KØBENHAVNS UNIVERSITET
INSTITUT FOR STATSKUNDSKAB

Kandidatspeciale

Alexander Hamann Bach og Jesper Svejgaard Jensen

Maskinlæring på skolebænken

En undersøgelse af maskinlæring anvendt til målretning af tiltag mod uddannelsesfrafald på Professionshøjskolen Metropol

Vejleder: Frederik Georg Hjorth

Afleveret den: 1. august 2017

Institutnavn:	Institut for Statskundskab
Forfattere:	Alexander Hamann Bach og Jesper Svejgaard Jensen

Titel:	Maskinlæring på skolebænken
Undertitel:	En undersøgelse af maskinlæring anvendt til målretning af tiltag mod uddannelsesfrafald på Professionshøjskolen Metropol
Vejleder:	Frederik Georg Hjorth
Afleveret den:	1. august 2017
Antal tegn:	329.191

Maskinlæring på skolebænken

En undersøgelse af maskinlæring anvendt til målretning af tiltag mod uddannelsesfrafald på Professionshøjskolen Metropol

Alexander Hamann Bach og Jesper Svejgaard Jensen

August 2017

Abstract

Machine learning has emerged from its computer science cradle and entered the spheres of empirical social science and public policy making. In this thesis, we explore the merits and pitfalls of applying machine learning with the aim of targeting policy interventions based on individual risk predictions. We do so by investigating the potential of using machine learning to predict students' dropout risk and target interventions at Metropolitan University College. We provide an overview of the debates concerning applied machine learning. Introducing central machine learning concepts, we emphasise the essential differences following from our occupation with prediction rather than estimation. In our analysis, we build and compare four different machine learning algorithms and delve into the workings of our best model based on Gradient Boosted Trees, which obtains an AUC of 0.727. Next, we trace out a stylized framework for applying the prediction model in practice. We conclude that machine learning has a potential to target interventions against dropout. We argue, however, that implementing the prediction model in an actual administrative context incurs challenges of both methodological, epistemic, and ethical nature, which previous studies fail to account properly for. Fleshing out these challenges, we revise our conclusion pointing to the considerable gap between technically predicting an outcome and translating these predictions into a sound basis for action in any administrative context across policy domains.

Keywords: Machine learning, prediction, student dropout, targeting policy interventions, Metropolitan University College.

Forord

Dette speciale tog sin begyndelse i en interesse for den stigende anvendelse af maskinlæring i den offentlige forvaltning. Vores case er anvendelse af maskinlæring på Professionshøjskolen Metropol. Dette samarbejde kom i stand via kontorchef Jens Storm i Uddannelses- og Forskningsministeriet, som pegede os i retning af et pilotprojekt om at forudsige frafald på Metropol. Her havde Metropol sammen med konsulenthuset Deloitte taget de første skridt mod at udvikle en prædiktionsmodel på statiske frafaldsdata.

Vi skylder først og fremmest en stor tak til Metropol, som har stillet deres data og faciliteter til rådighed, mod at vi udviklede en prædiktionsmodel direkte i deres administrative systemer. Vi har fået praktisk bistand af flere medarbejdere og har samtidig haft frie rammer til at arbejde med modellen samt publicere resultater. Særligt vil vi takke konsulent Kåre Degn, der har en stor del af æren for samarbejdets konstruktive rammer. Han har været vores primære kontaktperson og løbende ryddet praktiske udfordringer af vejen. Vi vil desuden takke konsulent Andreas Keller Leth for sparring på kodearbejdet i R. I denne henseende står vi også i gæld til open source-fællesskabet på StackOverflow, der altid har haft rede svar, når koden gjorde kvaler.

For inspiration til specialets problemstilling vil vi gerne rette en tak til Oskar Harmsen og Thyge Ryom Enggaard, som fra et økonomisk perspektiv har undersøgt maskinlærings potentiale til målretning i sundhedsvæsenet.

Endelig vil vi gerne takke vores vejleder Frederik Hjorth for kløgtig og konstruktiv feedback hele vejen igennem processen.

Indhold

Alexander Hamann Bach har haft hovedansvar for sektionerne:
2.1, 2.3.1-2, 3.1.1, 3.2.1-2, 3.2.7, 3.3.3-4, 4.1.1, 4.1.4, 4.2.1, 5.1, 5.3, 5.4.1, 6.1.2, 6.2.2, 6.3.1, 6.3.3, 6.4.1, 6.4.2

Jesper Svejgaard Jensen har haft hovedansvar for sektionerne:
2.2, 2.3.3, 3.1, 3.1.2, 3.2.3-6, 3.3.1-2, 4.1.2-4.1.3, 4.2.2, 5.1.1, 5.2, 5.4, 5.4.2, 6.1.1, 6.2.1, 6.3.2, 6.4.3, 6.4.4

Begge har forfattet kapitlerne 1 og 7 og har desuden bidraget omfangsrigt til hinandens sektioner.

Figurer

Tabeller

Kapitel 1

Indledning

Maskinlæring stammer fra datalogien og har frembragt teknologiske landvindinger som spamfiltre, talegenkendelse og selvkørende biler. Betegnelsen maskinlæring dækker over en vifte af tilgange til databehandling, hvor algoritmer er udviklet til at kortlægge komplicerede mønstre i store datamængder. Set i lyset af denne egenskab er det ingen overraskelse, at maskinlæring som tilgang har spredt sig fra datalogiske til samfundsvidenskabelige institutter, og fra IT-virksomheder til den offentlige forvaltning, hvor efterspørgslen på analyse og evidens er stor (Rieder & Simon 2016; Dahler-Larsen 2011).

Denne efterspørgsel er ikke ny. Lige fra de moderne nationalstaters fødsel har magthavere og embedsmænd efterspurgt data, de kunne føre politik og administrere ud fra. Den statistiske videnskab er uløseligt forbundet med opbygningen af en centraliseret administration (Foucault 2007; Rose 1991). Det nye er, at den eksplosive vækst i data og computerkræfter radikalt har ændret mulighederne for, hvordan vi kvantitativt kan undersøge samfundet (Imai 2017; James et al. 2013). En særlig interesse samler sig om muligheden for at bruge maskinlæring til forudsigelse (Hofman et al. 2017). Idet algoritmerne i maskinlæring giver mulighed for at modellere data mere fleksibelt end traditionelle statistiske metoder, åbner de muligheden for at finde komplekse mønstre i data ned på individniveau. Individualiserede forudsigelser giver mulighed for en mere præcis målretning af politiske tiltag.

Maskinlæring er fx forsøgt anvendt på sundhedsområdet til at forudsige tilbagefald blandt patienter for at muliggøre en differentieret behandling (Harmsen & Enggaard 2016). På retsområdet er der i USA udført forsøg med at forudsige sandsynligheden for, at en sigtet vil begå flere forbrydelser, for at assistere dommere i beslutninger om vare-

tægtsfængsling og prøveløsladelse (Berk 2012). På uddannelsesområdet er maskinlæring blevet brugt til at forudsige frafald på danske gymnasier (Şara 2014). Eksemplerne er mange, og de illustrerer maskinlærings begyndende indtog i samfundsvidenskaben og forvaltningen. Spørgsmålet er dog, om hypen er berettiget. Er gevinsterne ved tilgangen kun teoretiske, eller rummer den et potentiale i forvaltningen i praksis? Kan målretningen af tiltag have utilsigtede konsekvenser? Og er det overhovedet etisk forsvarligt at forskelsbehandle individer på baggrund af algoritmer?

Det er spørgsmål som disse, der trænger sig på, når en potentielt omkalfatrende tilgang til databehandling – og som følge heraf borgerbehandling – bevæger sig fra sit videnskabelige ophav i datalogien og ind i samfundsvidenskaben og forvaltningen. Og det er spørgsmål som disse, vi i denne opgave vil belyse i en case-undersøgelse, hvor vi tager afsæt i følgende problemstilling:

Har maskinlæring potentiale til at målrette tiltag mod uddannelsesfrafald?

Vi vil besvare problemstillingen med udgangspunkt i en maskinlærings-model til forudsigelse af frafald. Vi udvikler modellen i samarbejde med Professionshøjskolen Metropol, som vil anvende modellen til at målrette tiltag mod individuelle studerende.

Når vi i problemstillingen skriver, at vi vil undersøge *potentialet* ved maskinlæring, har vi to ting for øje. For det første vil vi analysere potentialet ud fra metodiske kriterier, hvilket handler om, hvor gode forudsigelser, modellen kan levere. For det andet vil vi diskutere modellens potentiale under hensyn til de konsekvenser, det kan have at anvende den i praksis.

Konkret vil vi i analysen sammenligne fire forskellige maskinlærings-algoritmer. Vi går i dybden med den, der leverer de mest præcise forudsigelser, og demonstrerer, hvordan modellen kan tages i anvendelse. De fire algoritmer, vi sammenligner, er logistisk regression, klassifikationstræer, *Random Forest* og *Gradient Boosted Trees*. Her repræsenterer logistisk regression en traditionel måde at lave forudsigelser på i samfundsvidenskaben. Anderledes repræsenterer de øvrige algoritmer nybrud inden for kvantitative metoder i samfundsvidenskaben. Vores analytiske fokus i specialet er således på maskinlærings metodiske potentiale.

Analysen giver anledning til en bredere diskussion af de konsekvenser, som følger, når maskinlæring tages i anvendelse i praksis. Det er omdrejningspunktet i vores diskussion, hvor vi belyser en række metodiske, epistemiske såvel som etiske dilemmaer ved tilgangen. Det er i lyset af disse, at vi diskuterer, om potentialet ved målretning

står mål med konsekvenserne. Hermed tager vi et skridt videre fra den klassiske maskinlæringslitteratur, som primært stammer fra datalogien og fokuserer på tilgangens tekniske aspekter. Med vores speciale skriver vi os ind i den spæde, men voksende politologiske litteratur om maskinlærings potentiale i den offentlige forvaltning.

I det følgende afsnit vil vi udfolde nogle metodiske overvejelser om vores valg af undersøgelsesdesign og valg af case. Dernæst følger en begrebsafklaring og endelig et overblik over opgavens opbygning.

1.1 Casevalg, teori og undersøgelsesdesign

Specialets kerne er en frafaldsmodel, som vi udvikler i et samarbejde med Professionshøjskolen Metropol. Modellen er baseret på en stor mængde data om ca. 23.000 studerende ved Metropol i perioden 2009-2017. Modellens formål er at forudsige studerendes individuelle risiko for at frafalde. Det kan anvendes af Metropols studievejledning til at målrette tiltag mod frafald, såsom visitationsmøder og individuelle mentorforløb. Casen giver os mulighed for at undersøge potentialet til at målrette tiltag baseret på maskinlæring i en konkret, administrativ praksis. Med opgaven søger vi dermed at undersøge en metodisk problemstilling med en konkret forvaltningscase.

Emnet i vores case er uddannelse og uddannelsesfrafald. Det er en klassisk problemstilling, hvordan et velfærdssamfund investerer i og sætter rammerne for udbuddet af uddannelse. Langt de fleste borgere i et samfund som det danske har selv modtaget uddannelse, ligesom deres forældre har og deres børn har udsigt til – uddannelse har derfor både stor betydning på mikroplan for den enkelte, såvel som på makroplan for samfundet som helhed. Uddannelsesfrafald er derfor betydningsfuldt og interessant i sig selv.

Ud over at have betydning for den enkelte og samfundet som helhed har frafald også en stor økonomisk betydning for Metropol, som hovedsageligt er finansieret via afregning per gennemførte studieår (Metropol 2017). Med vores metodiske fokus afgrænser vi os dog fra mulige organisationsteoretiske perspektiver på vores problemstilling. Derfor går vi fx ikke i dybden med Metropols motiver til at reducere frafaldet, men anskuer det som en ramme, der er givet.

Opgavens teoretiske fundament består af metodisk litteratur, som dels er klassisk politologisk, dels stammer fra datalogien. Centralt i teoriapparatet står en metodisk distinktion mellem estimation og prædiktion, idet vi anvender maskinlæring til at forudsige frem for at forklare frafald, som er en mere vanlig tilgang i samfundsvidenskaben.

Derudover vil vi i diskussionen bringe en politisk teoretisk litteratur i spil, som på forskellig vis kan bruges til at problematisere anvendelsen af maskinlæring i komplekse sociale kontekster. Det er en litteratur, som rejser spørgsmål af videnskabsteoretisk og etisk karakter, som vi mener vil finde resonans på tværs af policy-områder.

I opgaven går vi i dybden med Metropol som case på en uddannelsesinstitution. I hovedtræk deler danske uddannelsesinstitutioner rammebetingelser, og her skiller Metropol sig ikke ud. Datagrundlaget varierer selvsagt mellem institutionerne, men grundlæggende er dataindsamlingen standardiseret i Danmark, og alle professionshøjskoler deler fx administrativt system. Vi anser derfor Metropol for at være repræsentativ for en bredere kreds af uddannelsesinstitutioner og dermed en *typisk case* (Seawright & Gerring 2008: 296–300). I det omfang vi finder et potentiale til målretning af tiltag mod uddannelsesfrafald i vores case, forventer vi derfor at kunne generalisere potentialet til andre uddannelsesinstitutioner.

Implikationerne af casen rækker samtidig ud over uddannelsesområdet. I første instans er Metropol som nævnt ovenfor en case på en uddannelsesinstitution, som ønsker at målrette tiltag mod uddannelsesfrafald. I anden instans kan Metropol endvidere ses som en case på et mere generelt fænomen: en offentlig institution, som ønsker at målrette et tiltag. Ved at fokusere på dette generelle træk ved casen udvider vi betragteligt den kreds af enheder, vi kan generalisere til.

Samtidig er det dog klart, at der er grænser for, hvor vidtgående konklusioner vi kan drage, når kredsen af enheder udvides fra uddannelsesinstitutioner til en bredere gruppe af institutioner, der kun deler nogle formelle karakteristika (Gerring 2004: 347–348). Der kan fx være væsensforskelle på at forudsige uddannelsesfrafald og kriminalitet, selvom både skoler og politi er offentlige institutioner. Derfor er vi nødt til at medtænke betydningen af den konkrete policy-kontekst for generaliserbarheden. Med begrebet policy-kontekst refererer vi til de politiske hensyn, som sætter rammerne for den enkelte case. Det er de rammer, som en beslutningstager må forholde sig til i udformningen og målretningen af konkrete tiltag som led i at indfri en politisk målsætning. Det kan både være økonomiske rammer og hensyn til forskellige politiske værdier og principper om fx ligebehandling. I opgavens diskussionsafsnit vender vi tilbage til refleksionerne omkring generaliserbarhed.

Der er endnu relativt få casestudier og dermed begrænset viden om målretning af tiltag mod uddannelsesfrafald med maskinlæring. I vores analyse af casen Metropol anlægger vi derfor en eksplorativ tilgang. Dertil er casestudiet som undersøgelsesdesign særligt velegnet – eksemplets kraft kan have stor betydning for udviklingen af ny viden (Flyvbjerg 2006: 228). Casestudiet som design sætter dog også nogle begrænsninger

for, hvilken type af konklusioner, vi kan drage. Casestudier er oftest dårligt egnede til at afprøve hypoteser, og vi kan derfor ikke be- eller afkræfte en hypotese om maskinlærings potentiale (Gerring 2004). Med casedesignet kan vi derimod nuanceret belyse forskellige aspekter af maskinlæring, som kan have betydning for potentialet til at målrette policy-tiltag. Den dybdegående og detaljerede undersøgelse er casestudiets styrke og kan føre til indsigter, som kan generaliseres til lignende cases (Gerring 2004: 349–350). Casestudier, som kommer tæt på den virkelighed, de forsøger at beskrive, er en forudsætning for en grundig forståelse af et fænomen (Flyvbjerg 2006: 236). Vores argument er således, at et dybdegående, kontekstualiseret casestudie er velegnet til at konkretisere og belyse styrkerne og svaghederne ved maskinlæring som tilgang.

1.2 Begrebsafklaring

I dette afsnit vil vi kort afklare to af specialets centrale begreber: maskinlæring og algoritmer.

Maskinlæring Maskinlæring dækker over en tilgang til at finde mønstre i data, som siden 1950'erne er blevet forfinet inden for datalogien. En tidlig definition på maskinlæring blev givet i 1959, hvor maskinlæring forstås som computeres evne til at lære uden at være eksplicit programmeret til det (Samuel 1959). I maskinlæring anvendes en række konkrete metoder, som i datalogien omtales algoritmer. Nogle af metoderne er nye i samfundsvidenskaben, mens andre er velkendt inventar i den politologiske værktøjskasse, såsom lineære og logistiske regressionsmodeller (Friedman et al. 2009; Murphy 2012). Selvom nogle af algoritmerne er nye i samfundsvidenskaben, ligger nybruddet ikke i selve algoritmevalget, men i den måde de anvendes på.

Når vi bruger betegnelsen maskinlæring i denne opgave, refererer vi derfor til en bestemt tilgang, som har spredt sig fra datalogien. Med denne tilgang er formålet ikke at forstå eller forklare mønstrene i data, men at bruge dem til at komme med prædiktioner om nye data (Athey & Imbens 2016: 41–44; Mullainathan & Spiess 2017: 88). Maskinlærings læringselement består i, at modellerne selv "lærer" mønstre af data frem for, at vi på forhånd begrænser processen ved at udvælge variable og teoretisk specificere, hvordan vi tror eller antager, at variablene i et datasæt hænger sammen. Læringen består endvidere i, at modellerne over tid selv kan tilpasse sig nye data og dermed potentielt forbedre deres forudsigelser af fremtidige outcomes.

Vi kan illustrere tilgangen med vores egen case. Her er målet ikke at forklare eller forstå årsagerne til frafald på Metropol. Vi forsøger i stedet at udvikle en model, der kan opdage komplekse mønstre i data og bruge dem til effektivt at forudsige frafald

blandt nye studerende. I takt med at nye studerende begynder på Metropol og bliver en del af datasættet, får modellen stadigt flere mønstre at lære af og kan tilpasse sine fremtidige forudsigelser.

Maskinlæring kan hermed defineres som *en tilgang, der søger at opdage og omsætte sammenhænge i data til modeller, der kan anvendes til prædiktioner om nye data*. Med denne definition læner vi os op ad forskere, der anvender maskinlæring til at betegne denne, for samfundsvidenskaben, nye metodologi (Kleinberg et al. 2017: 13; Mittelstadt et al. 2016; Mullainathan & Spiess 2017). I opgavens teoriafsnit uddyber vi, hvordan maskinlæring foregår i praksis.

Algoritmer I opgaven oversætter vi i videst muligt omfang begreber fra maskinlæringslitteraturen til ækvivalente begreber fra den samfundsvidenskabelige tradition. Vi anvender dog betegnelsen *algoritme*. En algoritme definerer vi som et sæt af regler for, hvordan et specifikt problem løses. De algoritmer, vi anvender i opgaven, har forskellige regler for, hvordan de håndterer observationer for at finde mønstre og sammenhænge i et datasæt. Det kunne eksempelvis være et sæt af regler, der beskriver, at et datasæt skal to-deles på baggrund af de uafhængige variable, så længe der er mindst fem observationer i begge grupper ved den sidste to-deling. I opgaven her kan begrebet algoritme i vid udstrækning læses synonymt med *model* eller *metode*, men refererer mere præcist til en models indre regelsæt.

Der anvendes en lang række forskellige algoritmer til maskinæring (se fx Friedman et al. 2009; Murphy 2012). Ingen er dog universelt bedst på tværs af undersøgelser og datasæt (Murphy 2012: 24–25). I denne opgave sammenligner vi fire forskellige algoritmer for at gå videre med den bedste af dem.

1.3 Opgavens opbygning

For at besvare vores problemstilling situerer vi i kapitel 2 specialet i den eksisterende **litteratur** om maskinlæring til målretning af policy-tiltag. Vi vil dels undersøge erfaringerne fra tidligere casestudier, hvor maskinlæring anvendes i den offentlige sektor, dels opridse de vigtigste akademiske debatter om maskinlærings potentialer og faldgruber. Vi kommer herunder ind på, hvordan maskinlæring kan ses som en del af en ny strømning inden for samfundsvidenskaben. Maskinlæring indebærer et skift i fokus fra kausalestimation til prædiktion, som har omfattende metodiske konsekvenser. Det er temaet for kapitel 3, der bidrager med det **teoretiske fundament** for at forstå maskinlæring og introducerer opgavens centrale teoretiske koncepter. Herunder

diskuterer vi, hvordan vi kan måle en models prædiktions-performance, og kigger i maskinrummet på de fire algoritmer, som vi anvender i analysen.

Kapitel 4 tjener den dobbelte rolle at dokumentere vores håndtering af **data** samt illustrere, hvordan maskinlæring som tilgang tillader en anderledes håndtering af data, end vi kender det fra kausalestimation. I kapitel 5 sammenligner vi de fire algoritmers performance og går herefter i dybden med den bedste prædiktionsmodel. Med vores **analyse** begrænser vi os ikke til at vurdere modellens evne til at forudsige frafald, men undersøger også, hvorvidt forudsigelserne kan danne grundlag for at målrette tiltag i praksis. Det gør vi ved at opstille et framework, der inddrager den konkrete policy-kontekst. Hermed besvarer vi den del af vores problemstilling, der vedrører, hvorvidt maskinlæring rent metodisk har potentiale til at målrette tiltag mod frafald. I kapitel 6 diskuterer vi validiteten af denne konklusion, og i hvilket omfang vi kan generalisere ud over vores egen case. Derefter udvider vi vores **diskussion** af tilgangens potentiale til målretning af tiltag ved at inddrage kritiske perspektiver af metodologisk, epistemisk og etisk karakter. Dermed nuancerer vi vores konklusion om maskinlærings potentiale i lyset af de dilemmaer, som følger af, at maskinlæring som tilgang tages fra ét felt, datalogien, og anvendes i et andet, samfundsvidenskaben.

Kapitel 2

Litteratur-review

Med denne opgave lægger vi os i forlængelse af en spirende litteratur om maskinlæring til målretning af policy-tiltag. Selvom opgaven handler om uddannelsesfrafald, er det maskinlæring som tilgang, der er dens substantielle genstandsfelt. Den omfattende litteratur om frafald vil vi derfor kun behandle kursorisk. Afsnit 2.1 opridser centrale indsigter fra frafaldsforskningen.

Udbredelsen af maskinlæring i den offentlige forvaltning er endnu i sin vorden, men i de seneste år er der udgivet til stadighed flere casestudier om potentialet for at bruge maskinlæring på forskellige policy-områder. I afsnit 2.2 følger et kort review med fokus på de af studierne, som handler om uddannelsesfrafald. Hovedparten af denne litteratur er skrevet fra et datalogisk perspektiv med fokus på de tekniske muligheder. Kun få af studierne har også et politologisk perspektiv på de mulige udfordringer ved at anvende maskinlæring i denne nye sammenhæng. Disse udfordringer bliver til gengæld udforsket i en separat litteratur, der forholder sig mere kritisk til udbredelsen af maskinlæring i samfundsvidenskaben og til stadigt flere policy-områder. Vi kortlægger debatterne på området i afsnit 2.3.

2.1 Hvorfor dropper studerende ud?

Frafald og lang gennemførselstid har de seneste år fyldt en del i den politiske debat om uddannelse i Danmark. Det er dog langt fra nyt terræn for uddannelsesforskningen. Siden 1970'erne har det stået højt på dagsordenen at kortlægge faktorerne bag frafald og fastholdelse (Tinto 2012). Især i USA har et væld af undersøgelser beskæftiget sig med det høje frafald på *high schools* og *colleges*, hvor omtrent 25 henholdsvis 45 procent

aldrig gennemfører (Rumberger & Lim 2008: 1; Braxton 1997: 1). Af de studerende, som starter en videregående uddannelse i Danmark, falder ca. 30 procent fra igen – halvdelen inden for det første år (EVA 2017: 5).

Forskningens udgangspunkt er som regel, at frafald er en omkostning for både den enkelte studerende, uddannelsesinstitutionen og samfundet. Først og fremmest er det spild af tid og økonomiske ressourcer, men forskning har også kædet frafald sammen med lavere selvværd og andre menneskelige omkostninger for den enkelte studerende (EVA 2017: 13–14; Troelsen 2011). Ikke desto mindre er det en forsimpling at anskue alle frafald som uhensigtsmæssige. En del studerende finder måske ud af, at en anden uddannelse bedre modsvarer deres evner, personlighed og ønsker for fremtiden – og det er nok en urealistisk forventning, at alle kan nå frem til denne indsigt før påbegyndelsen af deres studier. Derfor er rationalet bag forskningen heller ikke nødvendigvis at eliminere, men blot at nedbringe den relativt høje andel af frafald (EVA 2017: 14).

Den amerikanske forsker Vincent Tintos teorier har opnået nærmest paradigmatisk status på området (Braxton 2000: 7). Meget af frafaldsforskningen stammer i det hele taget fra USA, og der kan selvfølgelig være væsensforskelle til en dansk kontekst, hvor gratis uddannelse og SU-systemet sætter nogle andre rammer for uddannelsessektoren. Ikke desto mindre er det også Tinto, som er det fælles referencepunkt for forskningen i Danmark (EVA 2017). Tinto opstiller et framework til at forstå frafald som en proces snarere end en isoleret begivenhed – og som et institutionelt ansvar snarere end blot et udtryk for den enkelte studerendes vedholdenhed (Tinto 2012). Den studerendes beslutning om at droppe ud af sit studie er afhængig af den studerendes *commitment* til uddannelse og til sin specifikke uddannelsesinstitution (Tinto 1993). De studerendes niveau af commitment er ved studiestart bestemt af:

- Socioøkonomisk baggrund
- Forudgående skolegang
- Individuelle egenskaber

Hos Tinto er forståelsen, at baggrundsfaktorerne ikke i sig selv har betydning for frafaldet. Sådanne variable vil ganske vist have prædiktiv værdi i en model for frafald, men kun igennem deres indflydelse på det indledende niveau af commitment, som er den egentlige forklarende variabel. Undervejs i et uddannelsesforløb stiger og falder commitment på baggrund af henholdsvis akademisk og social integration på studiet (Tinto 1993). Begge aspekter kan i sig selv være tilstrækkelige for gennemførsel. Hvis den studerende fx har mange forbindelser til andre studerende, kan det kompensere for

en manglende oplevelse af meningsfuldhed med det akademiske arbejde og vice versa. Det er denne akademiske og sociale integration, som optager forskningen mest, fordi det er faktorer, som institutionerne kan forsøge at ændre rammerne for (Braxton 2000). I en dansk kontekst samler meget opmærksomhed sig om studiestartens betydning for den sociale integration (EVA 2017).

Allerede i 1980 udkommer det første studie, som med baggrund i Tintos teori forsøger at forudsige frafaldsrisiko (Pascarella & Terenzini 1980). På baggrund af faktoranalyse af 34 survey-spørgsmål danner de fem indeks til at afspejle dimensioner af akademisk og social integration på et amerikansk college. Derefter bruger de klassisk regressionsanalyse til at påvise en sammenhæng mellem disse indeks og sandsynligheden for frafald. Sidenhen har meget forskning benyttet sig af frameworket til at skelne analytisk mellem forskellige grupper af faktorer (Braxton et al. 1997). Der er dog stadig begrænset empirisk belæg for betydningen af akademisk og social integration, og hvad som fremmer den (Braxton et al. 1997; EVA 2017). De fleste undersøgelser konvergerer omkring en række faktorer, der er befordrende for gennemførsel: høje forventninger, akademisk og social støtte, feedback samt involvering i studiemiljøet (Tinto 2012: 6–9). Den eneste sikre konklusion er imidlertid, at der er mange faktorer, som spiller sammen og indvirker på beslutningen om at fortsætte eller afbryde en uddannelse (Rumberger & Lim 2008: 66). Der er med andre ord mønstre bag frafaldet, men ingen variable er endeligt determinerende.

Manglen på endegyldig viden om enkeltfaktorers betydning er dog ikke afgørende for vores opgave. Hvor undersøgelsen af Pascarella & Terenzini (1980) forsøgte at forudsige frafald ved at operationalisere Tintos teoriapparat, står maskinlæring som tilgang i kontrast hertil ved ikke at have en solid teoretisk fundering. Det er en pointe, som vi folder ud i afsnit 3.1.1 om forskelle mellem estimation og prædiktion.

