identifikation af potentielle sammenhænge mellem risikoaversion og afkast, og bidrager derfor til en dybere forståelse af den underliggende dynamik på de finansielle markeder. Tilsvarende tjener inkluderingen af SENT et sammenligneligt formål ved at fange den overordnede stemning (eller stemning hos investorer) på de finansielle markeder. SENT er konstrueret ved at trække på en række forskellige kilder, såsom undersøgelser, medierapporter og sentiment-relaterede data. Disse kilder giver værdifuld indsigt i investorers holdninger, overbevisninger og adfærdsmæssige biases, der kan påvirke deres investeringsbeslutninger. Indekset har til formål at fange det psykologiske aspekt af markedsdeltagere og deres følelsesmæssige disposition over for forskellige investeringsmuligheder. Samlet set forbedrer brugen af både RAbex og SENT i evalueringen af porteføljepræstationerne forståelsen af sammenhængene mellem risikoaversion, sentiment og afkast (genereret fra algoritmerne).

Tabel F.2 Høj-risiko versus lav-risiko måneders H-L-afkast

Model	Mkt_{up}	Mkt_{down}	RAbex_{high}	RAbex_{low}	SENT_{high}	SENT_{low}
Panel A: Ligevægtet						
OLS [%]	3,99	4,21	4,40	3,65	4,17	3,67
RF [%]	2,43	3,55	3,31	2,33	2,77	2,76
GB [%]	3,57	4,22	4,57	3,04	3,79	3,63
$\text{ENS}^{(1)}$ [%]	3,42	4,38	3,0	2,89	3,62	3,62
$\text{ENS}^{(2)}$ [%]	3,94	4,83	3,62	3,57	4,25	4,03
Panel B: Værdivægtet						
OLS [%]	2,74	3,76	3,59	2,61	3,64	2,42
RF [%]	1,64	2,17	1,96	1,69	2,24	1,35
GB [%]	2,89	2,88	3,68	2,10	3,24	2,39
$\text{ENS}^{(1)}$ [%]	2,47	3,43	3,50	2,11	3,14	2,32
$\text{ENS}^{(2)}$ [%]	2,89	4,02	4,04	2,45	3,52	2,81

Note: De anvendte risiko-proxyerne er, hvorvidt det skandinaviske aktiemarked som helhed går op eller ned (Mkt), hvorvidt RAbex og SENT er større eller mindre end deres median.

Appendiks G

Supplerende til: Algoritmer og metodologi

G.1 OLS-antagelser

OLS-estimationsmetoden benyttes, da OLS-estimatorerne er de bedste[1] lineære unbiased[2] estimatorer (BLUE) under en række antagelser i relation til data (Wooldridge 2019). Antagelserne betegnes samlet Classical Linear Model- (CLM-) antagelserne.

Antagelse (1) Linearitet i parametrene - lineært forhold mellem den afhængige og de uafhængige variable.

Antagelse (2) Stikprøven er tilfældig.

Antagelse (3) Ingen perfekt kolinearitet, hvilket indebærer, at ingen af de forklarende variable er konstante eller perfekt korrelerede med hinanden. Ved perfekt kolinearitet forekommer uendeligt mange OLS estimater, hvilket betyder, at regressionen ikke kan estimeres. Desuden er estimaterne forbundet med høj usikkerhed i tilfælde af høj multikolinearitet (Wooldridge 2019).

Antagelse (4) Betinget gennemsnit er nul, $E(\varepsilon_t \mid \boldsymbol{X}) = 0$, hvilket indikerer et korrekt specificeret forhold mellem den afhængige og de uafhængige variable.

OLS er unbiased ved opfyldelse af antagelserne **1 til 4**.

[1] Bedste betyder her, at OLS estimatorerne har mindst mulig varians blandt alle lineære, unbiased estimatorer, hvilket betyder, at $var(\hat{\beta}_j) \leq var(\tilde{\beta}_j)$. Er samtlige Classical Linear Model-antagelser opfyldt har OLS-estimatorerne tilmed den mindste varians af alle unbiased estimatorer (Wooldridge 2019).

[2] Unbiased estimatorer betyder, at disse er centreret omkring den sande underliggende tidsserieproces (Wooldridge 2019).

Antagelse (5) Homoskedastiske fejlled, hvilket betyder, at:

$$var(\varepsilon_t \mid \boldsymbol{X}) = var(\varepsilon_t) = \sigma^2, \quad t = 1, 2, ..., T \tag{G.1}$$

hvor variansen af fejlledet, ε_t, er den samme for alle tidsperioder, t, betinget af de forklarende variable, $\boldsymbol{X}$. Opfyldes denne antagelse ikke, er fejlledende heteroskedastiske, hvilket betyder, at $var(\varepsilon_t)$ varierer afhængigt af forklarende variables værdier (Wooldridge 2019).

OLS er BLUE ved opfyldelse af antagelserne **1 til 5**.

Antagelse (6) Ukorrelerede fejlled, hvilket omhandler mere konkret ingen autokorrelation, hvilket betyder, at fejlleddene i to forskellige tidsperioder er ukorrelerede betinget af $\boldsymbol{X}$:

$$corr(\varepsilon_t, \varepsilon_s \mid \boldsymbol{X}) = 0 \tag{G.2}$$

for alle $t \neq s$. Opfyldes antagelsen ikke undervurderes den sande varians af OLS-estimatoren, hvormed estimatet præcision overvurderes (Wooldridge 2019).

Slutteligt indebærer CLM-antagelserne en antagelse omkring normalfordelte fejlled, det vil sige: $\varepsilon_t \sim N(0, \sigma^2)$, der er nødvendig for hypotesetest (Wooldridge 2019).

G.2 Faktormodeller

RMRF-faktoren angiver merafkastets følsomhed overfor markedsporteføljens volatilitet. Hvis $\beta_1 = 1$ modsvarer afkastets samt markedets risikopræmie hinandens, mens $\beta_1 < 0$ medfører, at afkastet realiserer en risikopræmie, der er lavere end markedets, og omvendt ved $\beta_1 > 1$ (Hillier et al. 2012). SMB-faktoren foreligger relativt høje afkastpræmier for små selskaber uafhængigt af volatilitet (Banz 1981) og er dermed afkastforskellen mellem selskaber med lave og høje markedsværdier. Basu (1977) finder HML-faktoren, og at selskaber med lave Price/Earnings (P/E-) ratioer repræsenterer relativt høje risikojusterede afkast. RMW-faktoren er afkastforskellen mellem selskaber med henholdsvis robust og svag rentabilitet, mens CMA-faktoren er afkastforskellen mellem selskaber med henholdsvis konservative og aggressive investeringsstrategier (Fama & French 2015). WML-faktoren beskriver en højere afkastpræmie for aktier med høje forudgående afkast end for aktier med lave forudgående afkast. Denne faktor besidder empirisk robusthed, men dens årsagsforklaring debatteres fortsat, omend den i høj grad beskrives af psykologiske faktorer med afsæt i behavioral finance (Jegadeesh & Titman 1993).