2.2 Eksisterende casestudier om maskinlæring i den offentlige forvaltning

Flere offentlige institutioner eksperimenterer i disse år med potentialet for maskinlæring til målretning af policy-tiltag. Det er fx velkendt, at politiet kan bruge maskinlæring til at forudsige mønstre i kriminalitet og dermed målrette proaktiv patruljering (Perry et al. 2013; Kulager 2016). I halvdelen af de amerikanske jurisdiktioner bliver maskinlæring anvendt til at støtte beslutninger om prøveløsladelse ved at forudsige risikoen for, at en indsat vil begå nye forbrydelser (Berk 2012: 4–6; Mayer-Schönberger & Cukier

2013: 158). Et andet eksempel fra USA er målretning af sociale interventioner mod udsatte unge, som forudsiges at være i risiko for at blive skudt (Chandler et al. 2011).

Det er dog et fåtal af studierne, der på denne måde dokumenterer maskinlæring anvendt i en konkret forvaltningspraksis. De fleste af studierne afsøger *mulige* anvendelsesområder og eksperimenterer med data, som offentlige myndigheder på sigt kan tænkes at få gavn af. Eksempelvis er det blevet demonstreret, hvordan tekstanalyse af online anmeldelser kan bruges til at forudsige hygiejneforhold på restauranter – og dermed til at målrette myndighedernes kontrolbesøg (Kang et al. 2013). Et af de områder, hvor maskinlæring kan vise sig at få størst indflydelse, er på sundhedsområdet, hvor der finder mange forsøg sted med personaliseret medicin og diagnosticering (Obermeyer & Emanuel 2016). Et nyligt dansk eksperiment undersøgte, om man ved at forudsige risikoen for tilbagefald blandt KOL-patienter kunne målrette en dyrere og mere intensiv efterbehandling mod de svageste (Harmsen & Enggaard 2016). Den samme tankegang ligger bag et forsøg med at forudsige risikoen for alvorlige komplikationer for at prioritere blandt patienter til dyr hoftekirurgi (Kleinberg et al. 2015).

På uddannelsesområdet er maskinlæring blevet brugt til at målrette forslag om relevante kurser mod bestemte studerende samt ansætte lærere og udvælge studerende ud fra deres forudsagte performance (Romero & Ventura 2010). Vi vil herunder stille skarpt på erfaringerne fra de studier, som har beskæftiget sig med forudsigelsen af frafald.

Endelig kan det også bemærkes, at maskinlæring selvsagt kan finde anvendelse i offentlige institutioner uden at være akademisk veldokumenteret eller offentligt kendt.

2.2.1 Forudsigelse af frafald

Det første studie, som anvender maskinlæring til forudsigelse af frafald, er fra 2003. Der er tale om et mindre studie omfattende 354 studerende på et langdistancekursus i informatik på det græske Hellenic Open University (Kotsiantis et al. 2003). Forskerne afprøver en række populære algoritmer fra datalogien, herunder klassifikationstræer, neurale netværk, support vektor-maskiner og naive bayesianske klassifikatorer. De konkluderer, at algoritmerne performer næsten ens. Den bedste model lykkes med at forudsige 63 procent af de studerende korrekt ud fra baggrundsdata alene og 83 procent, når de inkluderer akademiske præstationer halvvejs inde i studiet (Kotsiantis et al. 2003). Studiet udmærker sig endvidere ved at udvikle en prototype på et webbaseret værktøj, som automatisk skal identificere fremtidige frafaldstruede studerende.

Studiet er blandt de tidligste på uddannelsesfeltet til at udforske potentialet for maskinlæring her (Romero & Ventura 2010). Siden er der udgivet en håndfuld lignende casestudier fra andre universiteter (Herzog 2006; Dekker et al. 2009; Kovačić 2010; Bayer et al. 2012; Kristoffersen 2015). Fire ud af de fem stammer også fra polytekniske eller datalogiske fakulteter, hvilket formentlig afspejler maskinlærings ophav i computervidenskaben, og at anvendelsen til forudsigelse af uddannelsesfrafald stadig er på forsøgsstadiet.

Herzog (2006) klassificerer 84 procent af de studerende korrekt i et studie med 8018 amerikanske universitetsstuderende. Han finder endvidere, at klassifikationstræer kun performer en smule bedre end logistisk regression. Han mistænker datagrundlaget og antallet af variable for at være for småt til fuldt at udnytte potentialet i denne type af algoritmer. Kovačić (2010) bruger klassifikationstræer på 453 newzealandske universitetsstuderende og viser, at det er muligt at forudsige frafald med klassifikationstræer kun ud fra variable, som er tilgængelige ved studiestart – men også, at man ikke kan forvente sig mirakler. Frafaldet i dette case oplyses at være 50 procent. Modellen forudsiger 60 pct. af de studerende korrekt, hvilket svarer til en forbedring på 10 procentpoint, sammenlignet med hvad vi kunne have opnået ved tilfældige gæt. I de tidligere studier mangler der en sådan baseline at vurdere modellens nøjagtighed op imod, og det hindrer en egentlig vurdering af modellernes performance.

Modellernes nøjagtighed (på engelsk *accuracy*) er i det hele taget et utilstrækkeligt mål til at sammenligne performance på tværs af cases – særligt når værdierne ikke er ligeligt fordelt på outcome-variablen (Mollineda et al. 2007). Det er ofte tilfældet med frafald, fordi det trods alt er et fåtal at studerende som dropper ud. Dekker et al. (2009), som forudsiger frafald blandt hollandske studerende, påpeger som de første denne udfordring. De peger på AUC som et bud på et bedre performancemål, og siden er flere casestudier begyndt at orientere sig i denne retning (Aulck et al. 2016; Şara 2014; Kristoffersen 2015). AUC tager en værdi mellem 0,5 og 1 og er efterhånden blevet at regne for standarden for måling af prædiktions-performance (Kleinberg et al. 2017: 16). Prædiktion står så centralt i opgaven, at der følger en grundigere teoretisk indføring i forskellige performancemål i afsnit 3.1.2.

Det mest lovende frafaldsstudie til dato – målt ved både accuracy og AUC – er på de danske gymnasier (Şara 2014). Dette studie benytter sig af data fra det omfattende intranet Lectio, som anvendes af hovedparten af danske gymnasier, og hvor lærerne registrerer fravær og karakterer, og eleverne uploader afleveringer og holder sig opdateret på skemaændringer og læseplan. Med en model baseret på 36.000 gymnasieelever og et halvt års data opnås en accuracy på 93 procent og en AUC-værdi på 0,96. I et opfølgende studie viser Kristoffersen (2015), at modellen også kan klare sig med data fra

blot den første måned til at nå en relativt fin performance – her opnås en AUC-værdi på 0,86.

På baggrund af de ovennævnte casestudier kan vi ikke konkludere noget entydigt om forskellige variables potentiale til at forudsige frafald. I ét studie har etnicitet fx den største prædiktive værdi (Kovačić 2010), i et andet studie ingen (Şara 2014: 53). Mere interessant end specifikke variable er forsøgene på at måle social integration. Bayer et al. (2012) udgiver det første studie, som forsøger at anvende observationsdata frem for surveyspørgsmål til at indikere social integration. Konkret anvendes sociale data fra universitetets intranet i modellen. Det omfatter fx variable som "antal beskeder i offentlige fora" og "deling af filer med andre studerende". Det forbedrer modellens accuracy med ca. 10 procentpoint. På de danske gymnasier har den sociale integration målt ved bl.a. "antal beskeder sendt" ikke meget betydning; her er registreret fravær langt den stærkeste prædiktor (Kristoffersen 2015: 48).

Mens de tidligste studier ikke drog nogen entydige konklusioner om bedste algoritmevalg (Kotsiantis et al. 2003; Herzog 2006), har de seneste undersøgelser på området fundet, at familien af træbaserede algoritmer leverede de mest nøjagtige forudsigelser (Bayer et al. 2012; Kristoffersen 2015). De træbaserede algoritmer, som bliver anvendt, er dog også blevet mere avancerede i løbet af perioden. Casestudierne af frafald på danske gymnasier opnår fx de bedste resultater med en *Random Forest*-algoritme, som bygger på et såkaldt ensemble af klassifikationstræer (Şara 2014). Også uden for uddannelsesområdet vinder de træbaserede modeller frem – i særlig grad ensemble-modeller. Additive ensemble-modeller som fx *Gradient Boosted Trees* har opnået prominens ved at vinde mange online maskinlærings-konkurrencer (Athey & Imbens 2016: 50–51; Harmsen & Enggaard 2016: 55). Inden for samfundsvidenskaben er Gradient Boosted Trees fx blevet anvendt til at forudsige risikoen for tilbagefald blandt KOL-patienter (Harmsen & Enggaard 2016) og til at forudsige risikoen for, at varetægtsfængslede vil begå ny kriminalitet eller ikke møde op i retten (Kleinberg et al. 2017). Både Random Forest og Gradient Boosted Trees er blandt de algoritmer, vi tester i denne opgave. Alle anvendte algoritmer bliver introduceret i afsnit 3.3.

2.3 Debatter om maskinlæring

I de forgående afsnit har vi foretaget et review af litteraturen om frafald og casestudier, der anvender maskinlæring til målretning af policy-tiltag. I dette afsnit vil vi vende os mod en litteratur, som problematiserer udbredelsen af maskinlæring. Mens det er

relativt ukontroversielt at klassificere e-mails som spam[1], står de etiske dilemmaer i kø, når det handler om at klassificere mennesker som sandsynlige skattesnydere, langtidsledige eller frafaldstruede. Det første underafsnit vedrører maskinlærings plads i samfundsvidenskabens metodiske landskab. Det er en debat, som er metodologisk, idet den vedrører fordele og ulemper ved maskinlæring som tilgang sammenlignet med andre tilgange i samfundsvidenskaben. I de næste to underafsnit ser vi på henholdsvis epistemiske og etiske problemstillinger, hvormed vi strukturerer den akademiske debat med inspiration fra Mittelstadt et al. (2016). Refleksioner fra disse tre afsnit vil vi senere bringe i spil, når vi i kapitel 6 diskuterer frafaldsmodellens potentiale.

En del af den litteratur, vi trækker på, diskuterer implicit eller eksplicit maskinlæring med big data som det egentlige omdrejningspunkt (boyd & Crawford 2012; Kitchin 2014b; Mayer-Schönberger & Cukier 2013). Det er en litteratur, som handler om de muligheder og risici, som følger af en øget kapacitet til at søge, aggregere og krydse store datasæt (boyd & Crawford 2012: 663–664). Maskinlæring er en tilgang, som ofte benyttes til at bearbejde og undersøge store datamængder, og derfor har det en naturlig plads i litteraturen om big data[2]. Mange indsigter fra denne litteratur er dermed også relevante i vores diskussion.

2.3.1 Metodologiske debatter

En central strømning i den empiriske politologi er den såkaldte *credibility revolution* (Clark & Golder 2015). Det er en strømning, som er kendetegnet ved særlig interesse for kausal inferens og udbredelsen af en design-baseret forskningstilgang (Angrist & Pischke 2010). Det er studier, som lægger vægt på eksperimentelle forskningsdesigns og omhyggelig teoretisk specifikation, der kan sikre betingelserne for at estimere troværdige kausale sammenhænge.

Samtidig argumenterer mange for, at vi står midt i en *big data-revolution*, der indvarsler en ny æra for både videnskaben og samfundet (Anderson 2008; boyd & Crawford 2012; Einav & Levin 2014; Mayer-Schönberger & Cukier 2013). I denne strømning er fokus på de stigende muligheder for at analysere uanede mængder af data ved hjælp af maskinlæring og andre datadrevne tilgange. Forventningen er, at det kommer til at transformere vores forståelse af den politiske og sociale verden.

[1] Alt er som bekendt relativt. Burrell (2016) demonstrerer, hvordan også klassifikation af spam indebærer etiske dilemmaer.

[2] Big data skal her forstås som et bredt fænomen, som både har en teknisk, social og kulturel dimension. Begrebet big data står ikke centralt i opgaven her, hvorfor vi ikke definerer det nærmere. For en detaljeret afgrænsning af begrebet, se fx boyd & Crawford (2012).

Disse to metodologiske strømninger står på nogle måder i et spændingsforhold (Clark & Golder 2015). Sat på en spids handler det om, hvorvidt den nye datadrevne tilgang markerer begyndelsen på enden for den empiriske samfundsvidenskab, som vi kender den (Burrows & Savage 2014). De stærkeste fortalere for dette synspunkt mener, at videnskabens optagethed med kausalitet og hypotetisk-deduktive tilgang til forskning er forældet. I det lettere polemiske indlæg *The End of Theory* argumenterer Anderson (2008) for, at videnskabelig teori er overflødig, når datamængderne er så store, at korrelationer i data taler for sig selv. Her er synspunktet, at vi ikke længere behøver at kere os om kausalitet. Det er ikke afgørende, hvad årsagen til en sammenhæng er, men blot at sammenhængen eksisterer. Mayer-Schönberger & Cukier (2013) argumenterer ligefrem for, at det er mere ydmygt at handle ud fra de korrelationer, som data viser os, frem for at støtte os på teorier om kausalitet, der ofte viser sig at være fejlagtige.

Ikke overraskende støder mange sig på disse vidtrækkende konklusioner. Det er da også et synspunkt, som primært nyder opbakning inden for datalogien og i mediernes hype omkring big data – inden for samfundsvidenskaben er de fleste mere forbeholdne omkring de vidtløftige forventninger (Clark & Golder 2015). De vil medgive, at der ligger et potentiale i at udnytte data til at finde korrelationer, men at det ikke erstatter behovet for teori og kausal inferens (Grimmer 2015; Athey 2017). Anken er, at klassiske statistiske faldgruber som fx selektionsbias og spuriøse sammenhænge ikke bliver taget alvorligt nok i den nye datadrevne forskning (Grimmer 2015; Clark & Golder 2015). Disse velkendte problemer kan ikke løses ved blot at skrue op for mængden af data, lyder et argument – det er stadig problematisk at handle alene på baggrund af korrelationer. I praksis vil det ofte være umuligt at vide, hvad den rette handling er, hvis vi ikke har en dyberegående forståelse for, *hvorfor* ting hænger sammen, som de gør (Athey 2017; Cederman & Weidmann 2017).

Derfor er det formentlig mere frugtbart at se den nye datadrevne tilgang som komplementær til teoridreven forskning snarere end som et opgør (Hofman et al. 2017; Burrows & Savage 2014). Tilgangene er komplementære i den forstand, at de har forskellige målsætninger. Den teoridrevne, hypotetisk-deduktive tilgang, som stadig er dominerende på feltet, har som mål at forklare og forstå komplekse sammenhænge. Den datadrevne, induktive tilgang, der vinder frem som et alternativ, har ofte som mål at komme med præcise forudsigelser (Hofman et al. 2017). Betydningen af denne skelnen mellem forklaring og forudsigelse, mellem kausal estimation og prædiktion, beskriver vi i flere detaljer i afsnit 3.1.

Endelig er det på sin plads at understrege, at vi her har begrænset os til den del af debatten, som vedrører konsekvenserne af at anvende maskinlæring til forudsigelse. Der er en stigende interesse for, hvordan metoder fra maskinlærings-litteraturen også har et

potentiale i studier, hvor formålet er kausal inferens (Athey & Imbens 2016; Grimmer 2015; Varian 2014). Det ligger imidlertid uden for rammerne af denne opgave, hvor vi interesserer os for maskinlærings anvendelse til prædiktion.

2.3.2 Epistemiske debatter

En del af debatten om maskinlæring har epistemisk eller videnskabsteoretisk karakter. Den angår med andre ord karakteren af den viden, som vi opnår gennem datadreven, algoritmisk vidensproduktion.

Eksempelvis kan en epistemisk kritik rettes mod ovennævnte position om, at data "kan tale for sig selv" (Anderson 2008). Når big data-entusiaster profeterer om teoriens endeligt, antager de, at data kan være neutrale og værdifri (Kitchin 2014a). Men data bliver ikke genereret i et teoretisk vakuum. Data bliver indsamlet og kategoriseret i systemer, der allerede er formet af teoretisk viden, historiske erfaringer og tilgængelige teknologier (Kitchin 2014a). De bliver præsenteret som objektive og teorifri, men kan reelt skjule menneskeligt bias, som føder videre over i de modeller, som data bliver indarbejdet i (Barocas & Selbst 2016). Deres aura af objektivitet gør det endvidere svært at udfordre modellernes resultater – og det bliver ikke bedre af, at algoritmerne anvendt til maskinlæring ofte er uigennemskuelige (Rieder & Simon 2016).

Meget af den epistemiske kritik går derfor på at udfordre kvantificerbare datas ophøjede status. Datas ophøjethed er dog ikke noget særegent for maskinlæring; kvantitative beskrivelser af vores samfund har længe haft en vidensmæssig forrang (Mayer-Schönberger & Cukier 2013: 163–170). Kvantificeret information tillæges en særlig betydning og autoritet mange steder i hverdagslivet, inklusive i den offentlige administration (boyd & Crawford 2012). Imidlertid er italesættelsen af maskinlæring som et brud eller en revolution med til at give tilgangen appel og diskursiv vægt (Kitchin 2014b: 113). Der er en udbredt forestilling om, at denne nye tilgang muliggør en højere form for viden og kan bibringe hidtil utilgængelige indsigter (Rieder & Simon 2016). Der er fokus på forandringen frem for kontinuiteten. Flere har vist, at en måde at udfordre denne forståelse på er ved netop at kontekstualisere maskinlæring som en del af en længere historisk udvikling (Rieder & Simon 2016; Kitchin 2014b).

Det er en historisk udvikling, som tager sin begyndelse med fremkomsten af den statistiske videnskab. Statistikken er uløseligt forbundet til moderne statsbygning, idet numerisk beskrivelse muliggør en sammenhæng og generalitet, der er nødvendig for centraliseret administration (Desrosières 1998; Foucault 2007). Foucault (2007) beskriver, hvordan regeringsmagten gradvis forandrer sig fra forvaltning af territorier

til forvaltning af en population af numerisk beskrevne individer. I begyndelsen får den numeriske viden sin autoritet fra prestigefyldte statslige institutioner, men i det 20. århundrede udvikler idéen sig om maskinen som den perfekte "objektive outsider", der uden menneskelig indblanding kan afdække sandheden i data (Porter 1996: 85–86, 137–138). Tal får deres egen selvstændige aura af saglighed, neutralitet og sikkerhed (Porter 1996: 90–93). Denne udvikling styrkes med computeriseringen af samfundet (Rose 1991) og kan samtidig sættes i relation til en svigtende tillid i de politiske og sociale systemer, der nærer efterspørgslen på objektive holdepunkter (Rieder & Simon 2016). Den nuværende hype omkring big data og maskinlæring kan således situeres i en bredere socio-politisk kontekst, hvor offentlige beslutningstagere søger numerisk evidens at støtte deres handlinger på (Dahler-Larsen 2011). Det gør dem villige til at eksperimentere med nye datadrevne former for governance (Rieder & Simon 2016).

Problemet er, at beslutningstagning funderet i data og algoritmer kan give en falsk forestilling om sikkerhed og upartiskhed. I en diskussion af maskinlærings potentiale er det en nødvendig problemstilling at adressere.

2.3.3 Etiske debatter

Udbredelsen af maskinlæring i den offentlige forvaltning har også afstedkommet en debat om normative implikationer af at målrette policy-tiltag. Debatten rejser forskellige spørgsmål, som vedrører tilgangens fairness eller retfærdighed. Eksempelvis frygter nogle, at maskinlæring vil gøre forvaltningen mere teknokratisk (Mittelstadt et al. 2016). Algoritmiske beslutninger kan være svære at udfordre og reducerer nogle gange komplekse problemer til rent tekniske beslutninger. En anden udbredt bekymring ved algoritmisk beslutningstagning handler om håndteringen af data. Det rejser spørgsmål om datasikkerhed og om retten til privatliv (Foster et al. 2016; Nissenbaum 2011). Er det etisk forsvarligt, at den offentlige forvaltning inddrager en lang række personoplysninger i beslutningstagningen? Denne debat trækker også tråde til klassiske diskussioner om effekterne af overvågning (Foucault 2012).

Det måske mest centrale stridspunkt i normative diskussioner om maskinlæring går på, om det kan retfærdiggøres at forskelsbehandle individer. Er det rimeligt, at særlige udsatte gruppe tilbydes – eller endog påtvinges – særlige policy-tiltag? Hvor nogle mener, at diskrimination på baggrund af gruppetilhørsforhold altid er problematisk, argumenterer andre for, at positiv diskrimination er en forudsætning for reel ligebehandling (Barocas & Selbst 2016; Lippert-Rasmussen 2011).

Det er ikke entydigt, om målretningen af tiltag bryder med eller bidrager til at reproducere eksisterende mønstre i data (Rieder & Simon 2016; Barocas 2014). Det er en problematik, som naturligvis hænger sammen med karakteren af de tiltag, som modellen skal benyttes til at målrette. Samtidig kan selve sorteringen af individer i kategorier have selvstændige, uforudsete effekter. En gren af litteraturen beskæftiger sig med disse såkaldt konstitutive effekter af klassifikation (Dahler-Larsen 2014; Johnson 2014; Mittelstadt et al. 2016).

Kapitel 3

Teori

3.1 Prædiktion i samfundsvidenskab

I naturvidenskaben er prædiktion udbredt og ukontroversielt (Hofman et al. 2017). Anderledes har samfundsvidenskaben historisk eftersøgt forklaringer frem for forudsigelser (Hofman et al. 2017; Breiman et al. 2001). En del af grunden kan være et syn på sociale og samfundsmæssige systemer som fænomener, der er iboende komplekse og foranderlige, og som vi har relativt begrænsede data til rådighed om (Hofman et al. 2017). Det kan have betydning for, hvordan forskningsspørgsmål formuleres i samfundsvidenskaben. I stedet for at spørge, om en given teori kan forudsige et fænomen, så spørges typisk, om en bestemt variabel i en statistisk model kan forklare et outcome, og resultatet afhænger af, om en koefficient er statistisk signifikant og peger i samme retning, som teorien og hypoteserne tilsiger (Hofman et al. 2017; Breiman et al. 2001).

Men forudsigelser kan være nyttige (Kleinberg et al. 2017; Silver 2012; Tetlock & Gardner 2016). Eksempelvis spørger Nate Silver polemisk, "*If political scientists couldn't predict the downfall of the Soviet Union – perhaps the most important event in the latter half of the twentieth century – then what exactly were they good for?*" (Silver 2012: 51). Samtidig kan mange forskningsspørgsmål om samfundet med fordel anskues som prædiktionsproblemer, hvor målet er at forudsige frem for at forklare et outcome (Kleinberg et al. 2015). Kleinberg et al. (2017: 41) formulerer, at prædiktionsproblemer er en videnskabeligt interessant og socialt vigtig klasse af problemer, hvor en empirisk tilgang til prædiktion har stort potentiale.

Netop en empirisk tilgang til prædiktion har de seneste år vundet udbredelse i samfundsvidenskaben (Hofman et al. 2017). Det skyldes blandt andet den massive forøgelse i volumen og typer af data om sociale forhold, der dels har fået dataloger til at kigge i retning af samfundsvidenskaben, dels fået politologer som os selv til at undersøge mulighederne i nye metoder fra datalogien (Hofman et al. 2017; Athey 2017).

Hofman et al. (2017) ser udviklingen som en kærkommen mulighed for at genbesøge det historiske skel i samfundsvidenskaben mellem forklaring og forudsigelse, mellem estimation og prædiktion. Det er emnet for næste afsnit. Samtidig står forskellen mellem estimation og prædiktion som et helt centralt omdrejningspunkt gennem opgaven her, og næste afsnit er derfor også det videre afsæt til hele det nærværende kapitel, hvor vi vil stille skarpt på den kaskade af konsekvenser, det har for vores metode og analyse i praksis, at vi griber casen Metropol an som et prædiktionsproblem frem for et estimationsproblem.

3.1.1 Estimation vs. prædiktion

Formålet med kvantitative empiriske studier i samfundsvidenskaben er ofte at etablere kausalsammenhænge og estimere kausale effekter (Wooldridge 2009: 12–13; Angrist & Pischke 2015: xii–xv). Det gøres ved at undersøge kontrafakta gennem sammenligninger under *ceteris paribus*, hvor alt andet er eller antages at være lige.

I politologisk forskning har idéen om kontrafakta en intuitiv appel, fordi idéen lægger op til at kunne sammenligne to forskellige scenarier, hvor den eneste forskel eksempelvis er, hvorvidt en given policy er blevet implementeret eller ej (Kleinberg et al. 2015). Ude i virkeligheden genfindes kun ét af de to scenarier. Det andet scenarium, det kontrafaktiske, vil for altid forblive en idé uden empirisk modstykke, hvad der omtales *det fundamentale problem i kausal inferens* (Rosenbaum & Rubin 1983; Holland 1986; Imbens & Rubin 2015: 3–21). Den kontrafaktiske situation har imidlertid interesse, fordi den giver mulighed for at estimere effekten af den policy, som i dette eksempel er den eneste forskel mellem de to tænkte scenarier. Det er nemt at forestille sig, at beslutningstagere kunne være interesserede i at kende effekten af en given policy, før den implementeres, eller at man som politolog kunne være interesseret i at vide, om øget indkomst har en effekt på ideologisk højre-venstre-placering, eller noget helt tredje. Det er denne identifikation af kausale effekter, som vi her vil omtale *estimation*, og som vi stiller over for *prædiktion*, der anderledes handler om at forudsige et outcome.

Set i lyset af den samfundsvidenskabelige tradition for estimation, og estimationens intuitive appel, hvorfor så lave prædiktion? En typisk situation, hvor prædiktion er

interessant, er, når vi gerne vil kende et givent outcome, y, men ikke har mulighed for at måle det (James et al. 2013: 17-18). Det kan skyldes, at det er et fremtidigt outcome, som har interesse, fordi man politisk ønsker at rette et tiltag mod det. I vores case ønsker vi at kende de studerendes sandsynlighed for at frafalde for at kunne målrette tiltag mod de mest frafaldstruede studerende. Problemet er her, at vi ikke direkte kan måle de studerendes risiko for at frafalde. I stedet kan vi forsøge at prædiktere denne risiko $\hat{y}$ som en tilnærmelse af y.

En central forskel mellem prædiktion og estimation er, at prædiktions-performance kan måles, mens estimations-performance ikke kan. Vi kan aldrig måle estimations-performance på grund af ovennævnte *fundamentale problem i kausal inferens*. Vi vil aldrig få adgang til den kontrafaktiske situation. Kausal estimation hviler til syvende og sidst på en række antagelser, der skal underbygge, at alt andet er lige, og at modellen derfor er unbiased og rigtig i gennemsnit.

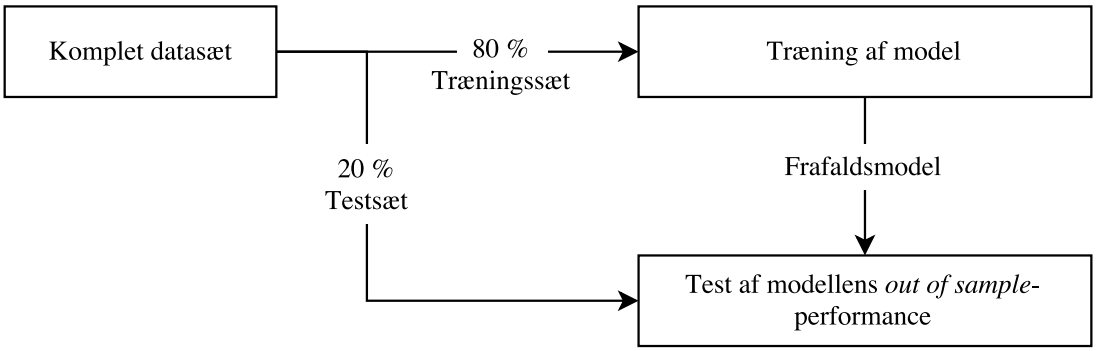
Prædiktionsmodeller er derimod baseret på korrelationer, som vi kan teste empirisk, fordi vi ikke behøver at gøre os videre antagelser om eventuelle årsagssammenhænge. Man kan derfor teste en prædiktionsmodels forudsigelser ved at træne den på kendte, historiske data og teste den på nye, fremtidige data. I vores case kan vi fx udvikle en model, som finder mønstre i tidligere studerendes frafald. Derpå kan vi bede den komme med forudsigelser om en ny årgang af studerende, og efter et par år kan vi så se, om forudsigelserne holdt stik. I praksis kan vi naturligvis ikke vente så længe med at teste modellens performance. Derfor deler vi i stedet det eksisterende datasæt op i to tilfældige samples: et træningssæt bestående af 80 pct. af det data, vi har til rådighed, og et testsæt med de resterende 20 pct. af data. Ved at træne[1] modellen på træningssættet og afprøve den på testsættet får vi et estimat over modellens prædiktions-performance. Logikken fremgår af figur 3.1 på den følgende side.

At prædiktions-performance kan måles, udgør en helt central væsensforskel mellem estimation og prædiktion, hvilket skyldes de to tilganges forskellige formål. Når man beskæftiger sig med estimation er formålet at maksimere *in-sample* performance i træningssættet, så den model, man opstiller, modellerer data bedst muligt. Når man beskæftiger sig med prædiktion, er formålet at maksimere *out-of-sample* performance i testsættet – dvs. opnå de mest præcise forudsigelser på det nye data.

Denne væsensforskel mellem estimation og prædiktion har stor betydning for den praktiske tilgang. Samtidig illustrerer den, hvorhenne vi med vores fokus på prædik-

[1] At *træne* en model er blot et synonym fra maskinlæringslitteraturen for at *fitte* en model til datasættet. Vi anvender ofte denne betegenelse i opgaven, fordi det understøtter logikken om, at modellen fittes til data i et træningssæt og testes på ukendte data i et testsæt.

Figur 3.1: Opsplitning af data i et trænings- og testsæt



tion placerer os i relation til den empiriske samfundsvidenskab som felt. En central strømning i feltet, som vi stiftede bekendtskab med i litteratur-reviewet, er den såkaldte *credibility revolution* (Angrist & Pischke 2010). Med vores korrelations-baserede prædiktionsmodeller adskiller vi os fra tankegodset i denne, og om dette kan siges at være et skridt fremad eller et skridt tilbage er ikke givet på forhånd. Vi vender tilbage til refleksioner om fordele og ulemper ved estimation og prædiktion i opgavens diskussionskapitel.

Forskellen mellem empirisk målbar og ikke-målbar performance har stor betydning for teoriens rolle i forskningen. Fordi antagelserne bag kausalestimation ikke kan testes empirisk, spiller teori en central rolle i opstilling af modellerne. Eksempelvis bygger modeller til kausalestimation på antagelser om den datagenererende proces; opstiller man fx en OLS-model, antager man, at den datagenererende proces er lineær i parametrene[2] (Wooldridge 2009: 84, 157). Fordi antagelsen ikke kan testes empirisk, anvendes teori til at sandsynliggøre denne antagelse. På samme måde bruges teori eksempelvis til at guide valget af, hvilke variable som inkluderes i modellen. Udeladelse af relevante variable vil give *omitted variable bias*, og inklusion af for mange kan give problemer med multikollinearitet og misvisende høje standardfejl som følge. Sidstnævnte leder især til type II-fejl, mens førstnævnte kan lede til fejl af både type I og type II (Agresti & Finlay 2009: 159–163; Wooldridge 2009: 89–101).

I denne henseende er prædiktion på sin vis simplere end estimation. I prædiktion er vi interesserede i $\hat{y}$ og ikke i de enkelte parameter-estimater, $\hat{\beta}_k$. Når vi opstiller og træner en given prædiktionsmodel $\hat{y} = f(\mathbf{X})$, hvor det kapitale og fede $\mathbf{X}$ angiver, at der er tale om en matrix af uafhængige variable, kan vi derfor også tillade os at betragte selve funktionen f som en *black box*. Det kan vi, fordi vi ikke er interesserede i funktionen eller parametrene i sig selv – vi er blot interesserede i at lave gode prædiktioner

[2] Fx $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$

(James et al. 2013: 17). Som følge heraf kan vi også arbejde med en bredere palet af mere fleksible funktionelle former i prædiktionsmodeller, såsom modeller med meget komplekse interaktioner mellem parametrene. Det skyldes, at gode prædiktioner ikke behøver at tage hensyn til parsimoni og muligheden for forståelse og fortolkning, som vi kender det fra estimationsmodeller (James et al. 2013: 19, 25; King et al. 1994: 20). Hvis vi genkalder os frafaldsstudierne fra litteratur-reviewet, er forskellen slående. Hvor tidligere estimationsstudier forsøgte at finde teoretisk orden i en rodet empirisk forskning (Pascarella & Terenzini 1980), søger vi med maskinlæring orden i empirien uden skelen til teoretisk parsimoni. Det er ikke afgørende om en sammenhæng mellem variable er korrekt teoretisk specificeret, så længe den bidrager til præcise forudsigelser.

I prædiktionsmodellen kan valget af den funktionelle form derfor være datadrevet frem for teoridrevet (Breiman et al. 2001). Muligheden for datadrevet at vælge funktionel form har den væsentlige fordel, at det kan levere bedre prædiktioner (James et al. 2013: 20). En antagelse om linearitet i den datagenererende proces er typisk en forsimpling, og lineære modeller vil derfor være ringere til at fitte data end mindre restriktive modeller (Breiman et al. 2001: 204; James et al. 2013: 23). Den store ulempe er dog, at vi med de mere fleksible modeller risikerer at *overfitte* til data. Begrebet overfitting dækker over en situation, hvor vi fitter til støj i data, forskelligt fra kun at fitte til signalet, dvs. til den strukturelle, datagenerende proces (James et al. 2013: 32). I denne situation ville en mere restriktiv model have leveret bedre prædiktioner, når vi prædikterer på nye data, såsom nystartede studerende på Metropol. Håndteringen af overfitting står meget centralt ved brug af maskinlæring. I afsnittene 3.2.4 til 3.2.7 viser vi, hvordan en central del af maskinlærings-processen handler om at håndtere overfitting ved empirisk at justere vores prædiktionsmodeller. Før vi når dertil, vil vi afklare, hvordan vi måler prædiktions-performance. Det er temaet for næste afsnit.

3.1.2 Hvordan måler vi prædiktions-performance?

Når vi skal forholde os til, hvor gode prædiktioner vores modeller leverer, har vi behov for et mål for deres performance. Fra en overfladisk betragtning er det måske overraskende, at vi har behov for et selvstændigt afsnit til at redegøre for, hvordan vi måler performance. Hovedparten af de casestudier, som vi gennemgik i forrige kapitel, evaluerer blot deres models accuracy, dvs. andelen af korrekte forudsigelser. Accuracy kan intuitivt forekomme som et godt mål. Problemet er imidlertid, at det afspejler fordelingen af outcomes i den underliggende data, hvilket er et problem, når der er en ulige fordeling på outcome-variablen (Witten et al. 2005: 161–163). Det er tilfældet i vores case, hvor der trods alt er et fåtal, som falder fra. Derfor kunne vi opnå en ret god accuracy blot ved at gætte på, at alle studerende gennemførte. I et studie

med 5 procents frafald, ville en models accuracy fx være på 95 procent, hvis den blot forudsagde alle studerende til at gennemføre.

Den underliggende antagelse for accuracy er således også, at et korrekt forudsagt frafald og et korrekt forudsagt ikke-frafald er "lige meget værd". Denne antagelse er formentlig uholdbar i de fleste policy-kontekster. Vi har derfor brug for et performancemål, som rummer mere nuanceret information om vores models forudsigelser. I dette afsnit vil vi vise, hvordan en såkaldt *confusion-matrix* er et godt værktøj til at vurdere en model i detaljer. På baggrund heraf kan vi aflede nogle mere retvisende performancemål til at sammenligne forskellige modellers performance. Vi lægger her i opgaven vægt på *ROC-kurver* og *AUC*, som med oprindelse i klinisk forskning har vundet stor udbredelse i maskinlæringslitteraturen[3] (Fawcett 2006; Kleinberg et al. 2017: 16–17; Witten et al. 2005: 169–173).

Confusion-matricer

En confusion-matrix viser et udfaldsrum for, om en models prædiktioner er rigtige eller forkerte (Fawcett 2006). Det kan vi illustrere ved at tage udgangspunkt i vores egen case. Her er frafald det outcome, y_i, vi ønsker at prædiktere. Fordi studerende enten kan frafalde eller ikke frafalde, er der tale om et binært prædiktionsproblem. Formelt ønsker vi at prædiktere $(\hat{y}_i|\mathbf{X}_i) = \{0, 1\}$, hvor $y_i = 1$ angiver, at den studerende er frafaldet sin uddannelse, og $y_i = 0$ angiver det modsatte. Hatten over $\hat{y}$ angiver, at der er tale om en prædiktion af outcome til forskel fra det faktiske outcome y. I praksis prædikterer vi dog ikke enten $\hat{y}_i = 0$ eller $\hat{y}_i = 1$. Vi prædikterer derimod en sandsynlighed mellem 0 og 1 for, at den studerende frafalder, dvs. $\hat{y}_i = \hat{p}(y_i = 1|\mathbf{X}_i)$. Det betyder, at vores prædikterede outcome, $\hat{y}_i$, ikke er kategorisk ligesom det faktiske outcome, y_i, men derimod er kontinuert i intervallet 0 til 1. For at transformere en estimeret sandsynlighed til en kategorisk prædiktion fastsætter vi en *tærskelværdi*, δ, og introducerer følgende prædiktions-regel:

$$\hat{y}_i < \delta \Rightarrow \hat{y}_i = 0$$

$$\hat{y}_i \geq \delta \Rightarrow \hat{y}_i = 1$$

På baggrund af den prædikterede sandsynlighed klassificerer vi altså de studerende i to grupper, som vi prædikterer henholdsvis vil og ikke vil falde fra. Et eksempel kan være, at vi sætter tærsklen til $\delta = 0,30$. Hermed klassificerer vi alle studerende, der

[3]Det bør dog bemærkes, at der ikke er konsensus om ét rigtigt performancemål – for en kort introduktion til forskellige mål, se fx Hofman et al. (2017).

har en prædikteret sandsynlighed for frafald over 30 pct. som frafaldstruede ($\hat{y}_i = 1$). Vi gør det omvendte for studerende, der har en prædikteret sandsynlighed for frafald under 30 pct.

Det kan i begge tilfælde være en empirisk rigtig eller empirisk forkert klassifikation. Det giver et udfaldsrum på fire felter, en confusion-matrix, som illustreret i tabel 3.1. I matricen er der to prædiktioner, som er empirisk rigtige (de sande), og to prædiktioner, som er empirisk forkerte (de falske). De to betegnelser for forkert klassifikation, FP og FN, svarer til type I- og type II-fejl. Det er en type I-fejl eller en falsk positiv, hvis vi fejlagtigt afviser en hypotese om ikke-frafald, og det er en type II-fejl eller en falsk negativ, hvis vi fejlagtigt fastholder en hypotese om ikke-frafald (Agresti & Finlay 2009: 159–161).

Tabel 3.1: Confusion-matrix ved binær klassifikation

		Prædikteret outcome	
		$\hat{y}_i = 1$	$\hat{y}_i = 0$
Faktisk outcome	$y_i = 1$	Sande positive (SP)	Falske negative (FN)
	$y_i = 0$	Falske positive (FP)	Sande negative (SN)

Her er det relevant at introducere lidt terminologi, som vi vil anvende løbende gennem opgaven. De fire udfald i matricen ovenfor, SP, FP, SN, FN, vil vi omtale under samlebetegnelsen *prædiktionstyper*. Vi vil desuden omtale studerende som *prædikteret frafaldstruede*, hvilket er dem, som vi forudsiger vil frafalde, dvs. SP + FP. Vi vil også omtale studerende som *faktisk frafaldstruede*, hvilket er dem, som rent faktisk frafalder, dvs. SP + FN.

Confusion-matricer er et godt værktøj til at forholde sig selvstændigt til performance af en given prædiktionsmodel, fordi de indeholder nuanceret information om, hvilke observationer vi klassificerer rigtigt og forkert. Som eksempel kan vi kaste et blik på tabel 3.2 på den følgende side, der viser antallet af prædiktionstyper for en af modellerne i vores analyse af Metropol. Her ses det, at vi forudsiger 228 frafald korrekt. Samtidig klassificerer vi dog fejlagtigt 518 studerende som frafaldstruede, selvom de ikke ender med at frafalde. Derudover er der 331 studerende, som falder fra, men som vi ikke fanger med modellen. Slutteligt er der 3287 studerende, som ikke falder fra, og som vi også korrekt forudsiger til at gennemføre.

Tabel 3.2: Eksempel på confusion-matrix

		Prædikteret outcome	
		$\hat{y}_i = 1$	$\hat{y}_i = 0$
Faktisk outcome	$y_i = 1$	228	331
	$y_i = 0$	518	3287

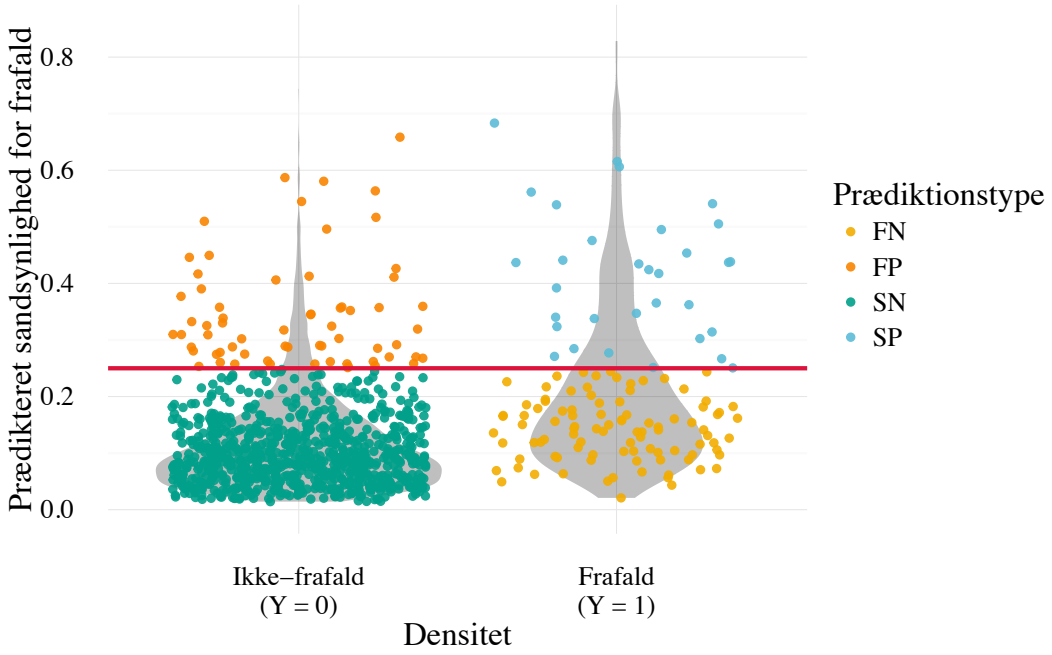
Når man skal fortolke resultaterne og fx forholde sig til, om 331 oversete frafald er mange eller få, afhænger det af konteksten for prædiktionsmodellen. Måske 331 oversete frafald ikke er så slemt sammenlignet med et alternativ, hvor man slet ikke prædikterer frafald og derfor "overser" alle frafald. Men havde outcome været et andet, såsom hvis vi havde prædikteret risikoen for recidiv hos en fængslet person, som vi overvejede at prøveløslade, så var det givetvis meget problematisk at overse 331 tilfælde. Konteksten – herunder den politiske kontekst – har altså betydning, når vi tolker en confusion-matrix.

Samtidig har fastsættelsen af tærskelværdien, δ, afgørende betydning for antallet af de fire forskellige prædiktionstyper. Når vi ændrer vores tærskelværdi, ændrer confusion-matricen sig også. Når vi fx sænker tærsklen, skal der "mindre til", for at vi klassificerer en studerende som frafaldstruet. Det vil resultere i, at vi prædikterer flere studerende som frafaldstruede. Det betyder på den ene side, at vi opfanger flere sande positive. Det betyder dog også, at vi får flere falske positive "med i købet". Denne sammenhæng mellem prædiktionstyperne og tærskelværdien kan vi illustrere med data fra Metropol som i figur 3.2 på næste side inspireret af Harmsen & Enggaard (2016).

I figuren inddeles alle observationer i to søjler på førsteaksen efter deres faktiske outcome, henholdsvis ikke-frafald ($y = 0$) og frafald ($y = 1$). På andenaksen ses observationernes prædikterede sandsynlighed for frafald. Vi er interesserede i at maksimere antallet af grønne og blå prædiktioner (SN + SP), og begrænse antallet af orange og gule prædiktioner (FP + FN). Den vandrette røde linje angiver en fastsat tærskelværdi. Det ses, at når vi hæver tærsklen, altså hæver den røde linje, så bliver orange prædiktioner til grønne prædiktioner (ønskeligt). Samtidig bliver de blå prædiktioner imidlertid til gule prædiktioner (ikke ønskeligt).

Hvis vores prædiktionsmodel havde været perfekt, ville vi kunne have sat tærsklen sådan, at der kun var blå og grønne prikker. Det ses af figur 3.2 på den følgende side, at det ikke kan lade sig gøre. Densiteten af observationerne fordeler sig dog forskelligt i de to søjler. Blandt den venstre søjle af studerende, der ikke frafalder,

Figur 3.2: Densitets-histogram over prædikteret vs. faktisk frafald



er densiteten af observationer meget høj ved de lave prædikterede sandsynligheder for frafald. Densiteten er mere jævnt fordelt for den højre søjle af studerende, der frafalder. Den forskellige fordeling af densiteterne har betydning for den hastighed, hvormed FP bliver til SN vis-a-vis den hastighed, hvormed SP bliver til FN, når vi ændrer tærskelværdien. Forholdet mellem de respektive prædiktionstyper ændrer sig ikke med samme hastighed, når vi ændrer tærsklen fra 0,1 til 0,2 sammenlignet med at ændre den fra 0,3 til 0,4.

Der eksisterer altså et forholdsvis kompliceret tradeoff mellem prædiktionstyperne. Det kan vi samle op med to performancemål, henholdsvis *sand positiv-raten* og *falsk positiv-raten* (Witten et al. 2005: 162–163). Målene er definerede i tabel 3.3.

Tabel 3.3: Prædiktionstyper og performancemål

		Prædikteret outcome		Performancemål
		$\hat{y}_i = 1$	$\hat{y}_i = 0$	
Faktisk outcome	$y_i = 1$	**SP**	FN	*Sand positiv-rate* $= \frac{SP}{SP+FN} = \frac{SP}{\sum y_i=1}$
	$y_i = 0$	**FP**	SN	*Falsk positiv-rate* $= \frac{FP}{FP+SN} = \frac{FP}{\sum y_i=0}$

Sand positiv-raten (SPR) er antallet af sande positive delt med antallet af alle faktiske positive. Vi husker på, at faktisk positive er defineret som $y_i = 1$, hvilket her vil sige faktiske frafald. I vores case betyder SPR derfor: Hvor stor en andel af de faktiske frafald fanger vi? Derfor kaldes SPR også for *probability of detection*, og vi vil gerne have en så høj SPR som muligt (Fawcett 2006). Falsk positiv-raten (FPR) er alle falske positive delt med antallet af faktiske negative. I vores case kan FPR derfor tolkes som: Hvor stor en andel af de faktiske ikke-frafald forudsiger vi fejlagtigt til at droppe ud? FPR kaldes også for *probability of false alarm*, og vi vil gerne have en så lav FPR som muligt (Fawcett 2006). For en given prædiktionsmodel vil SPR og FPR variere på tværs af forskellige værdier for tærskelværdien δ. Hver tærskelværdi modsvares med andre ord af en SPR og en FPR. I figur 3.2 på forrige side svarer SPR til de grønne prikkers andel af venstre søjle, mens FPR er de blå prikkers andel af højre søjle.

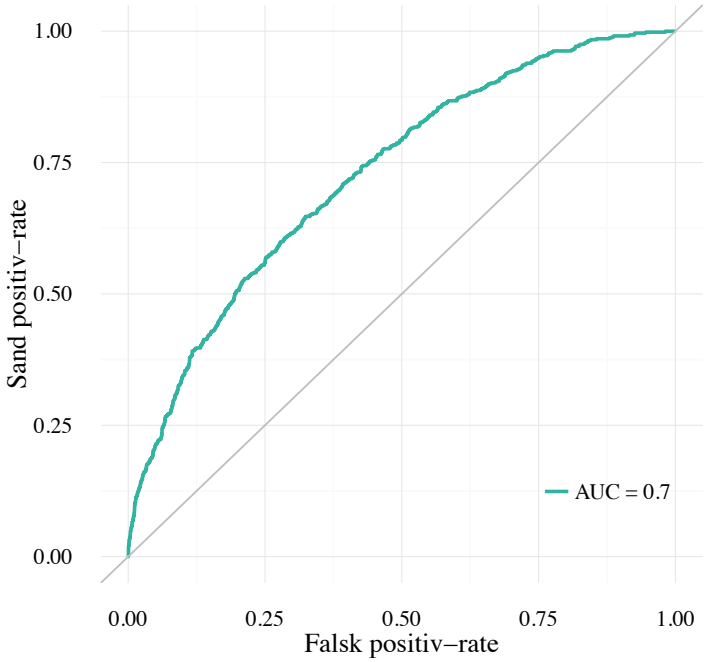
Hvis vi ville, kunne vi nemt sætte tærskelværdien, så vi med vores model sikrede en SPR på 1 og dermed indfangede alle frafald. Det kunne vi gøre ved at sætte tærskelværdien til 0 og gætte på, at alle studerende frafalder. Det ville dog modsvares af en meget høj FPR, fordi vi nu ville prædiktere en masse frafald blandt studerende, som ikke falder fra. Vi kunne også sagtens sikre en FPR på 0 ved at gøre det modsatte – sætte tærsklen til 1 og dermed forudsige, at ingen frafalder. Det ville dog give en meget lav SPR, fordi vi ikke ville identificere nogen af frafaldene. Situationerne illustrerer sammenhængen mellem SPR og FPR: Vi ønsker en høj SPR og lav FPR, men når vi sænker tærskelværdien, så stiger de begge, og når vi hæver tærskelværdien, så falder de begge. Der er med andre ord tale om et tradeoff på tværs af tærskelværdier.

ROC-kurver og AUC

Vi kan opsummere tradeoff'et mellem SPR og FPR ved at plotte dem mod hinanden i to dimensioner over hele spændet af tærskelværdier fra 0 til 1. Resultatet er en såkaldt *Receiver Operating Characteristic*-kurve (ROC-kurve) som angivet i figur 3.3 (Witten et al. 2005: 168-171).

ROC-kurven i figur 3.3 viser sammenhængen mellem SPR og FPR over spændet af tærskelværdier δ for en given prædiktionsmodel. Eksempelvis skærer den grønne ROC-kurve i figuren omtrent punktet (0.2, 0.5), hvor FPR = 0,2 og SPR = 0,5. Det kan vi tolke sådan, at vi med prædiktionsmodellen korrekt kan identificere 50 pct. faktisk frafaldstruede studerende, og at vi i samme ombæring fejlagtigt vil klassificere 25 pct. af de ikke-frafaldstruede studerende som frafaldstruede. Det kan bemærkes, at selve δ ikke kan aflæses af ROC-kurven, da den ikke indgår på akserne, selvom den modsvares af en SPR og FPR.

Figur 3.3: Eksempel på ROC-kurve



Når vi sænker tærskelværdien for vores prædiktionsmodel, dvs. gør tærsklen mere lempelig, vil SPR blive større, hvormed vi bevæger os op ad andenaksen i figuren. Samtidig vil FPR også stige, hvormed vi bevæger os ud ad førsteaksen. Disse to bevægelser gør, at vi samlet bevæger os "opad" på ROC-kurven mod punktet (1,1) i øverste højre hjørne. I dette hjørne forudsiger vi næsten alle til at falde fra, hvormed vi opdager alle frafald, men også får mange falske positive med i købet. Når vi omvendt hæver tærskelværdien, dvs. gør tærsklen mere restriktiv, bevæger vi os "nedad" på ROC-kurven mod punktet (0,0) i nederste venstre hjørne (Fawcett 2006: 862–863). Her får vi en mere forsigtig model, der kun opdager få frafald, men heller ikke udløser så mange falske alarmer.

I figuren repræsenterer den diagonale linje en *line of no discrimination*, som vil være resultatet af en prædiktionsmodel baseret på tilfældige gæt. Den har en hældning på én, hvilket vil sige, at FPR stiger lige så hurtigt som SPR. Det er vi ikke interesserede i – for en god prædiktionsmodel, som er god til at klassificere sine observationer rigtigt, vil SPR indledningsvist stige hurtigere end FPR, hvormed ROC-kurven ligger så meget over den diagonale linje som muligt. Derfor er en stejl ROC-kurve også eftertragtet. En perfekt klassifikationsmodel ville skære punktet (0,1) i øverste venstre hjørne, som angiver at SPR = 1 og FPR = 0, det vil sige perfekt klassifikation. Derved kan vi bruge arealet under ROC-kurven, *Area Under Curve*, AUC, som et samlet mål for en prædiktionsmodels performance over hele spændet af tærskelværdier (Witten et al. 2005: 173).

Her er AUC = 1 perfekt klassifikation, og AUC = 0,5 en baseline i form af arealet under den diagonale linje. En prædiktionsmodel vil derfor have en AUC i intervallet $0{,}5 < \text{AUC} < 1$. Det skal understreges, at AUC er et heuristisk mål, som er nyttigt til en umiddelbar sammenligning af modeller, men ikke er perfekt. Rationalet bag AUC er, at jo større areal under ROC-kurven jo bedre – og det er tilnærmelsesvist korrekt (Witten et al. 2005: 173). To meget forskelligt udseende kurver vil dog kunne have det samme underliggende areal. Intet tal vil alene kunne opsamle det komplicerede tradeoff mellem SPR og FPR (Witten et al. 2005: 173). Det kan kun lade sig gøre ved at supplere analysen med en todimensional fremstilling såsom ROC-kurven.

Med ROC-kurverne og AUC har vi nu et mål, som vi kan bruge til at evaluere en prædiktionsmodels performance i forhold til de fire prædiktionstyper over hele spændet af tærskelværdier. Det er nyttigt til at sammenligne modeller på tværs af algoritmer, datasæt og modelspecifikationer. Det får vi brug for i vores analyse af casen Metropol, hvor vi søger at finde frem til den bedste model.

Vi har i dette afsnit set, hvordan forskellige tærskelværdier hænger sammen med et forskelligt antal prædiktionstyper opsummeret ved SPR og FPR. Når man har fundet frem til den bedste prædiktionsmodel ved hjælp af AUC, har man behov for at fastsætte én bestemt tærskelværdi. I analysen vender vi tilbage til, hvordan denne udfordring kan gribes an, når vi skal anvende en prædiktionsmodel i praksis.

3.2 Maskinlæring som tilgang

Efter kun kort at have introduceret begrebet maskinlæring i indledningen er dette afsnit dedikeret til en mere uddybende gennemgang af maskinlærings metodiske aspekter. Vi tilstræber at begrænse den matematiske notation og i stedet fokusere på intuitionen bag centrale maskinlæringskoncepter.

3.2.1 Typer af maskinlæring

En ulempe ved betegnelsen maskinlæring er, at den anvendes med en flerhed af betydninger og i nogen grad overlapper med betegnelser som fx data mining, statistical learning og data science, der alle vedrører genkendelse af mønstre i data (Varian 2014: 5). Jævnfør afgrænsningen i indledningen forstår vi med maskinlæring en tilgang, hvor omdrejningspunktet for mønstergenkendelsen er prædiktion. Dermed refererer vi også til den gren af maskinlæring, som kaldes for *superviseret* maskinlæring, hvor der

eksisterer et *outcome*[4], ofte anført som y, som vi ønsker at prædiktere eller at klassificere data efter. Med superviserede metoder ønsker vi at finde frem til den funktion f, der bedst kan forklare eller prædiktere y ved $\hat{y} = f(\mathbf{X})$ givet et input af *variable*[5], ofte angivet ved en matrix $\mathbf{X}$.

Modstykket til superviserede metoder er usuperviserede metoder[6], som ikke har et prædefineret outcome, vi er interesserede i at forudsige. I stedet ønsker man at forstå mønstre eller klynger i data. Det kunne fx være at identificere ideologiske dimensioner i politikeres taler (Schwarz et al. 2015; Lowe & Benoit 2013; Goet 2016).

En anden opdeling i maskinlæringslitteraturen er mellem regressions- og klassifikationsproblemer (James et al. 2013: 28–29). Der er tale om et regressionsproblem, når y er kontinuert, og et klassifikationsproblem, når y er kategorisk – som i vores tilfælde, hvor outcome er enten frafald eller ikke-frafald. I vores analyse af casen Metropol vil vi således anvende algoritmer, der kan håndtere et superviseret klassifikationsproblem. I de følgende afsnit vil vi gå i dybden med generelle koncepter, der vedrører, hvordan algoritmerne konkret fungerer.

3.2.2 Algoritmer og loss-funktioner

I opgavens indledning definerede vi algoritmer som et sæt af regler for, hvordan et givent problem løses. Den opgave, som algoritmerne løser, er helt overordnet at minimere en *loss-funktion*. En loss-funktion kan forstås som en funktion, der udtrykker et loss, dvs. en fejl, når vi klassificerer observationen i som $\hat{y}_i$, givet den sande værdi er y_i. Dét koncept kender vi fx fra OLS, hvor vi minimerer en loss-funktion i form af de kvadrerede residualer (James et al. 2013: 61–63, 71–75):

$$\hat{f}_{OLS} = \arg\min_{f} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \sum_{j=1}^{J} \hat{\beta}_j x_{ij})^2 \tag{3.1}$$

På samme måde minimerer både modeller baseret på logistisk regression og træbaserede modeller også loss-funktioner[7]. I afsnit 3.3 gennemgår vi de fire algoritmer, som vi

[4] I maskinlæringslitteratur er der egentlig tradition for at omtale outcomes som *labels*. Vi har valgt at holde fast i termet outcome, nu hvor vi anvender maskinlæring i en politologisk kontekst.

[5] I maskinlæringslitteraturen er *features* egentlig betegnelsen for det, vi kender som regressor, prædiktor eller forklarende og uafhængige variable i almindelig regression. Vi holder fast i sidstnævnte at samme grund som ovenfor.

[6] Det kan også være en kombination af de to, men det vil vi begrænse os fra at komme ind på her.

[7] Strengt taget foretages logistisk regression typisk ved *maximum likelihood*, hvor en negativ loss-funktion minimeres (James et al. 2013: 130–134).

bruger i vores analyse af casen Metropol, og her vil vi komme ind på de enkelte algoritmers loss-funktioner i det omfang, det bidrager til den samlede forståelse af algoritmernes centrale karakteristika. Med vores fokus på intuitionen præsenterer vi ikke en grundig gennemgang for alle algoritmernes loss-funktioner; det er unødvendigt for vores analyse. Vi præsenterer dog nogle af loss-funktionerne, fordi de bidrager til at forstå, hvordan algoritmerne baseret på klassifikationstræer adskiller sig fra vores mere velkendte baseline-model baseret på logistisk regression.

Selve konceptet loss-funktioner er altså ikke noget nyt, der følger med brugen af maskinlæring – loss-funktionen angivet ovenfor i formel 3.1 er eksempelvis ikke fremmed for os. Hvad der til gengæld er nyt er den måde, som loss-funktionerne minimeres på. Det vil vi folde ud i de følgende afsnit, hvor vi først vil kaste et blik på den datagenererende proces for at forstå, hvad det er, vi fitter vores modeller til, når vi minimerer loss'et.

3.2.3 Den datagenererende proces og grænser for prædiktionsperformance

Når vi opstiller en prædiktionsmodel leverer vi forudsigelser af et outcome, $\hat{y}$, og formålet er, at disse i videst muligt omfang matcher det faktiske outcome, y. Når en prædiktionsmodel leverer prædiktioner på baggrund af en række observationer er $\hat{y}$ dog sjældent lig med y i alle tilfælde – prædiktionsmodellen er fejlbarlig. På baggrund af James et al. (2013) gennemgår vi her, hvordan man kan nedbryde denne fejlbarlighed i to mindre dele ved at sammenholde den datagenererende proces med den model, vi fitter. Det tjener dels til at vise, hvad formålet er med at teste flere forskellige algoritmer i vores analyse, og dels til at vise, at der er en grænse for, hvor god en prædiktionsmodel kan blive.

I prædiktion (såvel som estimation) kan vi antage en datagenererende proces, som vi gerne vil modellere. Den kunne fx se ud som følger:

$$y = f(\mathbf{X}) + \epsilon$$

Her er y det outcome, som vi er interesserede i. y genereres som en funktion af en række forklarende variable, $\mathbf{X}$, samt fejlleddet ϵ, der er uafhængigt af $\mathbf{X}$ og varierer tilfældigt, og derfor i gennemsnit er nul (James et al. 2013: 15–16). I vores case kan vi forstå den tilfældige variation som et resultat af, at frafald ikke er et deterministisk outcome. Når vi vil forudsige y kan vi opstille en prædiktionsmodel, hvilket vi kan skrive på følgende form:

$$\hat{y} = \hat{f}(\mathbf{X})$$

Her er $\hat{y}$ vores prædiktioner af y, og funktionen $\hat{f}$ er vores tilnærmelse af den datagenererende funktion f (James et al. 2013: 17–18). Hvor god en forudsigelse $\hat{y}$ er af y, afhænger af to faktorer, som henholdsvis er en *reducerbar* fejl og en *ikke-reducerbar* fejl. Den første faktor er, hvor god en tilnærmelse $\hat{f}$ er på f. Generelt er $\hat{f}$ ikke et perfekt estimat på f, men fejlen er reducerbar. Eksempelvis vil forskellige funktionelle former i større eller mindre grad svare til den funktionelle form i den datagenererende proces. Men selv hvis $\hat{f}(\mathbf{X})$ er et perfekt estimat for $f(\mathbf{X})$, så $\hat{f}(\mathbf{X}) = f(\mathbf{X})$, hvormed vores prædiktioner tager formen $\hat{y} = f(\mathbf{X})$, vil vores prædiktioner indeholde en fejl (James et al. 2013: 17–19). Det skyldes den anden faktor, som er, at y også er en funktion af ϵ, der per definition ikke kan forudsiges med $\mathbf{X}$. Variationen i ϵ har derfor også betydning for vores prædiktioner. Denne fejl omtales som ikke-reducerbar, fordi uanset hvor godt vi estimerer f, vil dette ikke reducere den fejl, der følger af variationen i ϵ. Denne tilfældige variation i ϵ omtales også som støj, mens den datagenererende proces også omtales som signal (James et al. 2013: 17–19).

Vi kan nu forstå afvigelsen mellem en datagenererende proces (som vi dog ikke kender i praksis og derfor heller ikke kan måle eller verificere) og en prædiktionsmodel som[8]:

$$E\left[y - \hat{y}\right]^2 = E\left[f(\mathbf{X}) + \epsilon - \hat{f}(\mathbf{X})\right]^2 = \left[f(\mathbf{X}) - \hat{f}(\mathbf{X})\right]^2 + Var(\epsilon) \tag{3.2}$$

Det første led i formlen er den forventede kvadrerede afvigelse mellem den faktiske værdi y og den forudsagte værdi $\hat{y}$. Udtrykket kaldes også *mean squared error* (MSE) og er et mål for, hvor god en prædiktionsmodel er til at levere forudsigelser[9], som svarer til de faktiske værdier i outcome y (James et al. 2013: 17–19, 29–33). To pointer er centrale at tage med videre. Den første er, at fejlen $[f(\mathbf{X}) - \hat{f}(\mathbf{X})]^2$ er reducerbar, og hele formålet med at afprøve forskellige algoritmer i opgaven her er at minimere denne fejl for at levere så gode prædiktioner som muligt. Den anden pointe er, at uanset hvor god en tilnærmelse en prædiktionsmodel $\hat{f}(\mathbf{X})$ er på den faktiske $f(\mathbf{X})$, så indgår fejlleddet ϵ i den proces, der genererer y, og da denne fejl er ikke-reducerbar sætter den en øvre grænse for, hvor god en prædiktionsmodel det er muligt at opstille (James et al. 2013: 17–19).

[8]Det bør bemærkes, at formaliseringen gælder under en antagelse om, at både $\mathbf{X}$ og $\hat{f}$ er fikserede. I afsnit 3.2.5 på side 36 om bias og varians folder vi en situation ud, hvor vi ikke gør os denne antagelse.

[9]Det kan bemærkes, at målet MSE typisk bruges i regressionsproblemer, men ikke i klassifikationsproblemer, hvor et mere retvisende mål er *error rate* givet ved $Err_i = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$. Vi har valgt at illustrere logikken med MSE, fordi MSE i høj grad minder om *ordinary least squares*, OLS, som vi allerede kender fra estimationssammenhæng. Intuitionen bag MSE og error rate er den samme.

3.2.4 In-sample vs. out-of-sample performance

Som vi så tidligere udtrykker loss-funktioner, hvor godt vores model fitter til data: jo lavere loss, jo bedre fit, og derfor minimerer vi loss-funktioner. Det gælder både i estimations- og prædiktionssammenhæng. Der er dog en helt central væsensforskel mellem tilgangene i de to sammenhænge. Mens formålet ved estimation er at minimere loss'et in-sample, er formålet ved prædiktion af minimere loss'et out-of-sample. In-sample data er det data, som modellen fittes til. Out-of-sample data er nyt og hidtil uset data (James et al. 2013: 29–30).

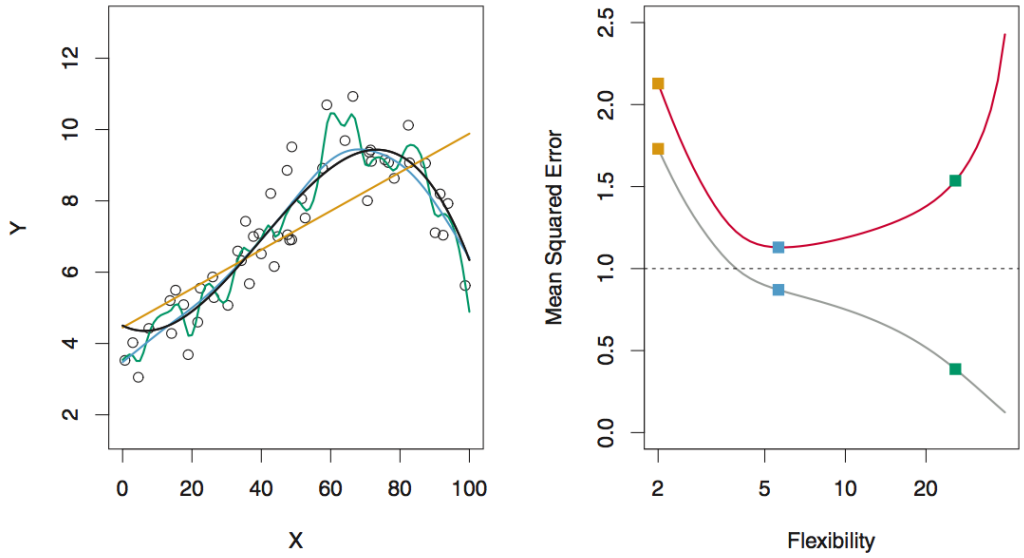
Årsagen til de forskellige formål er, at vi i estimation typisk er interesserede i at forstå og forklare sammenhænge, fx hvilke personlige karakteristika, som hænger sammen med frafald. Her vil et højt *in-sample* fit være et udtryk for en god performance for en model. Vi interesserer os for noget andet i prædiktionssammenhæng, hvor vi gerne vil prædiktere et outcome. I vores case er det fx et ønske om at forudsige frafald blandt nystartede studerende. Formålet er her et andet, fordi vi ikke er interesserede i at prædiktere frafald blandt de tidligere studerende i datasættet – blandt dem ved vi jo allerede, hvem der faldt fra, og hvem der ikke gjorde (James et al. 2013: 29–30). I stedet ønsker vi her at maksimere performance *out-of-sample*, sådan at vores model leverer de bedst mulige forudsigelser, når vi prædikterer frafaldet på nyt og uset data, fx data om de kommende studerende.

Der er imidlertid ikke nogen garanti for, at et godt in-sample fit er ensbetydende med et godt out-of-sample fit (James et al. 2013: 29–31). Årsagen er, at når vi maksimerer fittet in-sample, så fitter vi også til mønstre, som givetvis ikke vil findes i data om de nye studerende (James et al. 2013: 29–33). Det er samtidig den tekniske årsag til, at den vante estimationstilgang i en regressionsnalyse, hvor vi fx minimerer de kvadrerede residualer in-sample med OLS, ikke nødvendigvis er optimal i prædiktionssammenhæng.

Et smart kneb i prædiktionssammenhæng tillader os at maksimere out-of-sample performance. Det går jf. afsnit 3.1.1 ud på, at vi inddeler vores datasæt i henholdsvis et *træningssæt* og et *testsæt*. Træningssættet består fx af et tilfældigt sample på 80 pct. af hele datasættet, mens testsættet består af de resterende 20 pct. Testsættet holder vi helt uden for den proces, hvorigennem vi fitter vores model. Dermed behandler vi testsættet som out-of-sample. På den måde kan vi træne modellen på de 80 pct. af data, og teste prædiktions-performance out-of-sample på de 20 pct. (James et al. 2013: 29–33). Den underliggende antagelse er selvfølgelig, at testsættet er en god tilnærmelse af det "virkelige testsæt", dvs. de nye årgange af studerende, som vi af gode grunde ikke kan måle modellens performance for endnu. Hvorvidt denne antagelse er rimelig, vender vi tilbage til i diskussionen.

I sidste afsnit introducerede vi begrebet *mean squared error* (MSE) som et udtryk for en prædiktionsmodels fejlbarlighed. Begrebet er også nyttigt i afsnittet her til at illustrere sammenhængen mellem in-sample performance og out-of-sample performance. Høj in-sample performance betyder lav trænings-MSE, mens høj out-of-sample performance betyder lav test-MSE. Sammenhængene illustreres i figur 3.4.

Figur 3.4: Mean squared errors for trænings- og testsæt



Note: Baseret på (James et al. 2013: 31).

I venstre side af figuren vises et træningssæt. Den sorte kurve viser den datagenererende funktion, og datapunkterne ligger spredt omkring denne, da funktionen er givet ved $y = f(\mathbf{X}) + \epsilon$ og altså indeholder tilfældig støj i form af ϵ. I venstre side at figuren vises ydermere tre forskellige modeller, som er fittet til datapunkterne. Den mest restriktive model er den lineære model, den orange model, mens den grønne model er den mest fleksible model. Ind imellem de to ligger den blå model, som minder nogenlunde om den datagenererende sorte kurve.

I højre side af figuren vises modellernes MSE'er. De farvede punkter på kurverne modsvarer de farvede modeller i venstre side af figuren. Den grå kurve viser modellernes trænings-MSE. Det ses, at trænings-MSE'en kun bliver mindre, jo mere fleksible modellerne bliver. Den grønne model har således den laveste trænings-MSE, fordi den nærmest perfekt fitter alle punkterne i træningssættet. Den røde kurve viser modellernes test-MSE. Her er det anderledes den blå model, som har den laveste test-MSE. Den lineære orange model har en relativt lav test-MSE, fordi den lineære model er for restriktiv og ikke er en specielt god tilnærmelse på formen af den sorte datagenererende model. Vi siger, at modellen *underfitter*. Den meget fleksibe grønne

model har også en lav test-MSE, fordi den har fittet til støj i venstre side af figuren, hvilket giver dårlige prædiktioner out-of-sample i testsættet. Her siger vi, at modellen *overfitter* (James et al. 2013: 32–33).

Mens den grå kurve blot er aftagende og viser dalende trænings-MSE for mere fleksible modeller, viser den røde kurve anderledes en U-form i kurvernes test-MSE. Modellernes out-of-sample performance er dermed hverken optimal for en model, der er for restriktiv eller for fleksibel. Sammenhængen mellem den grå og den røde kurve er universel, når vi fitter modeller til et trænings- og testsæt, uafhængigt af datasæt og uafhængigt af hvilken model, som fittes (James et al. 2013: 31).

Når vi forsøger at finde frem til den model med den bedste out-of-sample performance, der som bekendt er formålet i prædiktion, er balancen mellem risikoen for underfitting og risikoen for overfitting afgørende. Særligt står begrebet overfitting centralt i maskinlæringslitteraturen. Det skyldes, at det med meget fleksible modeller – som dem, vi anvender i vores analyse – er meget nemt at fitte data meget præcist. Her kan en mindre fleksibel model potentielt levere en lavere test-MSE og højere out-of-sample performance (Friedman 2001: 1203; James et al. 2013: 32). For at blive i stand til at håndtere underfitting og overfitting vil vi i næste afsnit først vise, hvordan vi kan forstå balancen mellem de to.

3.2.5 Bias-variance tradeoff

Den U-formede sammenhæng mellem en models grad af fleksibilitet og dens performance out-of-sample skyldes to konkurrerende hensyn, når vi fitter en model til data. Der er tale om et *bias-variance tradeoff*, hvilket er emnet for afsnittet her.

Det kan vises, at den forventede MSE i testsættet for en given observation, x_0, kan nedbrydes til summen af tre elementer[10] (James et al. 2013: 33–36; Friedman et al. 2009: 37-38):

$$E\left[y_0 - \hat{f}(x_0)\right]^2 = \underbrace{E\left[\left(\hat{f}(x_0) - E\left[\hat{f}(x_0)\right]\right)^2\right]}_{Var(\hat{f}(x_0))} + \underbrace{\left(E\left[\hat{f}(x_0)\right] - f(x_0)\right)^2}_{Bias(\hat{f},f)^2} + Var(\epsilon) \quad (3.3)$$

[10]Den skarpe observatør vil bemærke, at formlen for MSE her, formel 3.3, adskiller sig fra den tidligere formel for MSE, formel 3.2 på side 33. Det skyldes, at $\hat{f}$ er tilladt at variere i nærværende formel, men var fikseret tidligere og derfor ikke varierede.

Her udtrykker det første led $E[y_0 - \hat{f}(x_0)]^2$ den forventede MSE i testsættet, dvs. et udtryk for den forventede fejl for prædiktion af y_0 out-of-sample med $\hat{f}(x_0)$. Variansen i $\hat{f}(x_0)$, $Var(\hat{f}(x_0))$, kan vi forstå som et udtryk for, hvor meget vores estimerede $\hat{f}$ ændrer sig, hvis vi estimerer modellen med et andet træningssæt. Når vi træner en model på forskellige træningssæt vil de resulterende modeller, $\hat{f}$, variere, fordi data varierer mellem træningssættene. Denne variation er dog ikke ønskværdig – vi foretrækker, at $\hat{f}$ i mindst muligt omfang er afhængig af det datasæt, som vi træner modellen på, fordi det sandsynliggør, at $\hat{f}$ er en tilnærmelse af den datagenererende f (James et al. 2013: 34).

En konsekvens af en høj varians i $\hat{f}$ er, at $\hat{f}$ vil være meget sensitiv over for ændringer i data. Det betyder, at hvis datapunkter ændrer sig datasættene imellem, så vil det have stor indflydelse på $\hat{f}$. Generelt har mere fleksible modeller en højere varians. Vi kan her kaste et blik tilbage på figur 3.4 på side 35. Den grønne kurve i figuren er meget fleksibel og har en høj varians. Det betyder, at hvis bare et eller få datapunkter ændrer sig, så vil det betyde en stor ændring i $\hat{f}$ for den grønne kurve. Og det er meget sandsynligt, at datapunkterne vil ændre sig i et nyt datasæt, eksempelvis testsættet, fordi punkterne i figuren varierer tilfældigt omkring den sorte kurve, fordi fejlleddet ϵ også indgår i den datagenererende proces (James et al. 2013: 34–35).

Bias i formlen 3.3 på forrige side refererer til den systematiske forskel mellem f og vores estimerede $\hat{f}$. Bias kan forstås som en fejl, der følger af, at vi forsøger at modellere en kompleks sammenhæng fra virkeligheden med en meget simplere model, fx en lineær model. Det vil introducere et bias i modellen, hvis sammenhængen ikke er lineær i virkeligheden, hvilket den formentlig sjældent vil være for fx komplekse sociale sammenhænge (James et al. 2013: 35). Her kan vi også kaste et blik tilbage på figur 3.4. Den orange kurve er lineær og meget lidt fleksibel, og det ses, at den funktionelle form er en dårlig tilnærmelse af den funktionelle form på den datagenererende proces, den sorve kurve. Det betyder, at uanset hvor mange flere observationer, vi tilføjer i træningssættet, så vil den lineære orange kurve stadig være dårlig til at fitte datapunkterne. Den lineære model er med andre ord biased. Generelt vil mere fleksible modeller i mindre omfang være biased end mere restriktive modeller (James et al. 2013: 35).

To yderligere pointer kan ses af dekomponeringen af den forventede test-MSE ovenfor i formel 3.3. For det første ses det, at siden både varians og kvadreret bias er positive termer, så kan den forventede test-MSE aldrig blive mindre end variansen i fejlleddet ϵ. Denne pointe er i tråd med afsnit 3.2.3 om grænser for prædiktions-performance. For det andet ses det, at når vi skal finde den model, som har den mindste, forventede test-MSE, og således den bedste out-of-sample performance, skal vi samtidigt minimere både modellens varians og bias.

Når vi minimerer bias og varians samtidigt opstår et tradeoff: mere fleksible modeller har højere varians, men lavere bias – og omvendt for mere restriktive modeller. Hvor hurtigt henholdvis variansen bliver større og bias bliver mindre, når en model gøres mere fleksibel, varierer fra model til model og fra datasæt til datasæt. Typisk vil det dog være sådan, at bias indledningsvist mindskes relativt hurtigere, end variansen bliver større, i takt med at modellen gøres mere fleksibel (James et al. 2013: 35–36). Derfor vil den forventede test-MSE samlet set blive mindre. Det vil forsætte indtil et punkt, hvor nytten af mere fleksibilitet ikke er specielt gavnligt, fordi formen på $\hat{f}$ er en udmærket tilnærmelse af f. Her vil bias ikke blive meget mindre, selvom vi øger modellens fleksibilitet. Til gengæld vil variansen begynde at stige kraftigt, fordi vi nu begynder at fitte til støj, hvormed den samlede forventede test-MSE vil begynde at stige. Det er denne sammenhæng, som giver U-formen på den forventede test-MSE som funktion af en models fleksibilitet, som vi så i figur 3.4 på side 35 (James et al. 2013: 35–36).

Når vi ønsker at maksimere out-of-sample performance ligger kunsten således i samtidig at opnå en så lille varians og et så lille bias som muligt. Tradeoff'et mellem bias og varians står meget centralt både i estimations- og prædiktionssammenhæng. Der er dog den vigtige forskel, at vi med maskinlæring som tilgang kan håndtere tradeoff'et empirisk, hvilket er temaet for det følgende afsnit.

3.2.6 Regularisering

Når vi skal finde frem til den optimale grad af fleksibilitet for en given model og dermed afveje bias og varians, kan vi bruge teknikker under fællesbetegnelsen *regularisering*. Teknikkerne har til formål at begrænse en models fleksibilitet og kompleksitet, og det er der flere måder at gøre på.

En måde at gøre det på er ved at kontrollere fleksibiliteten direkte med parametre, der for en given model begrænser, hvor fleksibelt modellen kan fitte data. Sådanne regulariseringsparametre er modelspecifikke og derfor forskellige fra model til model. Når vi senere introducerer de fire algoritmer, vi anvender i vores analyse, introducerer vi samtidig deres frie parametre til regularisering. Ud over disse parametre er en supplerende og ikke-modelspecifik måde at regularisere modellerne på ved *shrinkage*, hvor modellerne straffes for deres kompleksitet direkte i deres loss-funktion. Vi vil her illustrere tankegangen med såkaldt L2-regularisering af en lineær model. Vi bruger dette eksempel, fordi vi selv anvender L2-regularisering i vores analyse, og samtidig fordi teknikken er udbredt – eksemplet er intuitivt og logikken generisk (James et al. 2013: 203–204, 214–215).

Vi har set, at en model kan lide under højt bias som følge af at være for restriktiv, ligesom den kan lide under høj varians som følge af at være for fleksibel – begge med dårlig out-of-sample performance til følge. L2-regularisering fungerer ved at introducere en såkaldt regulariserings-term, som biaser modellen ved at begrænse dens fleksibilitet, men som følge heraf samtidig begrænser modellens varians. På denne måde introduceres en term, som indirekte modellerer modellens bias og varians. Lad os se på et konkret eksempel.

Vi kan forestille os en model, der beskriver frafald på en tænkt professionshøjskole. Lad y = *risiko for frafald*, x_1 = *alder* og x_2 = *køn*. Inspireret af Kleinberg et al. (2015) kan vi videre forestille os en regressionsmodel baseret på OLS, der eksempelvis giver koefficienterne $\hat{\beta}_1 = 1 \pm 0,02$ og $\hat{\beta}_2 = 15 \pm 40$. Det vil give prædiktionsmodellen $\hat{y} = x_1 + 15x_2$. Bekendte med den store usikkerhed på $\hat{\beta}_2$ vil vi dog være interesserede i at vægte koefficienten $\hat{\beta}_2$ lavere end $\hat{\beta}_1$, når vi foretager prædiktioner out-of-sample (Kleinberg et al. 2015: 492). Begge dele introducerer et bias i modellen, men med prædiktion for øje kan modellen tænkes at levere bedre prædiktioner out-of-sample ved at mindske eller fjerne støjen fra $\hat{\beta}_2$ (Kleinberg et al. 2015). Det kan vi gøre ved L2-regularisering, som introducerer en regularingseringsterm i loss-funktionen[11], hvormed vi minimerer følgende funktion for observationerne i og parametrene j:

$$\hat{f}_{L2} = \arg\min_{f} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{3.4}$$

Det første led er velkendt – det er *residual sum of squares*, RSS, loss-funktionen i OLS. RSS bliver mindst, når β-koefficienterne fitter data bedst muligt. Det andet led er regulariseringstermen. Da større β-koefficienter giver mere variable prædiktioner, er logikken, at vi straffer modellen for større β-koefficienter. Udtrykket i formlen 3.4 bliver mindst, når størrelsen på β-koefficienterne nærmer sig 0. På den måde tilskynder regulariseringentermen til at mindske β-koefficienterne. Det giver et bias i modellen, men det giver samtidig en simplere model med mindre varians. Derfor har vi nu med udtrykket ovenfor en funktion, hvor tradeoff'et mellem bias og varians modelleres (Kleinberg et al. 2015; James et al. 2013: 203–204, 214–215, 217–224).

Konstanten λ er en fri parameter, som justerer, hvor stor betydning regulariseringstermen skal have. Ved at justere denne kan vi bestemme "prisen" på kompleksitet og

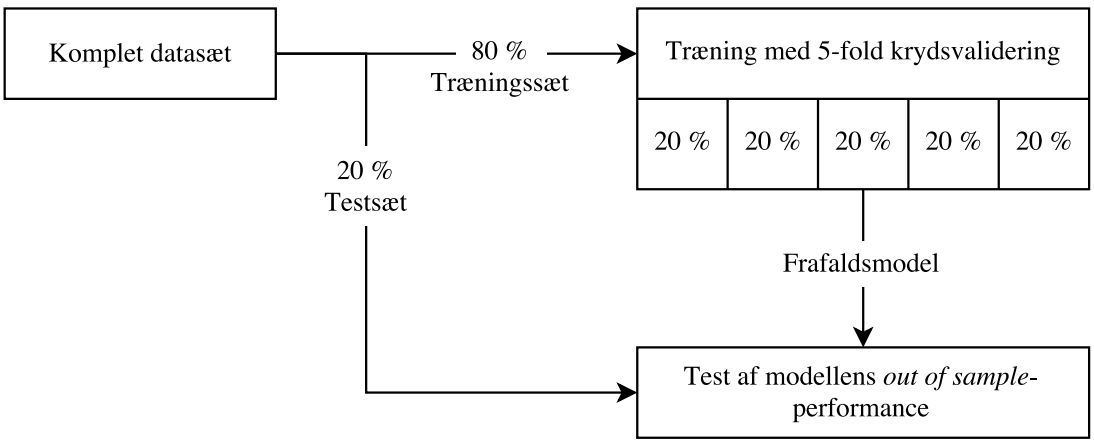
[11] L2-regularisering kaldes også for ridge-regression i en regressionssammenhæng. En anden udbredt regulariseringsteknik er L1-regularisering, også kaldet lasso-regression, der også kan bruges til at luge ud i de mindst betydningsfulde variable. Det er dog ikke et relevant hensyn i prædiktion, hvor vi problemfrit kan inkludere variable mhp. maksimal prædiktiv performance (James et al. 2013: 214–215, 217–220).

dermed prisen på højere varians – jo større λ, jo mindre varians og mere bias, og omvendt. Det er derfor afgørende for modellens performance out-of-sample, hvordan λ fastsættes. Selve idéen om regularisering er ikke et nyt bidrag fra maskinlæringslitteraturen – det er til gengæld fastsættelsen af parametren λ, fordi vi empirisk kan fastsætte parameteren, sådan at vi opnår den højeste out-of-sample performance. Det samme kan vi gøre med de øvrige modelspecifikke regulariseringsparametre nævnt først i dette afsnit. Processen kaldes *tuning* og er emnet for næste afsnit.

3.2.7 Tuning

Hvor man i estimationssammenhæng typisk vil fastsætte frie parametre teoridrevet eller ud fra antagelser, kan vi i prædiktionssammenhæng tune parametrene. Tuning går i al sin enkelhed ud på at afprøve forskellige kombinationer af parameterværdier, sådan at vi empirisk kan fastsætte den grad af fleksibilitet, der giver den laveste test-MSE og dermed den bedste performance out-of-sample (James et al. 2013: 175–183). I denne tuningproces ligger en del af maskinlærings empiriske læringselement. En udbredt metode er krydsvalidering, hvor man træner modellen på én del af træningssættet og evaluerer den på en anden (James et al. 2013: 32, 175–176). Dermed opnår man et estimat for out-of-sample performance ved forskellige kombinationer af parametre. Logikken i krydsvalidering er den samme, som ligger bag den overordnede opsplitning i trænings- og testsæt. Blot foretager man her en ekstra opsplitning af data inden for træningssættet. Testsættet holdes stadig uden for tuning-processen, da det jo skal fungere som nye data for den endelige model, og derfor ikke kan bruges til at fitte selve modellen. Den samlede logik bag vores fremgangsmåde med krydsvalidering fremgår af figur 3.5, hvor vi illustrerer, hvordan vi opdeler træningssættet i fem delmængder til brug i tuning-processen.

Figur 3.5: Overblik over håndteringen af vores datasæt



Ved at splitte træningssættet på denne facon laver vi såkaldt *k-fold krydsvalidering*, hvormed vi opdeler træningssættet i k folder, her 5. Typisk udføres krydsvalidering med $k = 5$ eller $k = 10$ (James et al. 2013: 181). Når modellen trænes med en given kombination af parametre udelader man én fold, og træner modeller på de resterende $k - 1$ folder af data. Herefter kan den udeladte fold bruges til at evaluere modellens performance, fx ved at beregne MSE i den k'te fold. Proceduren gentages k gange, én for hvert fold, hvorefter gennemsnittet af MSE'erne kan beregnes. Formålet med at gentage proceduren er at opnå robuste estimater for den MSE, som modellen kan ventes at have i det egentlige testsæt. Herefter kan parametrene justeres, hvorefter hele processen med krydsvalidering gentages, og modellernes MSE kan sammenlignes. Denne proces fortsætter, og slutteligt vælges den model med den kombination af parametre, som har den laveste MSE (James et al. 2013: 175–183, 227–228).

At nå frem til den optimale parameterkombination kan i praksis være både omstændeligt og udfordrende – særligt for algoritmer, der har mange frie parametre. I analysens afsnit 5.2.3 går vi i flere detaljer omkring, hvordan vi i praksis har tunet vores endelige frafaldsmodel. Den centrale indsigt i afsnittet her er, at vi med krydsvalidering empirisk kan tune vores modeller til den optimale kombination af parametre, som leverer den bedste forventede performance i testsættet.

3.3 Algoritmer

Vi har i det ovenstående gennemgået centrale teoretiske koncepter ved maskinlæring som tilgang. I dette afsnit vil vi beskrive de fire algoritmer, som vi afprøver i vores case. Når vi afprøver flere, skyldes det, at det ikke på forhånd er givet, hvilken model som performer bedst i det enkelte datasæt. Der findes ikke én algoritme, som er universelt bedst (James et al. 2013: 29).

Den første algoritme er logistisk regression. Ud over at være almindelig anvendt til maskinlæring er det en velkendt model i den politologiske værktøjskasse. Dermed fungerer den som en slags baseline for de tre andre algoritmer, som er nye i politologisk sammenhæng.

Den anden algoritme er et simpelt klassifikationstræ[12]. I sig selv har vi ikke høje forventninger til denne algoritme, der som regel bliver overgået af mere raffinerede metoder (Athey & Imbens 2016: 48–49). Vi gennemgår den imidlertid grundigt, idet klassifikationstræer er byggestenene for resten af de algoritmer, vi anvender i analysen.

[12] Algoritmen kan også anvendes til problemstillinger med et kontinuert skaleret outcome, i hvilket tilfælde den kaldes for et regressionstræ.

Hvor logistisk regression og klassifikationstræer kan kaldes for singulære algoritmer, kan de resterende to klassificeres som såkaldte ensembler. Frem for at basere deres prædiktion på en enkelt model tager de gennemsnittet af en lang række modellers prædiktioner. Ensembler leverer dermed, hvad man kan kalde konsensus-prædiktioner (James et al. 2013: 303). De to ensemble-algoritmer, vi anvender, er Random Forest og Gradient Boosted Trees, der på forskellig vis aggregerer prædiktionerne fra en lang række af simple klassifikatonstræer. Vi har valgt at fokusere på disse træbaserede algoritmer, fordi de er begyndt at vinde indpas i samfundsvidenskaben, og fordi eksisterende casestudier jf. afsnit 2.2 har vist, at de giver gode resultater i prædiktionssammenhæng[13] (Athey & Imbens 2016: 48–51).

3.3.1 Logistisk regression

I vores opgave tjener logistisk regression som en baseline-model, fordi den er velkendt i den politologiske værktøjskasse. Logistisk regression anvendes typisk i sammenhænge, hvor outcome-variablen er dikotom, såsom hvorvidt individer i et datasæt stemmer eller ej, eller hvorvidt studerende i et datasæt frafalder eller ej. Vi giver her en kort opsummering af algoritmens karakteristika.

Den primære grund til at bruge logistisk regression frem for en lineær model såsom OLS er, at den logistiske model – forskelligt fra OLS – kan give outputs, som kun ligger i intervallet fra 0 til 1[14]. Her kan 0 og 1 henholdsvis repræsentere ikke-frafald og frafald for en outcome-variabel y, hvormed det prædikterede outcome $\hat{y}$ tolkes som sandsynligheden for, at y er lig med 0 henholdsvis 1 givet $\mathbf{X}$. I en situation som denne er vi være interesserede i at begrænse outcome til intervallet fra 0 til 1, fordi det ikke er meningsfuldt, at sandsynligheden for $y = 1$ er over 1, eller at sandsynligheden for $y = 0$ er under 0 (James et al. 2013: 129–132).

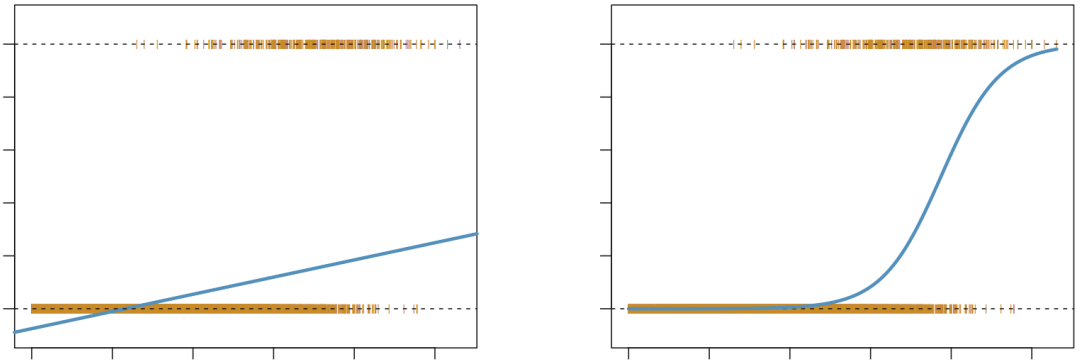
I figur 3.6 på næste side illustreres forskellen på at fitte et dikotomt outcome på baggrund af en kontinuert uafhængig variabel med henholdsvis en lineær og en logistisk model.

Som det ses af figuren giver den ikke-lineære funktionelle form i den logistiske model en S-formet kurve, som begrænser udfaldsrummet for $\hat{y}$ til intervallet fra 0 til 1.

[13] Andre klasser af algoritmer tæller bl.a. support vektor-maskiner og neurale netværk – for fyldestgørende overblik se fx Murphy (2012).

[14] Det følger af den logistiske regressions funktionelle form: $p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_k X_k}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_k X_k}}$ (James et al. 2013: 131–132).

Figur 3.6: Lineært og logistisk fit på data med dikotomt y og kontinuert X



Note: Baseret på (James et al. 2013: 131).

En logistisk regressionsmodel reguleres ikke som standard, og derfor er der heller ikke nogen parametre, som skal tunes. Modellen kan dog uden videre reguleres ved at tilføje en regulariseringsterm, eksempelvis med L2-regularisering, som vi så i afsnit 3.2.6. Når vi anvender modeller baseret på logistisk regression i analysen af casen Metropol, tjener de som nævnt som baseline, og derfor undlader vi at regularisere modellerne med det formål, at de i videst muligt omfang minder om den type modeller, som er almindelige i politologien. Den logistiske model ville dog ganske givet levere bedre resultater i analysen, hvis vi regulariserede og tunede den.

3.3.2 Klassifikationstræer

De træbaserede algoritmer adskiller sig væsentligt fra de regressionsmodeller, som vi er vante med i samfundsvidenskaben. Klassifikationstræer har en særlig appel, idet de kan modellere data meget komplekst og fleksibelt og alligevel bibeholde en intuitiv fortolkning. Helt overordnet fungerer klassifikationstræer ved at inddele observationerne i et datasæt på baggrund af en række ja/nej-spørgsmål om de uafhængige variable. Det kunne for eksempel være, om en studerende i datasættet er over 25 år gammel. Hvis ja, ryger den studerende over i den ene gruppe af observationer, og hvis nej, ryger den studerende over i den anden gruppe. Derefter stilles igen et nyt ja/nej-spørgsmål, og sådan fortsætter algoritmen indtil et givent stop-kriterium nås. Det, som algoritmen leder efter, er spørgsmål, som kan splitte datasættet i grupper, hvor de studerende inden for hver respektive gruppe har et ensartet frafaldsmønster – dvs. at gruppen enten har en meget lav eller meget høj andel frafald.

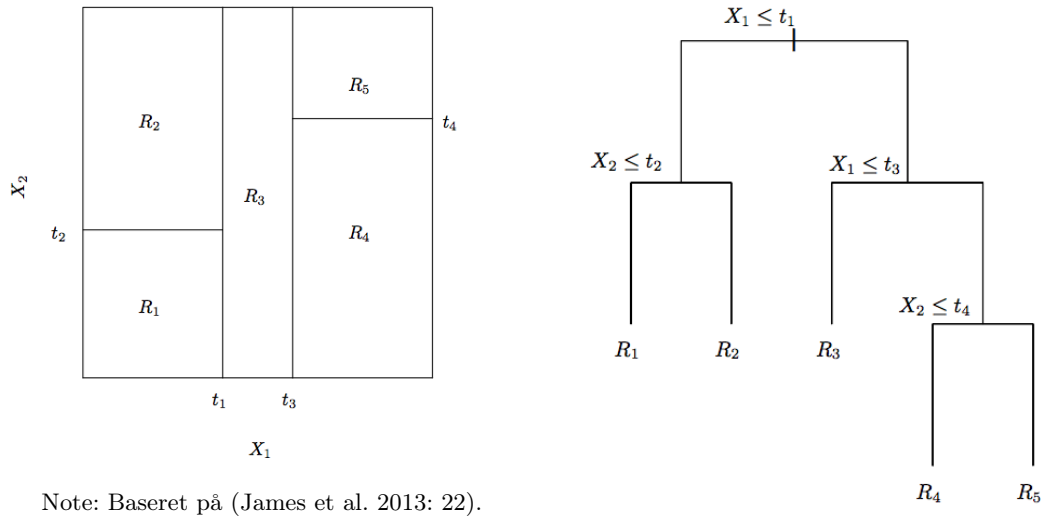
Klassifikationstræer fungerer ved at minimere en loss-funktion igennem segmenteringen af data på baggrund af de variable, som algoritmen har til rådighed (James et al. 2013:

303). Den første segmentering betinger senere segmenteringer, hvilket fortsætter nedad i træet. Det giver en fordel i prædiktionssammenhæng, fordi modellerne derved ikke bygger på en bestemt antagelse om fx linearitet mellem variable og outcome. Anderledes tillades et stort rum for kompleks interaktion mellem variablene, hvilket som nævnt i afsnit 3.1.1 om estimation og prædiktion kan give bedre prædiktioner.

Eksempelvis kunne datasættet i casen Metropol på baggrund af en variabel *køn*, x_1, blive inddelt i to regioner, R_m og R_k. Lad os sige, at frafaldsraten er 28 pct. blandt mændene og 24 pct. blandt kvinderne. I dette tilfælde vil en observation med $x_1 \in R_k$, dvs. en kvinde, forudsiges at have 24 pct. sandsynlighed for at frafalde (James et al. 2013: 303–316).

Efter algoritmen har foretaget et split, fx på variablen køn som i eksemplet her, gennemløbes alle variable og alle deres mulige split igen for hver af de to regioner. Det næste mulige split kunne eksempelvis være, om de studerende i datasættet er over eller under 25 år gamle. I figur 3.7 illustrerer vi logikken.

Figur 3.7: Illustration af et klassifikationstræ

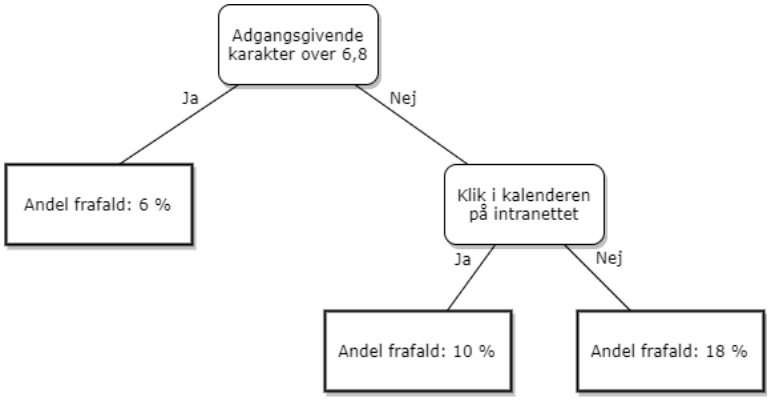


Note: Baseret på (James et al. 2013: 22).

I venstre side af figuren vises det, hvordan det såkaldte *predictor space* inddeles i regionerne $R_1 ... R_5$. Der er i alt foretaget fire split, $t_1 ... t_4$. Blandt andet deler t_1 hele datasættet ind i to dele – alle observationer, hvor hhv. $x_1 \leq t_1$ og $x_1 > t_1$. I regionen hvor $x_1 \leq t_1$ deles datasættet nu ved t_2 på baggrund af x_2, mens regionen hvor $x_1 > t_1$ deles ved t_3 på baggrund af x_1. Det samme klassifikationstræ vises i højre side af figuren, der er en anden måde at illustrere logikken på, og som minder om et flowchart. Illustrationerne tjener samtidig som eksempel på den fordel ved klassifikationstræer, at de – trods deres mulighed for at være meget fleksible og modellere data med mange

interaktioner – er ganske intuitive at fortolke. Eksempelvis tolkes regionen R_4 som alle de observationer, der har en værdi større end t_3 på variablen x_1 og en værdi mindre end eller lig med t_4 på x_2. Det kunne fx være alle studerende over 25 år og med et karaktergennemsnit fra gymnasiet mindre end eller lig med 9, hvis x_1 er alder og x_2 er karaktergennemsnit fra gymnasiet. I figur 3.8 fremgår et eksempel på et klassifikationstræ fra Metropol.

Figur 3.8: Illustration af et klassifikationstræ i casen Metropol



Træet i figur 3.8 viser frafaldsandelen inden for et halvt år for tre regioner af studerende. Risikoen for frafald er højest blandt den gruppe af studerende, som går på en uddannelse, hvor den adgangsgivende karakter ikke er over 6,8, og hvor de studerende ikke har klikket i deres kalender på intranettet i måneden forud for deres studiestart. Træet her er groet meget lidt dybt for eksemplets skyld – i praksis ville vi formentlig få bedre prædiktioner ved at gro træet dybere og splitte data flere gange for at nå frem til regioner, hvor frafaldsandelen er så høj eller lav som muligt.

Klassifikationstræer bliver groet oppefra og ned med en *topdown* og *greedy* tilgang, som kaldes *binary recursive splitting*. Heri ligger, at algoritmen starter med alle observationer i datasættet og splitter det i to, og i hvert af de to resulterende split foretages et nyt split – og så videre. Det gør tilgangen topdown, binær og rekursiv. Tilgangen er greedy, fordi algoritmen vælger det split, som er bedst her og nu – forskelligt fra at være mere "langsigtet" og foretage et split, som vil vise sig at være bedre længere nede i træet (James et al. 2013: 306). Når der ikke er flere split at foretage, siges algoritmen i træanalogien at være færdig med at *gro*, og de endelige regioner kaldes *blade*, ligesom forbindelserne i træet omtales *grene* (James et al. 2013: 305–307).

Mekanikken i klassifikationstræer

For at forstå, hvad der rent faktisk sker, når vi fitter et klassifikationstræ eller en træbaseret model mere generelt, er det nyttigt at kaste et blik på en generaliseret loss-funktion:

$$\ell_{tree} = \arg\min_{\ell} \left[\sum_{x_i \in R_1(j,s)} \ell(y_i, \hat{y}_{R_1}) + \sum_{x_i \in R_2(j,s)} \ell(y_i, \hat{y}_{R_2}) \right] \tag{3.5}$$

Her angiver ℓ_{tree} træmodellens samlede loss som en funktion af de faktiske outcomes, y_i, og de prædikterede outcomes, $\hat{y}_i$ for observationerne i i regionerne R_1 og R_2. Det udtryk, der minimeres, består af summen af to loss-funktioner, henholdsvis én for alle x_i i regionen R_1, og én for alle x_i i regionen R_2. Regionerne R_1 og R_2 er funktioner af j og s, hvor j angiver, hvilken variabel der splittes på, og s angiver, hvor variablen splittes[15]. Algoritmen søger altså at finde værdier for j og s, som minimerer udtrykket. Det vil sige: Hvilken variabel j skal splittes i to ved hvilken værdi s, for at summen af loss'et i de to resulterende regioner, R_1 og R_2, bliver mindst? På denne måde splittes datasættet i to regioner, R_1 og R_2, hvorefter processen gentages, og regionerne R_1 og R_2 hver splittes til to nye regioner. Det fortsætter indtil et stop-kriterium nås. Hermed er træet groet færdigt, og træmodellen er fittet til data (James et al. 2013: 303, 311–314; Friedman et al. 2009: 356).

Selve loss-funktionen, $\ell(y_i, \hat{y}_i)$, kan tage flere forskellige former, der alle er relativt komplekse mål for, hvor "rent" det givne split deler observationerne mellem de to mulige værdier på outcome-variablen[16]. Med renhed forstås, i hvor høj grad observationerne i regionen tilhører den samme klasse, fx frafald eller ikke-frafald (James et al. 2013: 303, 311–314; Friedman et al. 2009: 356).

En fordel ved klassifikationstræer er, at de som udgangspunkt kræver mindre præprocessering af data, bl.a. fordi skalering er unødvendigt, fordi de kan håndtere både numeriske og kategoriske variable, og fordi outliers ikke har den ekstreme indflydelse, som vi kender det fra OLS (Friedman et al. 2009: 352). Det skyldes, at alle variable behandles som dikotome i træmodeller, hvor observationerne på en variabel deles i to regioner, fx over og under en given tærskel. Dermed er der underordnet, om observationerne ligger lige over eller meget over tærsklen – de har den samme indflydelse

[15] Formelt: $R_1(j,s) = \{X|X_J < s\}$ og $R_2(j,s) = \{X|X_J \leq s\}$

[16] Eksempelvis kan *classification error rate* $= 1 - \underset{k}{max}(\hat{p}_{mk})$ minimeres som loss-funktion, hvor $\hat{p}_{mk}$ angiver andelen observationerne i træningssættet, der er i den m'te region og er fra den k'te klasse (James et al. 2013: 311–314). Ofte anvendes dog mere komplekse mål såsom et *gini index* og *cross-entropy*, fordi målet classification error rate ikke er tilstrækkeligt sensitivt (James et al. 2013: 312–313).

på modellen. Dette karakteristika ved træmodeller åbner samtidig for en ny måde at håndtere manglende data på, hvilket vi diskuterer særskilt i afsnit 4.1.4 (Friedman et al. 2009: 352).

Regularisering og tuning af klassifikationstræer

Fordi klassifikationstræer kan modellere data meget fleksibelt, er de tilsvarende sensitive over for overfitting. Et træ kan uden videre gro, indtil det korrekt klassificerer næsten samtlige observationer i et givent træningssæt. Men, som vi har set, vil dette være at overfitte til data, hvilket vil give dårlig performance out-of-sample. Derfor regulariserer vi træer gennem såkaldt *pruning* (beskæring). Ved pruning lader vi først træet *gro vildt*, hvorefter vi beskærer det. Det vil resultere i et *subtree*, der er en mindre kompleks udgave af det oprindelige, vildtvoksende træ. Grunden til, at vi først gror træet med meget lempelige kriterier for derpå at beskære det (i stedet for bare at gro et mindre lempeligt træ fra starten), er, at et umiddelbart nytteløst split kan give anledning til et nyttigt split længere nede i træet (James et al. 2013: 307–311).

Med vores implementering af algoritmen regulerer vi træets kompleksitet gennem tre frie parametre. De første to er stop-kriterier: `maxdepth`, som sætter en grænse for, hvor dybt træet kan gro, og `minbucket`, som sætter en nedre grænse for antal observationer i et blad. Den tredje parameter er α (`alpha`), som sætter en pris på kompleksitet i træets loss-funktion (Therneau et al. 2015). Kompleksitet defineres som antallet af blade (flere blade = mere komplekst træ). I stedet for loss-funktionen i formel 3.5 på forrige side vil vi nu minimere en regulariseret loss-funktion på formen:

$$\ell_{tree} = \arg\min_{\ell} \sum_{m=1}^{M} \sum_{x_i \in R_m} \ell(y_i, \hat{y}_{R_m}) + \alpha M \tag{3.6}$$

Her angiver M antallet af blade i træet, hvor regionen R_m er det m'te blad, og hvor $\hat{y}_{R_m}$ er det prædikterede outcome for observationerne x_i i bladet R_m. Når α sættes til 0, så gror vi det fulde, komplekse og uregulariserede træ. Når α stiger, så stiger prisen i loss-funktionen for at have flere blade i træet og dermed for at have et mere komplekst træ. På den måde minder det om L2-regularisering (James et al. 2013: 309).

De tre frie parametre kan fastsættes med tuning for at finde de parameterværdier, som leverer den bedste performance out-of-sample.

3.3.3 Random Forest

Random Forest (RF) er en tilgang baseret på et ensemble af klassifikationstræer. Forskelligt fra blot at fitte ét klassifikationstræ til vores data, fitter vi med RF en hel skov af træer. Derefter aggregerer vi træerne ved blot at tage det simple gennemsnit af alle deres individuelle prædiktioner (James et al. 2013: 319–321). Det kan vi formelt forstå som:

$$\hat{f}_{RF}(x) = \frac{1}{T}\sum_{t=1}^{T}\hat{f}_t(x) \tag{3.7}$$

Ensemblemetoder som Random Forest resulterer typisk i en forbedret prædiktionsperformance, men har den ulempe, at det kan være svært at fortolke den resulterende model (James et al. 2013: 319). Når vi fitter hundreder eller tusindvis af træer, er det ikke længere muligt at præsentere modellen som det simple flowchart eller predictor space, som vi så i figur 3.7 på side 44 for det simple klassifikationstræ. Det er også svært umiddelbart at sige, hvilke variable som er mest betydningsfulde på tværs af træernes segmenteringer af data. Der findes dog metoder til at beregne denne betydningsfuldhed, variablenes *importance*, hvilket vi gør for vores bedste model i analysen.

Bagging

Random Forest bygger videre på en teknik, som kaldes *bagging*. Det er en forkortelse for *bootstrap aggregation* og er en resampling-teknik, hvor vi for hvert nyt træ trækker et tilfældigt sample fra træningssættets observationer (James et al. 2013: 187-190). RF benytter sig af denne teknik, men går et skridt videre ved også at begrænse hvert træ til kun at have et tilfældigt subsample af datasættets variable til rådighed (James et al. 2013: 319).

Det smarte er, at alle variable dermed får mulighed for at "komme til orde", fordi der kun er en del af dem til rådighed for hvert træ. Ellers kunne der i datasættet have været et fåtal af stærke prædiktorer, som ville forekomme i det øverste split i alle træerne i ensemblet, hvilket ville få træerne til at minde om hinanden, dvs. korrelere. I stedet kan vi nu opnå en række forskelligartede, uafhængige og ukorrellerede træer. Når vi tager gennemsnittet af de ukorrelerede træer, bliver variansen af prædiktionerne mindre, ligesom når vi i andre sammenhænge tager gennemsnittet af en række uafhængige observationer. Samtidig er gennemsnittet af træerne ikke biased i forhold til det oprindelige sample, fordi de er baserede på tilfældige udtræk. Dermed opnår vi

en model, hvor vi – sammenlignet med det almindelige klassifikationstræ – mindsker variansen, uden at det er på bekostning af øget bias, hvilket er den væsentligste årsag til, at RF typisk giver gode prædiktioner out-of-sample (James et al. 2013: 316–321; Murphy 2012: 550–551).

Regularisering og tuning af Random Forest

Random Forest kan principielt reguleriseres med de samme parametre som simple klassifikationstræer, hvilket dog ikke er nødvendigt i praksis (Liaw & Wiener 2002). Lige så vel kunne vi på ensemble-niveau regulere antallet af træer i skoven med parametren `ntree`, men i praksis er det også overflødigt. Årsagen er, at vi med bagging-teknikken sikrer, at de enkelte træer er uafhængige af hinanden. En større skov fører derfor ikke til overfitting, men kræver blot flere computerressourcer (James et al. 2013: 321). Den eneste parameter, vi tuner for RF, er `mtry`, som regulerer antallet af variable, der skal subsamples for hvert træ (Liaw & Wiener 2002).

3.3.4 Gradient Boosted Trees

Gradient Boosted Trees (GBT) er en algoritme, der ligesom Random Forest bygger på et ensemble af klassifikationstræer. Mens RF er baseret på bagging, er GBT-modeller baseret på en teknik kaldet *boosting*. Teknikken går grundlæggende ud på at fitte en række klassifikationstræer *sekventielt*, således at hver iteration, dvs. hvert nyt træ, bygger videre på de forudgående ved at lægger størst vægt på de observationer, som blev klassificeret forkert ved de tidligere iterationer (Friedman et al. 2009: 337–342, 353–361; Kleinberg et al. 2017: 15). Det uddyber vi i afsnittet her.

Boosting

Logikken i boosting er ligesom ved RF at kombinere et ensemble af simple klassifikationstræer og foretage en afstemning blandt disse. Ved boosting er afstemningen dog vægtet sådan, at de træer, der leverer de mest præcise forudsigelser, tillægges en større indflydelse (Friedman et al. 2009: 337–340). Den funktion, vi minimerer, tager samme form, som vi kender:

$$\ell_{GBT} = \arg\min_{\ell} \left[\sum_{i=1}^{n} \ell(y_i, \hat{y}_i^{(b)}) + \sum_{t=1}^{T} \omega(f_t) \right] \tag{3.8}$$

Det første led er et loss som følge af at prædiktere $\hat{y}_i$, givet at det faktiske outcome for observationen i er y_i. Det andet led regulariserer udtrykket for hvert træ f_t i ensemblet. ω er en kompleksitets-funktion, der eksempelvis kan indeholde antallet af blade i et givent træ f_t, sådan som vi så i afsnit 3.3.2 om regularisering af det simple klassifikationstræ. Det særlige ved loss-funktionen ovenfor er, at der findes et prædikteret outcome $\hat{y}_i$ for hver iteration, b, i booosting-processen. Når vi træner modellen ved boosting foregår det således sekventielt (Friedman et al. 2009: 337–340; Chen 2014).

Det betyder, at hvert træ hele tiden bygger videre på træerne fra de tidligere iterationer. Det træ, som i GBT bliver tilføjet i iterationen b er valgt sådan, at det minimerer loss'et fra de foregående træer mest muligt. Det sker ved at fitte til de foregående træers residualer snarere end selve outcomet y_i, som vi gør med et almindeligt klassifikationstræ. I takt med at flere træer tilføjes som led i boosting-processen, bevæger vi os langsomt nedad mod et minimum for det laveste loss, som det er muligt at opnå med modellen anvendt på det givne datasæt. Denne tilgang kaldes for *gradient descent* og er ophavet til ordet *gradient* i Gradient Boosted Trees. Den sekventielle fitting-proces er den centrale forskel på GBT og RF. Ved RF er træerne i ensemblet uafhængige; i GBT bygger de videre på hinanden (Chen 2014; Friedman 2001; James et al. 2013: 321–323).

Regularisering og tuning af Gradient Boosted Trees

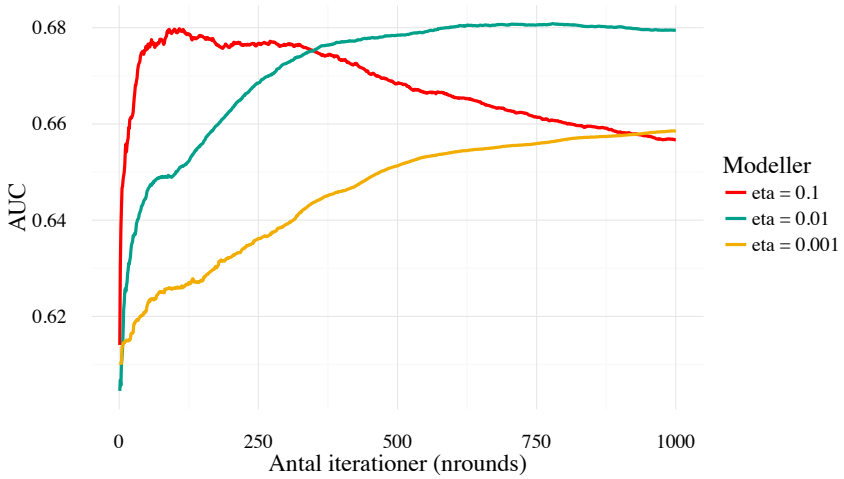
I vores analyse af casen Metropol anvender vi den implementering af Gradient Boosted Trees, som hedder `xgboost` (DMLC 2016a). Gradient Boosted Trees er den af vores algoritmer, som har flest frie parametre. De kan overordnet inddeles i tre grupper.

Den første gruppe er på træniveau. Her fastsætter vi de samme tre parametre som for individuelle klassifikationstræer: `max_depth` begrænser træernes dybde, `min_child_weight` sætter en nedre grænse for antal observationer i et blad, og kompleksitetsparametren `gamma` sætter en straf på antallet af blade. Dertil anvender `xgboost` også `lambda`, der svarer til L2-regularisering af loss-funktionen, som vi beskrev det i afsnit 3.2.6 på side 38 (Chen et al. 2016; James et al. 2013: 321–323).

Den anden gruppe vedrører sampling. Fra RF-tilgangen lånes `colsample_bytree`, der regulerer hvor mange variable, som skal være tilgængelige for hvert træ, mens `subsample` regulerer andelen af observationer, som skal samples. Intuitionen bag begge parametre er, at de tilføjer tilfældig variation til modeltræningen, som gør den mere robust over for støj (Chen et al. 2016; James et al. 2013: 321–323).

Den sidste gruppe af parametre vedrører selve boosting-processen. Det er henholdvis antallet af boosting-iterationer `nrounds` (dvs. antallet af træer) og den såkaldte læringsrate η (`eta`). Forskelligt fra Random Forest er der i boosting en risiko for overfitting, hvis man foretager for mange boosting-iterationer, fordi træerne ikke er uafhængige af hinanden. Hvor hurtigt dette sker er dog afhængigt af læringsraten, som styrer, hvor stor en indflydelse den nyeste iteration skal have på det foreløbige ensemble af træer. Antallet af `nrounds` og η er inverse, så det for en lavere lærings-rate er optimalt med et højere antal boosting-iterationer og omvendt (Friedman 2001: 1203–1206; James et al. 2013: 321–323). Det illustrerer vi i figur 3.9, der viser sammenhængen mellem antallet af boosting-iterationer og out-of-sample performance (målt ved AUC) ved tre forskellige læringsrater, η.

Figur 3.9: Performance som funktion af antal iterationer ved tre forskellige læringsrater



Det ses af figuren, at modellen med den højeste læringsrate, $\eta = 0{,}1$, relativt hurtigt når sit toppunkt for performance. Det skyldes, at en model med en højere læringsrate fitter hurtigere til data, fordi den i mindre grad regulariseres, hvormed nye iterationer tillades en større indflydelse på ensemblet af træer. Til gengæld begynder denne model også at overfitte hurtigt relativt til de øvrige modeller. Eksempelvis når modellen med den laveste læringsrate, $\eta = 0{,}001$, slet ikke sit toppunkt for performance i figuren (den underfitter). Ud over at vise sammenhængen mellem læringsraten og antallet af boosting-iterationer tjener figuren til at illustrere en mere generel pointe: I forskellige kombinationer kan de frie parametre substituere hinanden, og når én parameter ændres, har det betydning for, hvordan de øvrige parametre fastsættes optimalt. Når vi tuner GBT-modellens parametre, forsøger vi at finde frem til den empirisk bedste kombination af parametre, herunder læringsraten og antal boosting-iterationer. I analysens afsnit 5.2.3 beskriver vi, hvordan det foregår i praksis.

Kapitel 4

Data

I dette kapitel beskriver vi vores datasæt og diskuterer en række diskretionære valg og afgrænsninger. I afsnit 4.1 diskuterer vi først, hvilken forskel det gør for datasættet, at det skal bruges til prædiktion frem for estimation. Herefter beskriver vi vores datakilder, bearbejdningen af variable og håndteringen af manglende data. I afsnit 4.2 diskuterer vi, hvordan vi sætter forskellige tidsmæssige afgrænsninger for frafaldsmodellen, og hvilken betydning det har.

4.1 Konstruktion af datasæt

Vores datasæt tæller 23.107 studerende med information om 103 variable. Vi betragter datasættet som populationsdata, da datasættet ikke er et sample. Vores population af studerende omfatter alle studerende, som er begyndt på Metropol efter 1. august 2009 – eksklusive bestemte grupper som er uinteressante for frafaldsundersøgelsen, fx udvekslingsstuderende. I gennemsnit frafalder 27 pct. af de studerende.

4.1.1 Forskelle mellem data til estimation og prædiktion

Ofte er bearbejdning af data det mest tidskrævende arbejde ved maskinlæring, når man er interesseret i at konstruere et datasæt, som er velegnet til prædiktion (Foster et al. 2016: 150–151). Vi vil her fremhæve tre aspekter af, hvordan datasæt til prædiktion adskiller sig fra datasæt til estimation.

For det første kan datasættet indeholde et væld af forskelligartede data, som vi ikke kan give nogen kausal fortolkning, og som vi ikke på forhånd har teoretiske forventninger til. Det står i modsætning til estimationsstudier, hvor det er nødvendigt med en klar teoretisk forventning om sammenhængene i data. Hvis vi skal tro på en models effektestimater, må vi teoretisk kunne godtgøre, at alle relevante variable indgår, og at der ikke er spuriøse sammenhænge i data. Når formålet er prædiktion, bekymrer vi os derimod ikke om kausalitet og derfor heller ikke om spuriøsitet. Et godt eksempel fra vores eget datasæt til at illustrere denne forskel er data om de studerendes brug af intranettet. Vi forventer nemlig ikke, at de studerendes adfærd på intranettet har en selvstændig kausal effekt på frafaldsrisikoen. Den kan blot være et udtryk for bagvedliggende forhold såsom gode studievaner, engagement i studiet etc. I prædiktionssammenhæng bekymrer vi os ikke om denne spuriøsitet, men blot om sådanne data har prædiktiv værdi.

For det andet lægger man ved maskinlæring større vægt på at eksperimentere med datasættet. Et vigtigt trin i maskinlæring er den såkaldte *feature engineering*, hvor man gennemgår data for at eksperimentere med og generere nye variable (Foster et al. 2016: 150–151, 180–181). Som en del af denne proces afprøver man gerne mange forskellige specifikationer og kombinationer af variable for at hive mest mulig prædiktiv værdi ud af dem. Dette arbejde ville være en kilde til panderynken, hvis formålet med studiet var kausal inferens. I estimationsstudier bør udvælgelsen af variable være styret af teori og klare hypoteser, hvis estimaterne skal være troværdige. For meget efterfølgende datatransformation ville være dårlig latin og vække mistanke om, at man forsøgte at fifle sig frem til signifikante p-værdier – en uskik betegnet *p-hacking* eller *researcher degrees of freedom* (Gelman & Loken 2013). Dette fifleri er imidlertid ikke blot acceptabelt, men helt fundamentalt i maskinlæring, hvor p-værdierne ikke er afgørende, fordi modellen alene bliver vurderet på sin prædiktive performance.

For det tredje udmærker datasæt til maskinlæring sig typisk ved at have flere variable, end vi traditionelt arbejder med i samfundsvidenskaben. I et datasæt til estimation er det nødvendigt at begrænse antallet af beslægtede variable, der korrelerer, for at undgå problemer med multikollinearitet. Når flere variable er kollineære, vil det give misvisende store standardfejl for parameterestimaterne (Wooldridge 2009: 84–89; James et al. 2013: 99–102). Når vi har prædiktion for øje, er vi dog ikke interesserede i parameterestimaterne, og derfor er multikollinearitet ikke en bekymring her. Derfor er der ikke noget til hinder for at inkludere mange beslægtede variable og forskellige specifikationer af den samme variabel.

4.1.2 Datakilder

Alle analysens variable er genereret på baggrund af data, som Metropol selv råder over, og som også fremadrettet er tilgængelige for Metropol. Det betyder, at vi fx ikke medtager socioøkonomiske variable om de studerende, som af hensyn til persondatabeskyttelse kun kan indhentes gennem bekostelige engangsudtræk hos Danmarks Statistik. Ved at tage højde for sådanne begrænsninger på dataudveksling får vi et realistisk grundlag for at diskutere maskinlærings potentiale i den offentlige forvaltning mere generelt, frem for at basere modellen på en unik dataadgang.

En del af Metropols data stammer fra tre eksterne kilder. Fra CPR-registret kommer baggrundsoplysninger om fx de studerendes navn og bopæl. Den Koordinerede Tilmelding (KOT), som er det statslige organ for optagelse på videregående uddannelser, leverer data om de studerendes ungdomsuddannelser. Endelig bidrager Uddannelses- og Forskningsministeriet (UFM) med data om de studerendes prioritering af uddannelser. Disse eksterne data samkøres dagligt i et nationalt it-system, som Metropols database synkroniserer med.

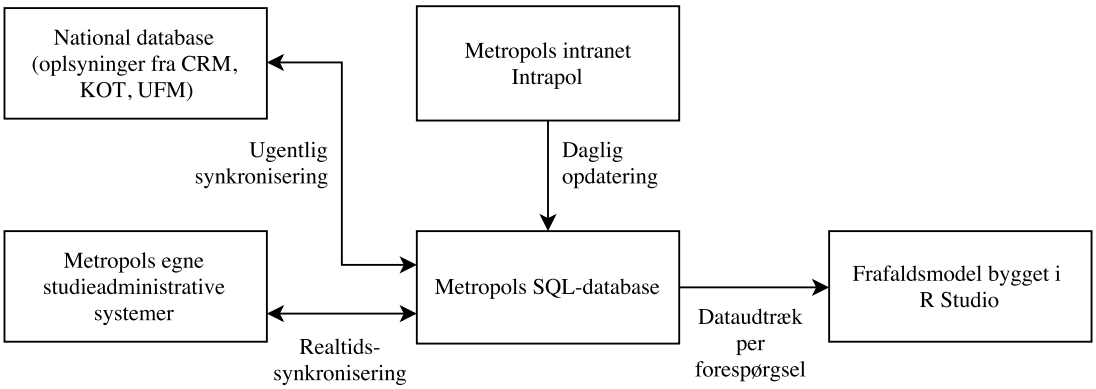
Langt det meste data bliver dog genereret på Metropol selv. En stor del bliver manuelt indtastet i studieadministrative systemer – det gælder fx de studerendes karakterer, holdplaceringer, orlov og udlandsophold. Derudover bliver en mængde data automatisk genereret om de studerendes brug af intranettet ved såkaldt sessionslogning. De forskellige datakilder bliver alle samlet i Metropols SQL-database, som det fremgår af figur 4.1. I statistikprogrammet `R Studio` sætter vi en direkte forbindelse op, så vi kan indhente data gennem en række forespørgsler til databasen. Modellerne har dermed hele tiden de nyeste data til rådighed.

4.1.3 Feature engineering

Feature engineering, dvs. udvælgelse og bearbejdning af variable, er en vigtig del af praktisk maskinlæring. Indtil nu har vi fremhævet, at maskinlæring som tilgang frigør os fra teori, men denne fremstilling er en smule forsimplet. Selvom vi ikke eksplicit begrænser frafaldsmodellen med teoretiske forventninger, spiller teoretisk, domænespecifik viden en implicit rolle for, hvordan vi indsamler, undersøger og genererer data (Foster et al. 2016: 150-151). Det er en pointe, vi senere reflekterer kritisk over i diskussionen.

Vi har konkret indsamlet, renset og koblet data i samarbejde med fagpersonalet på Metropol, og vi har endvidere ladet os inspirere af litteraturen om uddannelsesfrafald

Figur 4.1: Oversigt over dataflow



og de tidligere maskinlæringsforsøg på området. På den baggrund har vi fx fokuseret på at modellere de variable, som ifølge teorien er bestemmende for niveauet af de studerendes commitment ved studiestart. Det drejer sig navnlig om variable relateret til de studerendes forudgående skolegang, hvorfor vi fx har ladet karaktergennemsnit fra gymnasiet indgå i flere forskellige variable: den studerendes absolutte gennemsnit, den relative forskel til uddannelsens adgangsgivende snit og den relative forskel til studieårgangens gennemsnit. Sådanne variable vil selvfølgelig have en høj grad af kollinearitet, men da det ikke er en bekymring i prædiktive studier, lader vi dem alle indgå, da vi ikke kan vide, hvilken specifikation der er bedst.

Tabel 4.1 på den følgende side indeholder en oversigt over de variable, som indgår i det fulde datasæt. I første kolonne er variablene samlet i grupper med visse fællestræk. Enkelte beslægtede variable er slået sammen for overskuelighedens skyld; i så fald markerer sidste kolonne, hvor mange variable en række dækker over.

Af anden kolonne fremgår det, om variablene er kontinuert skalerede, dikotome eller kategoriske med flere end to kategorier. Kategoriske variable har vi omkodet, så hver kategori er repræsenteret ved en dummy-variabel[1]. Af tredje kolonne fremgår enten variablenes navne, i de tilfælde hvor flere variable er slået sammen i én række, eller kategoriernes navne for de kategoriske variable.

Mange data er allerede tilgængelige ved studiestart. Det gælder fx baggrundsdata, uddannelsesspecifikke data og data om de studerendes adgangsgrundlag. Derudover bliver noget data genereret løbende hos Metropol. Det gælder fx data om de stude-

[1]Der er i tabellens tredje kolonne sat parenteser omkring reference-kategorien for dummien.

Tabel 4.1: Oversigt over variable

Variabel	Type	Kategorier	# variable
Baggrundsdata			
Alder	Kontinuert		1
Køn	Dikotom	M, K	1
Optagelsesår	Kontinuert		1
Indskrevet via KOT	Dikotom	Ja, nej	1
Vinter- eller sommeroptag	Dikotom	Vinter, sommer	1
Tidligere frafaldet Metropol	Dikotom	Ja, nej	1
Uddannelsesspecifikke data			
Uddannelse	Kategorisk	Forskellige uddannelser	13
Campus	Kategorisk	Forskellige uddannelsesmatrikler	7
Praktik/klinik	Kontinuert	Andel, længde til første, antal uger før første	3
Karakteristika ved første studieår	Kontinuert	Andel teori, antal undervisningstimer, vejledning	3
Antal studerende på uddannelsen	Kontinuert		1
Størrelse af uddannelse	Dikotom	Stor, lille	1
Dimensionering af uddannelse	Dikotom	Ja, nej	1
Adgangskrav til uddannelse	Dikotom	Ja, nej	1
Adgangskrav om matematik	Dikotom	Ja, nej	1
Adgangsgivende karaktergennemsnit	Dik./Kont.	Ja, nej	2
Uddannelsesvarighed	Kontinuert		1
Uddannelsens alder	Dikotom	Gammel, ny	1
Data om adgangsgrundlag			
Ungdomsuddannelse	Kategorisk	Gymnasial, anden, (ingen)	2
Karaktergns. fra ungdomsuddannelsen [a]	Kat./Kont.	2–, 3, 4, 5, 6, 7, 8, 9, 10, 11+, (NA)	12
Antal år efter gymnasiet [a]	Kat./Kont.	0, 1, 2, 3+, (NA)	6
Forskel ml. karaktergns. og adgangskrav [a]	Kat./Kont.	<–1, <0, >0, >1, (NA)	6
Karaktergns. ift. årgangens gns. [a]	Kat./Kont.	Meget under, omtrent samme, meget over, (NA)	5
Optaget på prioritet [a]	Kat./Kont.	Førsteprioritet, andenprioritet, tredje eller over, (NA)	5
Antal prioriteter	Kontinuert		1
Hold-, orlov- og udlandsopholdsdata			
Gennemsnitlig holdstørrelse	Kontinuert		1
Gennemsnitsalder på hold	Kontinuert		1
Gennemsnitlig kønsfordeling på hold	Kontinuert		1
Haft udlandsophold	Dikotom	Ja, nej	1
Længden af udlandsophold	Kontinuert		1
Haft orlov	Dikotom	Ja, nej	1
Orlovslængde	Kontinuert		1
Orlovstype	Kategorisk	Barsel, frivillig, sygdom, tvungen studiefri, (ingen)	4
Karakterdata			
Haft eksamen	Dikotom	Ja, nej	1
Beståelsesgrad	Kontinuert		1
Karaktergennemsnit	Kontinuert		1
Antal eksamensforsøg	Kategorisk	Brugt to forsøg, brugt 3 forsøg, (ikke brugt flere forsøg)	2
Længde til første bestået eksamen	Kontinuert		1
Udvikling i karakterer	Kontinuert		1
Logdata fra intranettet [b]			
Antal logons	Kontinuert		1
Antal dage	Kontinuert	I logonperiode, fra første/sidste logon til studiestart	3
Samlet sessionstid	Kontinuert		1
Gennemsnitlig sessionstid	Kontinuert		1
Har klikket i bestemt sektion	Dikotom	Fildeling, kalender, lektionsplan	3
Antal klik i bestemt sektion	Kontinuert	Fildeling, kalender, lektionsplan	3

[a] Disse variable indgår kun i deres kategoriske version i den logistiske model. Se forklaring i afsnit 4.1.4.
[b] Logdata indgår kun i en selvstændig analyse. Se forklaring i afsnit 4.1.4.

rendes karakterer, hold, orlov og udvekslingsophold, som udvikler sig for den enkelte studerende gennem studielivet.

Den sidste gruppe af variable i tabellen, logdata, stammer fra de studerende brug af intranettet på Metropol. Fra denne datakilde har vi konstrueret variable såsom *gennemsnitlig sessionstid* og *antal klik i lektionsplanen*. Denne gruppe af variable er også tilgængelige fra studiestart, idet de studerende oprettes på intranettet, samtidig med at de får deres studieplads, og dermed kan anvende intranettet ca. en måned før undervisningens begyndelse.

Vi giver logdata en særlig opmærksomhed i analysen. Det skyldes, at det adskiller sig fra fx registerdata og surveydata, som vi er mere hjemmevant med at undersøge i samfundsvidenskaben. I kraft af omfanget, karakteren og hastigheden, hvormed logdata genereres, kan det siges at have big data-karakteristika[2]. For det første genereres logdata i realtid. For det andet er der tale om observationelt data bestående af de spor, som de studerende mere eller mindre ubevidst efterlader ved almindelig online-adfærd. For det tredje er data omfattende. På det tidspunkt vi laver udtrækket, er der ca. 160 mio. rækker med sideregistreringer, der kan aggregeres til 12 mio. unikke sessioner. Omfanget betyder, at computerkræfterne viser sig at være mest efficient brugt ved at lave al feature engineering på logdata direkte i SQL-databasen.

Set i lyset af hypen omkring big data er det interessant at undersøge den prædiktive værdi i denne del af datasættet. Den egentlige udfordring viser sig dog ikke at være omfanget af data – på en måde tværtimod. Metropols intranet har nemlig kun eksisteret siden sidste halvdel af 2014. Det betyder, at der kun er 7.563 studerende, som har logdata helt fra deres studiestart. Hvordan vi håndterer disse og andre manglende data, er emnet for næste afsnit.

4.1.4 Håndtering af manglende data

Manglende data er en velkendt statistisk hovedpine i samfundsvidenskaben, da observationer ikke kan indgå i en model, hvis de ikke har en registreret værdi for alle variable i modellen. Der er flere bud på, hvordan denne udfordring løses bedst (Little & Rubin 2014). I dette afsnit diskuterer vi, hvordan vi håndterer manglende data i casen Metropol. Herunder kommer vi ind på, hvordan de træbaserede algoritmer åbner for en ny måde at håndtere problemet på, som ikke er mulig for regressionsmodeller.

[2]Her har vi en ofte anvendt definition af De Mauro et al. (2016) i tankerne, som karakteriserer big data ved tre kendetegnende V'er: volume, velocity og variety.

Det kræver en vurdering af de enkelte informationshuller at finde den mest velegnede strategi til at håndtere dem. Først og fremmest afhænger strategien af, om den manglende data er tilfældigt fordelt blandt observationerne. Er der et mønster i, hvilke studerende som har manglende data, eller er det rent tilfældigt, hvilke data som mangler? Den vigtigste distinktion er, om den manglende data kan anskues som:

- MCAR (Missing Completely At Random)
- MAR (Missing At Random)
- MNAR (Missing Not At Random)

Den nemmeste situation er, hvis vi vurderer informationshuller til at være *MCAR*. Det betyder, at sandsynligheden for, at en værdi mangler, hverken afhænger af værdien selv eller af andre observerede variable (Little & Rubin 2014: 12). Mangler er med andre ord helt tilfældige. En mulig – og hyppigt anvendt – strategi til at håndtere denne type mangler er *listwise deletion*, hvormed vi simpelthen fjerner observationer fra datasættet, som har manglende værdier (Little & Rubin 2014: 41–43). Der er dog tre problemer ved denne tilgang. For det første risikerer vi at miste mange observationer og dermed få lavere statistisk *power* (James et al. 2013: 101). For det andet er det praktisk uhensigtsmæssigt i vores case, at nye studerende ikke kan indgå i modellen og få prædikteret en frafaldsrisiko, blot fordi de har ukomplette data. For det tredje er det svært at påvise, at antagelsen om MCAR er overholdt. Ofte antager man i stedet, at data er *MAR*, hvilket vil sige, at sandsynligheden for manglende data afhænger af observerede variable (men ikke af de manglende værdier selv). Når data er MAR, er der et mønster i, hvilke studerende som mangler information på en variabel, og derfor vil det introducere bias i modellen at udelade dem med listwise deletion (King et al. 2001: 51–52).

For MAR-data er *imputation* en bedre strategi, hvormed manglende værdier erstattes på baggrund af korrelationer mellem variable i datasættet (King et al. 2001). Det er en strategi, som løbende er blevet mere raffineret. Ved *multiple imputation* imputerer man flere gange og tager gennemsnittet af datasættenes prædiktioner for at tage højde for de imputerede værdiers usikkerhed (King et al. 2001: 53). Fordelen ved imputation i forhold til listwise deletion er, at man bevarer alle observationerne og undgår muligt bias. Det er imidlertid også en mere omstændelig procedure. Vi har tre informationshuller i vores data, hvor det kunne være relevant med multiple imputation, men på baggrund af en konkret vurdering af de manglende data har vi ikke anvendt denne fremgangsmåde. Vi gennemgår her de tre grupper af manglende data efter tur og forklarer hvorfor.

Det første informationshul består af tilsyneladende tilfældige mangler i de manuelt indtastede data. Der er samlet set tale om tre observationer, og derfor tillader vi os at lade disse udgå ved simpel *listwise deletion* frem for at foretage mere omstændelig imputation.

Det andet informationshul gælder logdata, som jf. forrige afsnit kun er tilgængeligt for en afgrænset periode, fordi intranettet kun har eksisteret siden 2014. Der er tale om MAR-data, fordi manglerne alene afhænger af observerede data: nemlig de studerendes optagelsesår. Til gengæld kan man diskutere, om det er logisk meningsfuldt at imputere værdier, som ikke eksisterer – det strider med en underliggende antagelse om, at manglende data dækker over en sand værdi (Little & Rubin 2014: 8, 10). Rent teknisk er det dog ikke noget problem, fordi imputationen her blot er et matematisk værktøj til at beholde data i modellen (Schafer & Graham 2002: 155). I princippet kunne vi derfor imputere hypotetiske logdata bagud i tid, men vi vurderer, at det ikke er umagen værd. Som vi skal se i et selvstændigt afsnit af analysen (afsnit 5.3.2), viser det sig nemlig, at logdata-variablene praktisk taget ikke bidrager med prædiktiv værdi for den del af datasættet, hvor de er til rådighed. Dermed giver det ikke mening at imputere værdier for resten af observationerne – også set i lyset af, at observationer med manglende værdier udgør langt hovedparten af datasættet. Vi vælger i stedet at udelade de berørte variable fra hovedanalysen, efter vi har vist, at de ikke bidrager med prædiktiv værdi. I denne sammenhæng er det vigtigt at huske på, at vores undersøgelses formål ikke er kausal estimation. Fremgangsmåden kunne i så fald have været en kilde til omitted variable bias. Idet vores formål er prædiktion, er vi dog ikke på samme måde bekymrede for bias i modellen – her er vi kun interesserede i prædiktiv værdi.

Det tredje og sidste informationshul i vores data forekommer i de variable, der er relateret til adgangsgrundlaget. En del studerende er fx optaget på kvote 2 uden at have en adgangsgivende ungdomsuddannelse. Derfor er det helt naturligt, at de ikke har et karaktergennemsnit, og derfor kan det være logisk misvisende at imputere ikke-eksisterende værdier. Lad os fx forestille os en situation, hvor man skulle forklare en studerende, at vedkommendes prædikterede frafaldsrisiko var baseret på en hypotetisk gymnasiekarakter. Det ville næppe være forvaltningsmæssigt acceptabelt og demonstrerer, hvordan forskningsidealer kan kollidere med realiteterne i en konkret anvendt forvaltningspraksis.

Samtidig er der en anden og vigtigere grund til, at vi ikke imputerer gymnasiedata. Måske disse data slet ikke er MAR, hvilket er den antagelse om tilfældighed, som multiple imputation bygger på (King et al. 2001). Nogle studerende søger ganske vist ind på kvote 2, fordi de ikke har gået på gymnasiet. Andre søger dog ind på kvote 2, selvom de har gået på gymnasiet, fordi de fx ikke har fået høje nok karakterer til

at bruge det som adgangsgrundlag. Dermed er data ikke MAR, men *MNAR*, hvilket vil sige, at de manglende data afhænger af deres egen værdi (Little & Rubin 2014: 12). Karakterernes eget niveau har med andre ord indflydelse på, hvor sandsynligt det er, at de mangler i datasættet. Vi kan derfor ikke imputere, idet karakterer og øvrige gymnasierelaterede variable kan være systematisk forskellige fra resten af de studerendes.

En bedre strategi kan her være at håndtere den manglende data som en selvstændig kategori. At data mangler er i sig selv informativt. Den klassiske måde at håndtere denne situation på er ved at lave en dummy-variabel for manglende data. Det fungerer fint for kategoriske variable, men betyder at man er nødt til at omkode kontinuerte variable til et mindre antal dummy-variable. For regressionsmodeller er det en udmærket løsning, fordi alle dummy-variable indgår "samtidig", når modellen fittes. Det er dog mindre hensigtsmæssigt for træbaserede modeller, der splitter på én variabel ad gangen (Twala et al. 2008). Her indeholder den enkelte dummy-variabel kun information om en mindre subgruppe af populationen og rummer således mindre potentiale til at forbedre modellens performance end en kontinuert skaleret variabel, der dækker hele populationen. Endvidere har dummy-variable kun ét muligt split, mens en kontinuert variabel kan splittes langs hele dens skala. Derfor er ønskværdigt at bevare kontinuerte variable i træmodeller frem for at opsplitte dem i dummier (Twala et al. 2008).

Vi implementerer derfor en anden strategi, der også håndterer manglende data som en selvstændig kategori, men uden at omkode de kontinuerte variable. Strategien består ganske simpelt i at kode manglende data til en ekstrem værdi og bliver derfor nogle steder benævnt MIA, *Missingness Incorporated in Attributes* (Twala et al. 2008; Kapelner & Bleich 2015). Vi genkalder os, at klassifikationstræer ved hvert split sorterer observationerne i to grupper ud fra, om deres værdi er højere eller lavere end det valgte brudpunkt. Med implementeringen af MIA får klassifikationstræer mulighed for at sende manglende data videre med enten den ene eller den anden gruppe i et givet split. Der er også mulighed for at lægge brudpunktet sådan, at de manglende data bliver splittet fra resten og på den måde opfattes som selvstændig kategori.

Denne fremgangsmåde har i flere eksperimenter vist sig at give bedre resultater end konkurrerende strategier for træmodeller (Twala et al. 2008; Ding & Simonoff 2010). MIA er dog forbeholdt klassifikationstræer, idet metoden ville resultere i helt misvisende koefficienter i den logistiske regressionmodel. I denne model indgår derfor kun dummy-versionerne af disse variable, sådan som det fremgår af tabel 4.1.

4.2 Optimal periode til modeltræning og prædiktion

Vores studiepopulation tæller i udgangspunktet de nævnte 23.107 studerende. Det er imidlertid ikke alle studerende, som indgår i alle specifikationer af frafaldsmodellen. Vores datasæt er bygget dynamisk, sådan at det ændrer sig på baggrund af to forhold: dels tidspunktet for forudsigelsen, og dels tidshorisonten, som vi er interesserede i at forudsige frafald inden for. Hvordan vi specificerer frafaldsmodellen og dermed populationen giver derfor anledning til en række diskretionære valg, som vi kort vil diskutere i de næste afsnit.

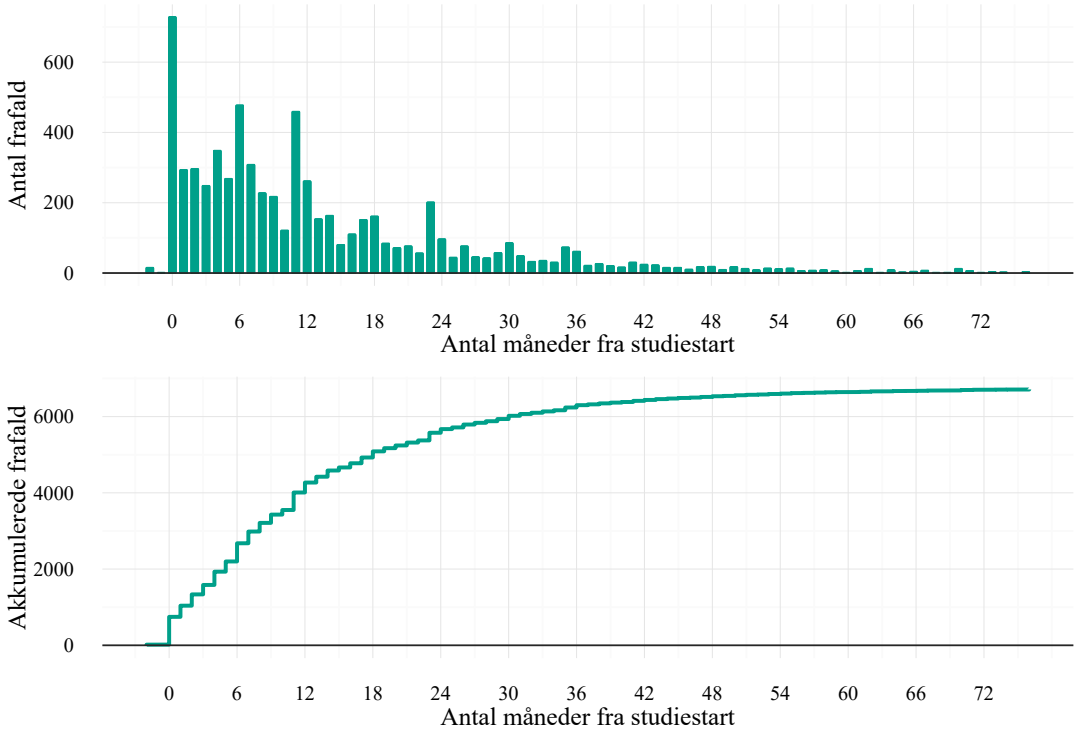
4.2.1 Forudsigelsestidspunkt

Hvornår er det bedste tidspunkt at forudsige frafald på Metropol? Fordi databasen opdaterer i realtid, kunne Metropol i princippet vælge at forudsige frafald hver eneste dag, sådan at studerende hele tiden fik opdateret deres forudsagte sandsynlighed for at frafalde studierne. I praksis bliver de fleste variable dog kun opdateret med en noget langsommere kadence, betinget af studieårets begivenheder. Nye karakterer og holdplaceringer følger fx den halvårlige semesterstruktur.

Derfor kan der fastsættes forskellige forudsigelsestidspunkter ud fra, hvornår nye variable bliver tilgængelige, hhv. ved studiestart, efter første semester og efter første studieår. Hvert halve år efter semesterstart kan det således give mening at prædiktere de studerendes frafaldsrisiko på ny. Det optimale tidspunkt er dog ikke givet alene ud fra datatilgængelighed. Jo længere man venter med at forudsige frafald, jo flere data er der til rådighed om den enkelte studerende – men jo flere studerende vil allerede være droppet ud. Det er altså en afvejning mellem datamængde og forpassede forudsigelser.

I denne opgave fokuserer vi på at forudsige frafald ved studiestart. Der er ganske vist færre tilgængelige variable på dette tidspunkt, men til gengæld finder en tredjedel af alle frafald sted inden for første semester, som det ses af figur 4.2 på næste side. Efter et år har over halvdelen af alle frafald fundet sted. Det er med andre ord ved studiestart, at potentialet for Metropol til at iværksætte tiltag mod frafald er størst. I analysen medtager vi også en model, der forudsiger frafald efter et år, for også at vurdere modellens potentiale senere i studieforløbet, når den har flere informationer til rådighed.

Figur 4.2: Frafald over tid



4.2.2 Modellens tidshorisont

Den anden parameter, som afgrænser datasættet, er den periode, som vi forudsiger frafald inden for. Skal vi forudsige frafald inden for første måned, første semester eller første år? Jo længere en tidsperiode vi træner modellen på, jo flere frafald kan indgå i modellen.

Imidlertid kan der også være grunde til at sætte perioden kortere. Måske forventer vi, at årsagerne til frafald sent på en uddannelse er væsensforskellige fra årsagerne til det tidlige frafald. Lad os for eksempel antage, at der er en regelmæssighed i baggrundsdata hos de studerende, som falder fra på første semester, mens de studerende, som falder fra efter et år, bedre kan detekteres ud fra udviklingen i deres karakterer, som er ukendt ved studiestart. Denne sidste gruppe af frafald ville i så fald kun bidrage til at sløre billedet, hvis vi bad modellen om at forudsige deres frafald ved studiestart. Hvorvidt det er tilfældet, afgør vi empirisk ved at træne modellen på forskellige perioder.

Derudover indebærer en lang tidshorisont også, at det nyeste data ikke kan indgå i modeltræningen. Hvis vi eksempelvis træner en model på frafald inden for det første studieår, frasorteres de studerende i datasættet, som endnu ikke har studeret et helt år. Vi træner derfor også vores modeller med forskellige tidshorisonter.

Kapitel 5

Analyse

I dette kapitel gennemgår vi resultaterne af vores prædiktionsmodeller. I afsnit 5.1 sammenligner vi resultaterne af de fire algoritmer: logistisk regression, et simpelt klassifikationstræ, Random Forest (RF) og Gradient Boosted Trees (GBT). I afsnit 5.2 går vi i dybden med GBT-modellen, der viser sig at være den bedste model. Vi viser, hvordan tuningen har betydning for performance, og undersøger, hvilke variable som har den største prædiktive værdi. I afsnit 5.3 undersøger vi andre modelspecifikationer. Her tester vi først betydningen af sample-størrelse, samt hvor stor en prædiktiv værdi, der er at hente i logdata fra intranettet. Derefter undersøger vi, hvordan modellen klarer sig efter et helt år, når den er blevet beriget med flere data om de studerendes akademiske præstationer og bruges til at forudsige frafald for den resterende studieperiode.

I afsnit 5.4 undersøger vi, hvordan modellen kan implementeres i praksis. Her opstiller vi et framework for, hvordan en beslutningstager kan fastsætte en tærskelværdi og tage frafaldsmodellen i anvendelse til målretning af tiltag. Her viser vi, hvordan anvendelsen af en prædiktionsmodel er aldeles afhængig af konteksten, som den implementeres i.

Alle vores modeller er udviklet i open source-programsproget `R` (R Core Team 2016). Koden er vedlagt i bilag D. Vi har brugt de indbyggede funktioner til at fitte den logistiske model, mens vi har benyttet os af ekstra pakker til at træne de forskellige træbaserede algoritmer. Til at gro simple klassifikationstræer har vi anvendt pakken `rpart` (Therneau et al. 2015). Random Forest er implementeret i pakken med det selvforklarende navn `randomForest` (Liaw & Wiener 2002). Endvidere har vi brugt `xgboost`, som er en populær implementering af Gradient Boosted Trees (Chen et al. 2016).

Alle modellerne er trænet på et træningsæt, der består af et tilfældigt sample på 80 pct. af det fulde datasæt. Modellerne er blevet tunet med 5-fold krydsvalidering som beskrevet i afsnit 3.2.7. De resterende 20 pct. af observationerne udgør testsættet. Det er vores out-of-sample data, som de trænede modeller skal afprøves på. Logikken kan genfindes illustreret i figur 3.5 på side 40. Når vi igennem kapitlet gennemgår modellernes resultater, er det udelukkende disse out-of-sample resultater, som vi referer til. Det gælder også, når vi i afsnit 5.4 anvender prædiktionsmodellen i praksis. Dermed evaluerer vi modellernes evne til at forudsige frafaldet blandt disse for modellen ukendte studerende.

5.1 Sammenligning af modeller

Vi vil i dette afsnit sammenligne resultaterne af de fire algoritmer. Som vi beskrev i afsnit 3.1.2, er AUC et godt mål til at sammenligne forskellige modellers performance. Tabel 5.1 oplister rækkevis AUC-værdierne for de fire modeller.

Tabel 5.1: AUC-værdier for fire modeller og tre tidshorisonter

Model	AUC-værdier		
	30 dage	Halvt år	Helt år
Logistisk regression	0,609	0,647	0,646
Klassifikationstræ	0,531	0,601	0,656
Random Forest	0,648	0,668	0,692
Gradient Boosted Trees	0,674	0,692	0,727
Antal observationer i sættet	23.107	23.066	21.820
Antal frafald i sættet	306	1711	2952
Andel frafald	1,3 %	7,4 %	13,5 %

Tabellen har tre kolonner, fordi hver af algoritmerne er trænet for tre tidsperioder på hhv. 30 dage, et halvt år og et helt år efter studiestart. På den måde kan vi empirisk afgøre, hvilken tidshorisont som giver bedst resultater. Som det fremgår af tabellen, falder det samlede antal observationer en smule, når vi forlænger perioden. Det skyldes, at der er studerende med i det fulde datasæt, som endnu ikke har gået et halvt eller et helt år på studiet. Omvendt gælder det, at jo længere en periode vi betragter, des flere studerende falder fra. Bortset fra denne forskel i outcome-variablen, er der ingen forskelle mellem datasættene for de tre perioder. Tabellens tre kolonner bygger altså på præcis de samme informationer om de studerende – informationer, som alle er tilgængelige ved studiestart.

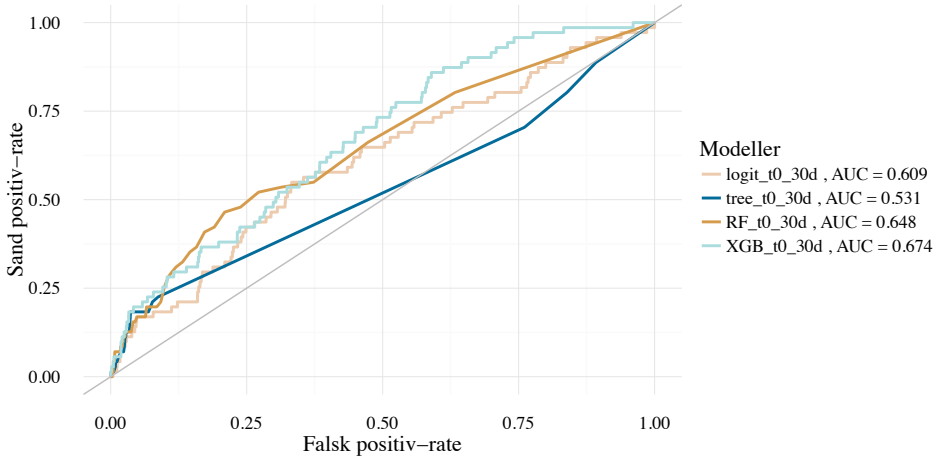
Det interessante er derfor, at alle modellernes AUC-værdi stiger med længere frafaldsperiode. Kun den logistiske regressionsmodels forbedring flader ud mellem et halvt og et helt år. Modellerne bliver altså bedre i takt med, at de får flere frafald at træne på. Det er ikke nogen stor overraskelse, at det forholder sig sådan. I begyndelsen er der trods alt relativt få frafald, som modellerne kan lære mønstre efter – kun 1,3 pct. af observationerne har et positivt outcome de første 30 dage. Imidlertid kunne det omvendte også have været tilfældet: at modellerne klarede sig bedre på den korte bane, når vi forudsiger frafald inden for de første 30 dage. Det virker sandsynligt, at det frafald, som sker lige efter forudsigelsestidspunktet, er nemmere at forudsige end senere frafald. Der kunne også være fællestræk blandt de tidligt frafaldne, som gjorde dem særligt identificerbare for modellerne. Det lader dog jf. tabel 5.1 ikke til at være tilfældet.

Vi bliver bestyrket i denne opfattelse ved at betragte ROC-kurverne for disse 30 dagesmodeller, som fremgår af figur 5.1 på den følgende side. Når vi sammenligner med den grå, diagonale linje, der repræsenterer tilfældige gæt, kan vi se, at særligt det simple træs forudsigelser er helt ude i skoven – for nu at tillade sig en letkøbt metafor. Over kurvens sidste stræk dykker den nedunder den diagonale linje. Havde man ønsket en sand positiv-rate (SPR) på over 60 pct., havde tilfældige gæt altså forudsagt flere rigtigt end klassifikationstræet. Det kan også bemærkes, at alle ROC-kurverne er relativt volatile og takkede, hvilket skyldes de få frafald inden for tidshorisonten, hvormed SPR er meget sensitiv over for forudsigelsen af de enkelte frafald.

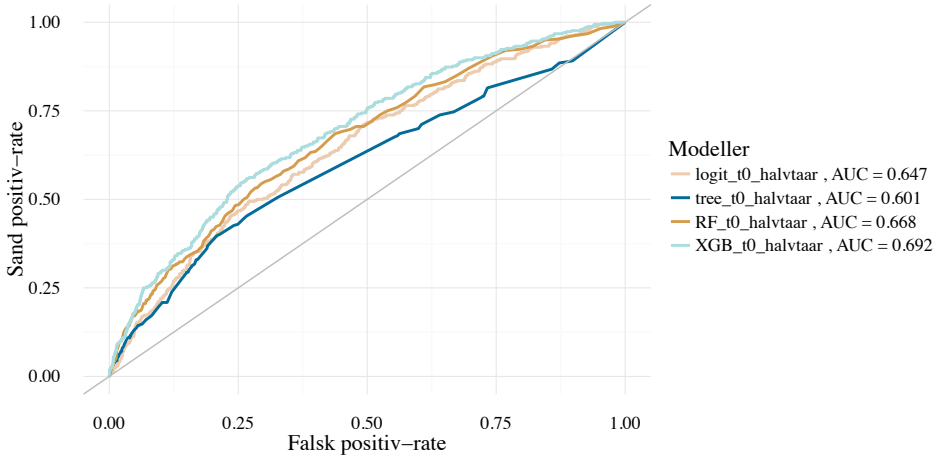
Figur 5.2 og 5.3 på næste side plotter ROC-kurverne for modellerne, der forudsiger frafald inden for henholdsvis det første halve og hele år på studiet. Sammenlignet med de 30 dage forbedres alle modellernes performance, når tidshorisonten bliver længere. Samtidig kan vi se, at det ikke er nemmere at forudsige det "nære" frafald inden for det første halve år end frafaldet inden for et helt år – de flere frafald i den længste tidsperiode opvejer altså, at frafaldet først sker længere fra forudsigelsestidspunktet.

GBT-modellen når i figur 5.3 op på en AUC-værdi på 0,727. Selvom den har den højeste AUC-værdi for alle tidshorisonter, er det først med et års frafald, at den performer entydigt bedst over hele ROC-kurvens strækning. Denne grafiske vurdering af ROC-kurverne er nødvendig, fordi AUC kun er et heuristisk mål for arealet under ROC-kurverne (Witten et al. 2005: 173). Hvis vi fx igen betragter figur 5.1 på den følgende side for de 30 dage, så kan vi se, at kurverne bugter sig forskelligt, og at to vidt forskellige kurver derfor kan give samme AUC-værdi. Til en tidshorisont på 30-dage klarer RF-modellen sig fx bedre på det første stræk af ROC-kurven. For GBT-modellen modsvares en sand positiv-rate på 50 pct. her af en falsk positiv-rate på 30 pct., mens RF-modellen opnår det samme med en falsk positiv-rate på kun 25 pct. Først når vi træner på et helt

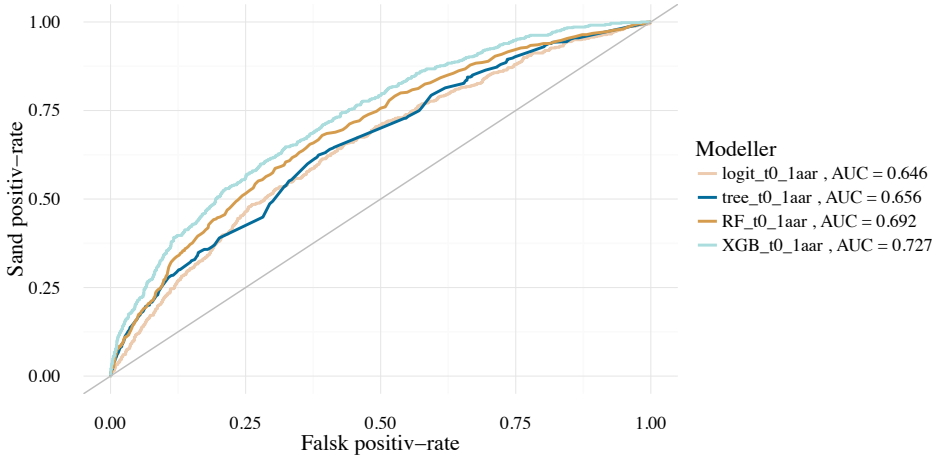
Figur 5.1: ROC-kurver for modeller på 30 dages frafaldsdata



Figur 5.2: ROC-kurver for modeller på første semesters frafaldsdata



Figur 5.3: ROC-kurver for modeller på et helt studieårs frafaldsdata



års frafaldsdata er GBT-modellen entydigt bedst, uanset hvordan man senere ønsker at sætte en konkret tærskelværdi og afveje sande og falske positive mod hinanden – et trin vi vender tilbage til senere i analysen (afsnit 5.4).

5.1.1 Usikkerhed og resultaternes pålidelighed

Ligesom der er standardfejl i estimationssammenhæng, er der også en statistisk usikkerhed forbundet med AUC-værdierne rapporteret i tabel 5.1 på side 64, hvor vi sammenligner modellernes prædiktions-performance. Der er dog ikke tradition for at beregne og rapportere usikkerhedsmål for prædiktionmodeller, hvilket kan skyldes, at usikkerhed forstås og behandles forskelligt i estimations- og prædiktionssammenhæng (Hofman et al. 2017: 1–3; James et al. 2013: 187).

Usikkerheden på AUC stammer fra to kilder. Den første er, at vi splitter datasættet i et trænings- og et testsæt. Den anden er, at to af algoritmerne anvendt i analysen, RF og GBT, er baseret på resampling-metoder (jf. afsnit 3.3). Selvom vi arbejder med populationsdata, introducerer de to kilder altså en sampling-usikkerhed, når vi træner modellerne, hvilket betyder, at AUC kan variere.

I maskinlæringslitteraturen er der ikke enighed om, hvordan usikkerhed bedst estimeres. Der findes forskellige bud, men de er relativt tekniske, og i forsøget løber man imod den problemstilling, at det ikke er klart, om AUC eksempelvis er normalfordelt omkring et gennemsnit, hvad der er fundamentet for at estimere og fortolke standardfejl i estimationssammenhæng. Derudover spiller usikkerhederne også en anden rolle ved prædiktion, fordi formålet her ikke er at estimere en strukturel, datagenererende proces. AUC er et mål for modellens fit out-of-sample. AUC-værdier kan derfor ikke sammenlignes med β-koefficienter fra en regressionsmodel, men har mere tilfælles med R^2, som vi heller ikke estimerer standardfejl for.

Årsagerne nævnt her kan være grunden til, at det ikke er almindelig praksis at estimere et mål for AUC-værdiers usikkerheder. Der estimeres eksempelvis ikke usikkerheder for AUC eller andre performancemål i nogen af de case-studier, som vi refererer til i litteraturgennemgangen. Ikke desto mindre er det relevant at holde in mente, at vi på baggrund af tabel 5.1 på side 64 ikke kan konkludere, om forskellene mellem GBT-modellens AUC og og de øvrige modellers AUC er statistisk signifikante. Hvorvidt forskellene er statistisk signifikante står dog ikke centralt for vores problemstilling, hvor vi ønsker at undersøge potentialet af maskinlæring til målretning af tiltag. Til dette formål ønsker vi at gå videre med den algoritme, som performer bedst – om forskellen også er statistisk signifikant, er ikke afgørende.

I den resterende del af analysen går vi i dybden med GBT-modellen for derpå at illustrere, hvordan modellen kan anvendes i praksis. Usikkerheden til trods omtaler vi modellen som "den bedste model", fordi den systematisk leverer de bedste prædiktioner out-of-sample for alle tre tidshorisonter, hvilket er grunden til, at vi går videre med denne model og ikke de øvrige.

5.2 Den bedste model

Vi går i dette afsnit i dybden med den endelige model, Gradient Boosted Trees-algoritmen trænet på et helt studieårs frafald. Først vil vi evaluere dens performance. Derefter vil vi demonstrere, hvilken betydning tuningen har for modellens performance, samt hvilke variable der har størst prædiktiv værdi.

5.2.1 Vurdering af modellens performance

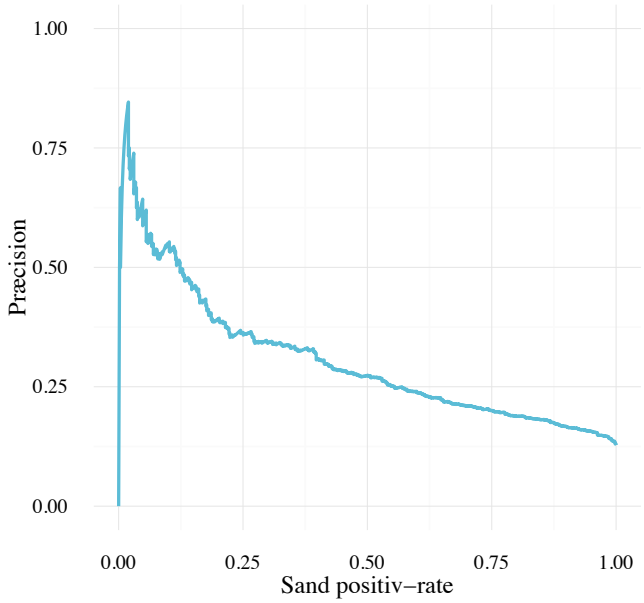
Gradient Boosted Trees-modellen er med en AUC-værdi på 0,727 en væsentligt forbedring i forhold til vores baseline-model, logistisk regression, som nåede op på 0,647. Dermed skriver vi os ind i rækken af studier, som jf. afsnit 2.2 på det seneste har opnået gode resultater med GBT-algoritmen. Samtidig er performance på niveau med lignende studier af maskinlæring i forvaltningen. Harmsen & Enggaard (2016) når eksempelvis op på en AUC-værdi på 0,702, da de forsøger at forudsige genindlæggelser af KOL-patienter på et hospital. Aulck et al. (2016) når en AUC-værdi på 0,729 for forudsigelsen af frafald på amerikanske universiteter, og Kleinberg et al. (2017) en AUC på 0,707 for forudsigelsen af recidiv blandt fængslede løsladt mod kaution.

Vores AUC-værdi er dog ikke prangende, hvis vi sammenligner med modellen, der blev udviklet til at forudsige frafald i danske gymnasier ved studiestart. Her når den bedste model (baseret på Random Forest) op på en AUC på 0,86 (Kristoffersen 2015: 50). En så høj AUC-værdi skal dog ses i lyset af de helt særlige dataforhold, som gymnasierne nyder godt af, idet fx elevernes afleveringer, faglige præstationer og fravær registreres på daglig basis. Endvidere bliver forudsigelserne først lavet efter en måneds skolegang, og det må forventes, at AUC havde været betragteligt lavere, hvis modellen som her skulle have været trænet før første skoledag.

For at vurdere GBT-modellens performance kan vi supplere ROC-kurven med et par andre nyttige visualiseringer. ROC-kurven viser som bekendt et tradeoff mellem SPR og FPR. Et tredje muligt performancemål er modellens *præcision* defineret som andelen af sande positive blandt alle forudsagte positive, dvs. SP / (SP + FP). Hvor SPR fortæller,

hvor mange af de faktiske frafald, som modellen opdager, fortæller præcisionsmålet, hvor mange af dens forudsagte frafald, som faktisk er sande. Det kan være et nyttigt mål i tilfælde som vores, hvor vi primært er interesserede i de positive outcomes. Ligesom de andre mål varierer præcisionen med tærskelværdien. I figur 5.4 har vi plottet præcisionen som funktion af sand positiv-raten. Det er en anden måde at illustrere tradeoff'et fra ROC-kurven ved at fokusere mere eksklusivt på de positive outcomes (dvs. frafaldene) (Witten et al. 2005: 171). Figuren illustrerer den grundlæggende sammenhæng, at jo flere af de frafaldne vi ønsker at identificere, jo større andel af falske positive følger med.

Figur 5.4: Præcision vs. sand positiv-rate

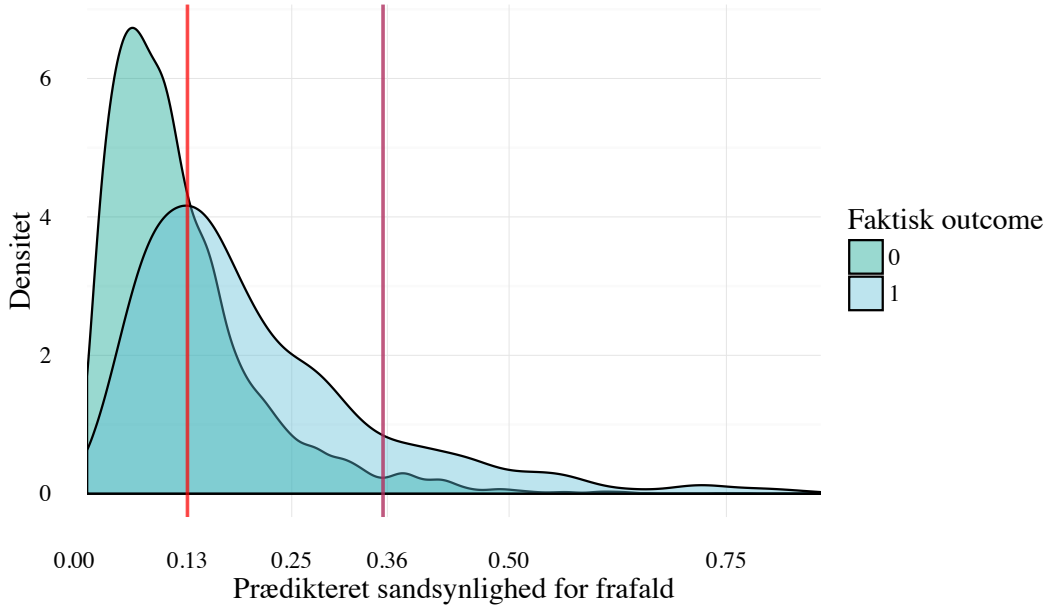


Det ses, at vi fx kan identificere 50 pct. af alle frafald (SPR = 0,5) med en præcision på 25 pct. Med andre ord udgør de sande positive kun en fjerdedel af alle dem, som vi forudsiger vil frafalde. Figuren viser, hvordan præcisionen falder drastisk for vores model, når vi øger sand positiv-raten. Det kommer med andre ord med et stort tab af præcision (flere falske positive), når vi gerne vil identificere så mange af frafaldene som muligt. Faldet af præcision er størst i begyndelsen, hvorimod tabet i præcision er relativt mindre, hvis vi fx ønsker at hæve andelen af de identificerede frafald, SPR, fra 50 pct. til 75 pct.

Figur 5.5 er en anden effektiv måde at visualisere dette forhold (Harmsen & Enggaard 2016: 87–89). Her har vi plottet et densitets-histogram over de forudsagte sandsynligheder for de to grupper: faktisk frafaldne (1) og ikke-frafaldne (0). Fremstillet på denne måde, er det tydeligt, at modellen langt fra formår at skelne perfekt mellem

de positive og negative tilfælde. De to histogrammer skulle i så fald have ligget i hver sin ende af spændet af tærskelværdier for prædikteret sandsynlighed. Der er i praksis et stort overlap, som betyder, at vi ikke kan få en høj SPR uden også at acceptere en høj FPR og dermed lav præcision. Sætter vi en høj tærskelværdi for prædikteret frafaldssandsynlighed (fx den lilla linje) frem for en lavere (fx den røde linje), har modellen relativt høj præcision. En stor andel af dem, som vi forudsiger falder fra, gør det rent faktisk. Til gengæld er SPR lav – dvs. modellen forudsiger kun en relativt lille andel af alle frafald. At histogrammet for de faktiske frafald ligger til højre og ikke overlapper fuldstændig, fortæller dog samtidig, at modellen i noget omfang kan bruges til at skelne mellem de to grupper.

Figur 5.5: Densitets-histogram af prædikteret vs. faktisk frafald



5.2.2 De vigtigste variable

I dette afsnit undersøger vi, hvilke variable som har størst indflydelse i modellens prædiktioner. Det gør vi på baggrund af tre såkaldte *importance*-mål, som kan beregnes for GBT-modellen: *gain*, *cover* og *frequence*. Gain er et mål for, hvor meget en given variabel i gennemsnit minimerer træernes loss-funktioner, for hver gang variablen tilføjes et træ som led i boosting-processen. Cover er et mål for, hvor mange observationer i data, som er knyttet til den givne variabel. Frequence er et mål for, hvor mange gange en given feature indgår i modellens træer. Alle målene er relative – hver variabel har en værdi for gain, cover og frequence mellem 0 og 1, og summen af alle varibles gain, cover

og frequence er 1 henholdsvist. Målene udtrykker derfor en given variabels importance relativt til de øvrige variable (Chen et al. 2016: 31).

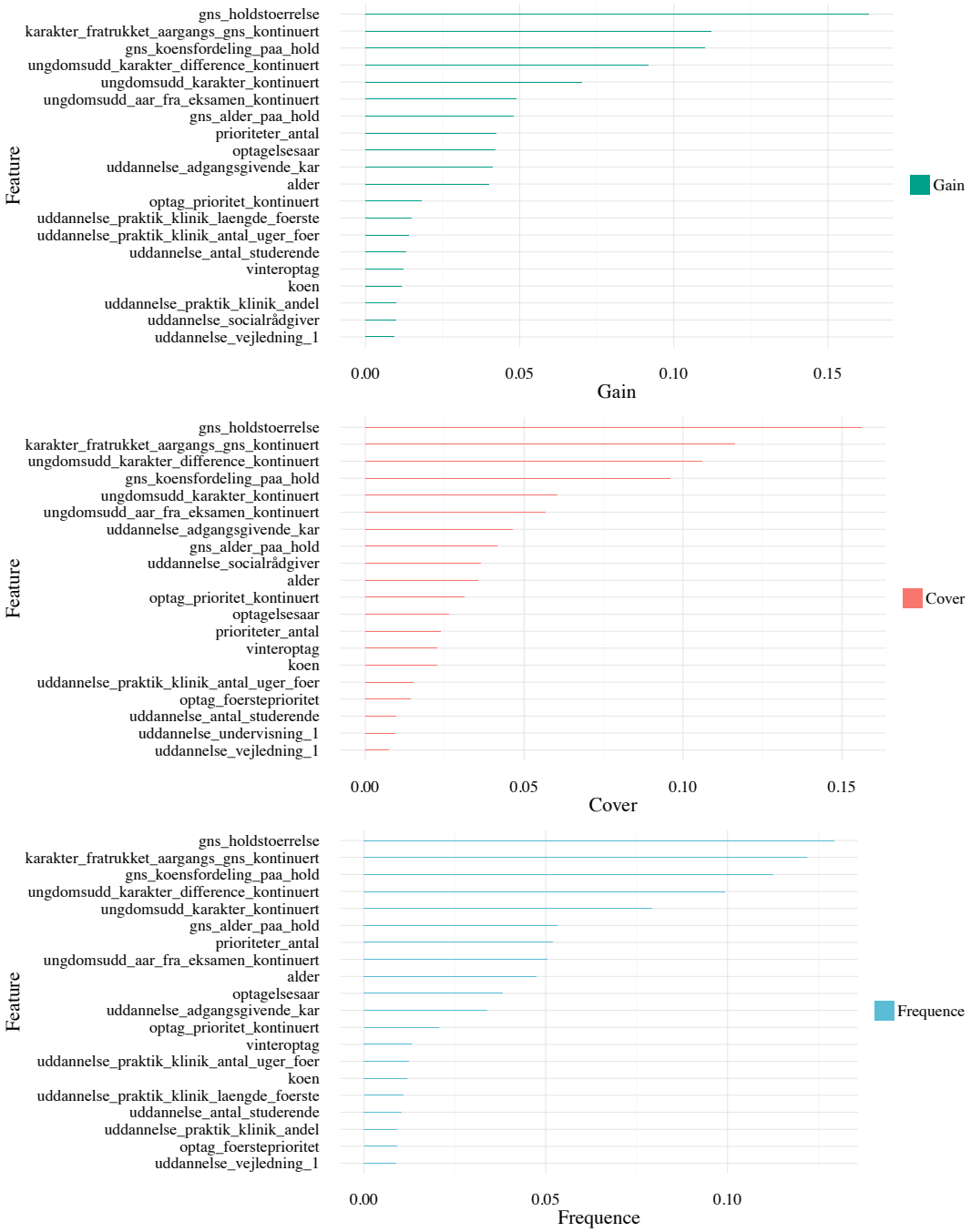
Importance-målene er interessante, fordi de indikerer, hvilke variable som driver vores prædiktioner. Det kan i noget omfang bidrage til at åbne den *black box*, som algoritmen GBT kan forekomme som. I fortolkningen af importance-mål skal man imidlertid have nogle forhold in mente. Det er vigtigt at huske, at træerne alene er korrelationsbaserede, og at de i øvrigt bygger på variable, som i vidt omfang kan være kollineære. Konsekvensen af førstnævnte er selvsagt, at variablene ikke kan tillægges en kausal fortolkning. Konsekvensen af sidstnævnte er, at en given variabel fx kan score højt på målet frequence, selvom den ikke har noget med outcome at gøre, hvis variablen blot er kollineær med andre variable, som hænger sammen med outcome. Derudover angiver importance-målene heller ikke retningen på en variabel. Hvis køn fx er en indflydelsesrig variabel, gør importance-målene os ikke i sig selv klogere på, om det er kvinder eller mænd (eller en øvrig kønskategori), som i størst omfang frafalder – eller hvilke andre variable i træerne denne sammenhæng er betinget af.

De tre søjlediagrammer på side 72 viser de 20 variable med højeste gain, cover og frequence, og i tabel 5.2 på side 73 fremgår de variable, som ligger blandt de 20 mest betydningsfulde variable i alle tre mål. For de 17 variable i tabellen angiver vi variablenes importance som summen af gain, cover og frequence.

Som det ses af figurerne og tabellen, er der flere variable, der går igen med stor indflydelse på tværs af de tre importance-mål. De fem variable med størst indflydelse er den gns. størrelse på den studerendes hold, den studerendes karaktersnit ift. sin årgang, den gns. kønsfordeling på holdene, den studerendes karaktersnit fra sin ungdomsuddannelse samt differencen ml. dette snit og uddannelsens adgangskvotient. Summen af gain, cover og frequence for de fem variable er 1,6, hvilket er lidt mere end halvdelen af modellens samlede forklaringskraft (summen af gain, cover og frequence for samtlige variable er 3).

Det er bemærkelsesværdigt, at de studerendes karakterer fra deres ungdomsuddannelse indgår i tre ud af de fem mest indflydelsesrige variable. Karaktersnittet indgår både som mål i sig selv og relativt til både årgangens snit og det adgangsgivende snit. De to sidstnævnte er variable, som vi konstruerede som en del af vores feature engineering. Som led i denne beregnede vi også de to øvrige variable i top fem, der drejer sig om karakteristika ved den studerendes hold, henholdsvis holdstørrelsen og kønsfordelingen. Holdstørrelsen har en markant højere importance end de øvrige variable. Det virker oplagt at drage en forbindelse mellem variablene kønsfordeling og holdstørrelse og de to faktorer social og akademisk integration, som jævnfør litteraturgennemgangen i afsnit

Figur 5.6: De 20 variable med højest importance



Tabel 5.2: De mest indflydelsesrige variable (sum af gain, cover og frequence)

Variabel	**Importance**
1. Gns. holdstørrelse	0,449
2. Karaktersnit fra ungdomsudd. fratrukket årgangens gns.	0,350
3. Gns. kønsfordeling på hold	0,319
4. Karaktersnit fra ungdomsudd. fratrukket adgangskvotient	0,297
5. Karaktersnit fra ungdomsuddannelse	0,210
6. Antal år fra eksamen på ungdomsuddannelse	0,156
7. Gns. alder på hold	0,143
8. Alder	0,123
9. Adgangskvotient på uddannelsen	0,122
10. Antal prioriteter ved ansøgning	0,118
11. Optagelsesår	0,107
12. Optaget på hvilken prioritet	0,070
13. Vinteroptag	0,048
14. Køn	0,046
15. Antal uger før første praktik	0,042
16. Antal studerende på uddannelsen	0,033
17. Andelen af vejledning på første år	0,026

2.1 er veletablerede som årsager til frafald. Forbindelsen synes plausibel. Vi kan dog ikke strække konklusionen meget mere vidtgående, da vi fx ikke ved, om en mere homogen kønssammensætning på holdene styrker den sociale og akademiske integration – eller om en meget homogen kønssammensætning betyder mindre integration og dermed større frafald for kønsminoriteten.

Problematikken ovenfor illustrerer det begrænsede rum til fortolkning af en korrelationsbaseret model og dennes importance-mål. Det illustrerer videre, at variablenes importance ikke kan stå alene som et grundlag for handling i praksis – en studievejleder kan ikke kaste et blik ned over variablene og målrette et tiltag på baggrund af dem. Det kan vedkommende for det første ikke, fordi importance-målene som sagt ikke fortæller noget om retningen på sammenhængen mellem en given variabel og outcome. For det andet fordi der kan være tale om komplekse interaktioner og betingede sammenhænge. Når en given variabel scorer højt på frequence ved vi eksempelvis ikke, om dette er betinget af en lang række split længere oppe i træerne og således komplekse interaktioner med øvrige variable. Målretning af tiltag bør derfor ikke baseres på enkeltstående variable med høj importance, men bør baseres på modellen i sin helhed.

Det kan bemærkes, at der tilsyneladende var ræson i vores håndtering af manglende værdier i kontinuerte variable jf. afsnit 4.1.4. De kontinuert skalerede varianter af variab-

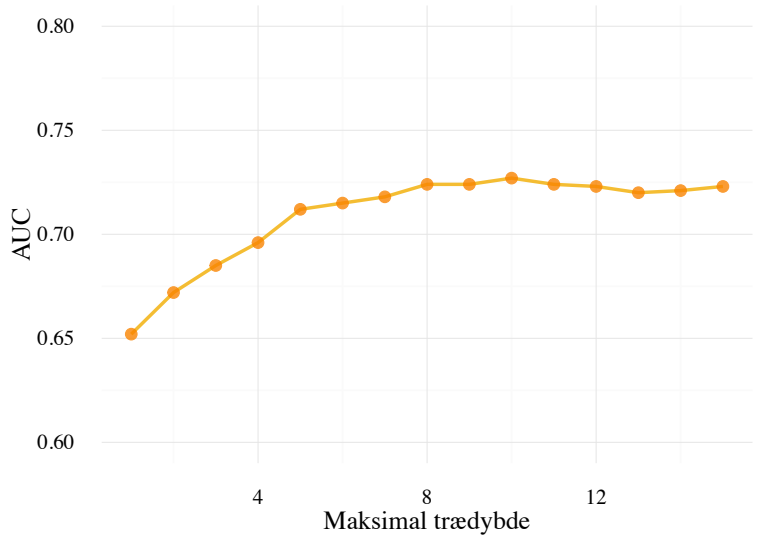
lene er anvendt hyppigere end de ækvivalente faktoriserede variable, hvilket illustrerer nytten af vores tilgang baseret på MIA (*Missingness Incorporated in Attributes*).

5.2.3 Tuning-processens betydning

Som vi genkalder os fra afsnit 3.3.4, har GBT-modellen en række frie parametre, som skal tunes for at finde det rigtige niveau af modelkompleksitet. I dette afsnit beskriver vi, hvordan vi fastsætter de frie parametre for GBT-modellen, og hvordan det påvirker dens performance.

Tuning-processen kan beskrives som mere en kunst end en videnskab (Snoek et al. 2012). I bund og grund består processen i at afprøve forskellige kombinationer af parametre og vælge den bedste. Den bedste kombination er den, som resulterer i den højeste AUC-værdi i krydsvalideringen. Intuitionen bag tuning kan illustreres med figur 5.7. Her har vi med udgangspunkt i vores endelige model holdt alle parametre konstant og kun varieret parametren `max_depth`, som regulerer klassifikationstræernes maksimale dybde. Som det ses, er AUC højest for en trædybde på 10. Skruer vi op eller ned for parametren, falder AUC. Den samme type figur kunne have været tegnet for alle modellens andre parametre. Udfordringen ligger i at finde den kombination af parametre, som tilsammen giver den højeste AUC.

Figur 5.7: AUC som funktion af trædybde

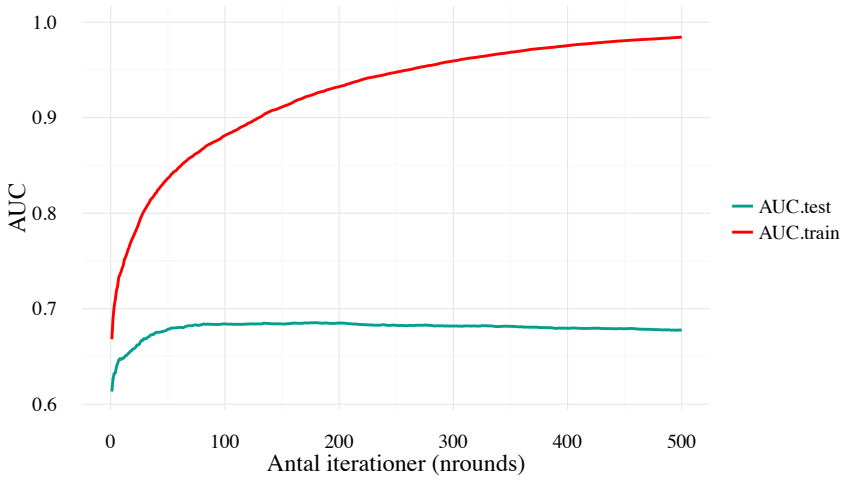


Jo flere frie parametre en model har, desto vanskeligere bliver det at finde den bedste kombination. Afprøvningen kan dog automatiseres på forskellig vis for at sandsynliggøre, at mulighedsrummet er ordentlig afprøvet. Ofte anvendes såkaldt *grid search*, hvor

man angiver en vektor af værdier for hver parameter og beder computeren om at afprøve alle de mulige kombinationer. For hver kombination af parametre skal modellen trænes med krydsvalidering – og det er en proces, som hurtigt kræver mange computerkræfter. Det sætter begrænsninger for, hvor fintmaskede intervaller af parameterværdier, som det er muligt at afprøve i praksis.

Vi bruger en anden tilgang til at tune GBT-modellen kaldet *bayesiansk optimering*, som er implementeret i R-pakken `rBayesianOptimization` (Yan 2016). Her er den grundlæggende idé, at man for hver afprøvning af en parameter-kombination får information, der kan bruges, når man skal vælge nye parametre til afprøvning. Hvis en bestemt kombination giver et lovende resultat, er det værd at prøve en lignende kombination med ganske få justeringer af parametrene. Omvendt hvis en kombination giver dårlige resultater, er det værd at prøve noget helt andet. En gennemgang af matematikken bag bayesiansk optimering ligger uden for rammerne af denne opgave. Men intuitionen er, at man udnytter information fra alle de tidligere evalueringer – frem for blot at evaluere den næste prædefinerede kombination af parametre (Snoek et al. 2012). Det første trin er at bede den bayesianske algoritme afprøve en masse tilfældige kombinationer af parameter-værdier for vores GBT-model. På baggrund af disse modellers performance beder vi den dernæst forudsige den optimale kombinationen af parametre. Disse nye parametre bliver nu afprøvet, og informationen taget til efterretning. Sådan fortsætter algoritmen med at afprøve og opdatere sin forudsigelse i et prædefineret antal iterationer.

Figur 5.8: Trænings- og test-AUC som funktion af antal iterationer



GBT-modellens mest betydningsfulde parametre er antallet af iterationer, `nrounds`, og læringsraten, η. Som vi viste i figur 3.9 på side 51 er balancen mellem disse to afgørende for ikke at under- eller overfitte. Sænkes den ene, skal den anden hæves.

Vi har holdt `nrounds` konstant på 100 og tunet η derefter. Det har vi gjort, fordi vi foretrækker at holde antallet af iterationer relativt lavt, da det er den computermæssigt mest omkostningsfulde parameter at øge. Figur 5.8 på forrige side illustrerer, hvordan tuningen af vores model har virket efter hensigten. De to grafer viser, at AUC-værdien for vores trænings- og testsæt stiger for hver ny boosting-iteration. Efter omtrent 100 iterationer er der ikke mere prædiktiv værdi at hente, og kurven for testsættet flader ud og daler endog en smule. Det passer med, at vi har tunet læringsratens størrelse til at matche 100 iterationer. Træningssættets AUC-værdi stiger ufortrødent i takt med, at nye iterationer føjes til modellen, der efterhånden fitter perfekt til træningssættet. Efter de første 100 iterationer er der dog reelt tale om overfitting til træningsdata – med dårligere performance på testdata til følge.

I tabel 5.3 har vi oplistet parametrene for vores bedste GBT-model med og uden tuning. Uden tuning bruger algoritmen et sæt standard-parametre. Som det ses af tabellen, hæver tuningen modellens AUC fra 0,702 til 0,727. Betydningen af parametre er altså ikke altafgørende, men nødvendig for at maksimere algoritmens performance. Uden tuning ligger GBT-modellens performance på niveau med fx Random Forest-modellens.

Tabel 5.3: Sammenligning af GBT-modellen med og uden tuning

Parametre	GBT-model	
	Med tuning	Uden tuning
`nrounds`	100	100
η	0.066	0.30
`max_depth`	10.0	6.0
`min_child_weight`	4.0	1.0
`gamma`	1.0	0.0
`lambda`	2.0	1.0
`colsample_bytree`	0.81	1.0
`subsample`	1.0	1.0
AUC-værdi	0.727	0.702

Den optimale kombination af frie parametre afhænger af den underliggende struktur i det enkelte datasæt. Derfor har vi tunet modellen særskilt til hver modelspecifikation, dvs. på frafaldsdata for forskellige tidshorisonter. De resulterende parameter-sæt fremgår af bilag A.

5.3 Andre modelspecifikationer

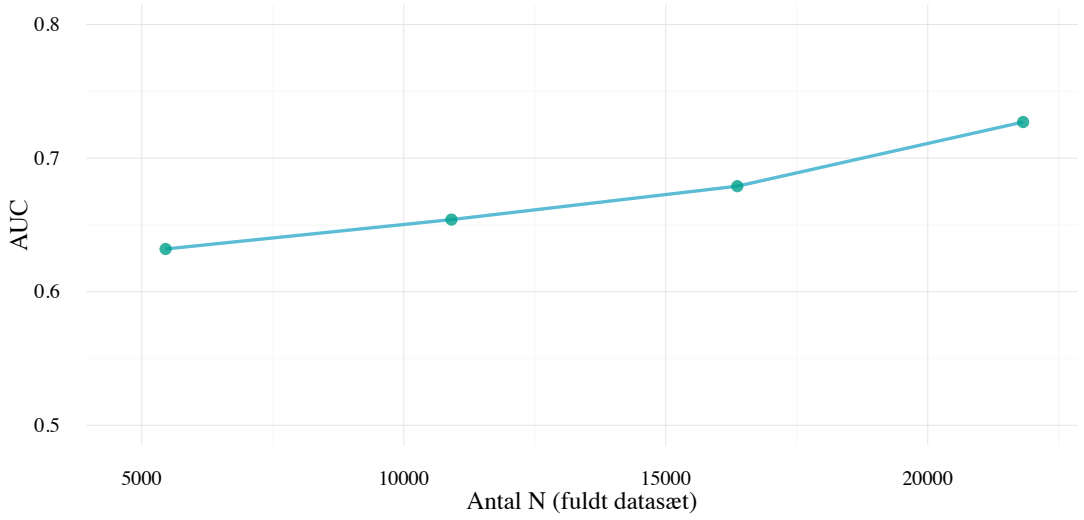
I dette afsnit undersøger vi, hvordan afgrænsninger af data og tidsperioder har betydning for modellen. Vi undersøger betydningen af datasættets størrelse, den prædiktive værdi af logdata og modellens performance, efter de studerende har gået på studiet et år.

5.3.1 Betydningen af datasættets størrelse

Vi kan undersøge betydningen af datasættets størrelse N ved at udtrække tilfældige samples i forskellige størrelser af det samlede datasæt og derefter fitte modellen på hvert subsample. Vi udtrækker samples på hhv. 75, 50 og 25 pct. af datasættes størrelse, og resultatet af at bruge modellen på disse subsamples fremgår af figur 5.9.

Det ses af figuren, at AUC falder fra ca. 0,73 til ca. 0,63, når antallet af observationer falder fra omtrent 20.000 (100 pct.) til omtrent 5.000 (25 pct.). Det er et ganske betydeligt fald i AUC – det svarer cirka til at bygge vores prædiktionsmodel på logistisk regression frem for Gradient Boosted Tress (som vist i tabel 5.1 på side 64).

Figur 5.9: AUC-værdier og sample-størrelser



En række studier har desuden peget på, at det ikke kun er datasættets antal observationer, der har betydning, men også fordelingen af observationerne på outcome-variablen (Mollineda et al. 2007; Chawla et al. 2002). I datasættet anvendt her, frafalder 15,6 pct. at de studerende inden for det første studieår, hvilket betyder, at 15,6 pct. af observationerne ligger i den ene kategori af den binære outcome-variabel. Datasættet er med

andre ord ubalanceret. Vi har forsøgt at balancere datasættet på tre udbredte måder, hhv. ved oversampling af minoritetskategorien, undersampling af majoritetskategorien og ved at justere vægten af fejlklassifikationerne, så det vejer relativt tungest, når en frafalden studerende klassificeres forkert (Chawla et al. 2002; Mollineda et al. 2007; Witten et al. 2005: 164–166). I flere af de tidligere casestudier på uddannelsesområdet har sådanne teknikker været taget i brug for at forbedre performance (se fx Dekker et al. 2009; Bayer et al. 2012; Kristoffersen 2015). Vores resultater ved brug af balancering findes i bilag B på side 140. Ingen af tilgangene forbedrede dog prædiktionsevnen for vores model.

Det tyder altså på, at frafaldsmodellen i casen Metropol ikke er specielt sensitiv over for balancen i datasættet, men dog er sensitiv over for datasættets størrelse. Sammenhængen mellem N og AUC er velkendt i maskinlæringslitteraturen (Kleinberg et al. 2017). Det indikerer, at det generelt som beslutningstager vil være relevant på forhånd at overveje, hvilken *mængde* af data, der er til rådighed, før der i en given forvaltning træffes beslutning om at udvikle og implementere en prædiktionsmodel.

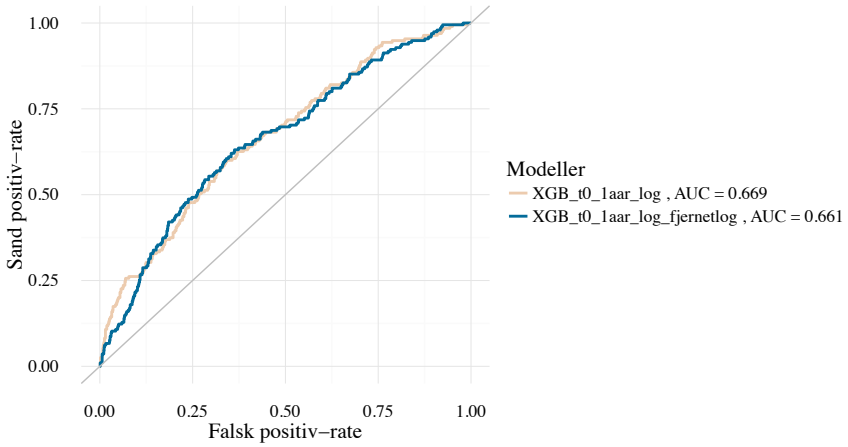
5.3.2 Betydningen af logdata

I dette afsnit vil vi kort undersøge potentialet af de studerendes logdata. Det er ikke medtaget i vores modeller ovenfor, fordi det kun er tilgængeligt for en begrænset tidsperiode. Som vi skal se nedenfor, er der så begrænset prædiktiv værdi at hente, at vi ikke har fundet det værd at inkludere logdata i den samlede model ved imputation. På grund af logdatas særlige karakter og hypen om big datas potentiale i forvaltningsøjemed foretager vi her en kort særskilt analyse.

I figur 5.10 på den følgende side viser vi ROC-kurverne for to modeller, som kun er trænet på den del af datasættet, hvor der eksisterer logdata. I den ene model er de 12 variable om de studerendes adfærd på intranettet medtaget. I den anden model er de udeladt.

Det ses, at begge modeller har en langt mindre AUC end vores bedste model. Det er ikke så overraskende. Vi har netop konkluderet, at AUC er afhængigt af N, og når vi kun ser på den del af datasættet, hvor der er logdata, falder antallet af observationer til under en tredjedel. Det overrasker os til gengæld, at det målt ved AUC er nærmest uden betydning at fjerne logdata. Modellen med variable om logdata leverer kun marginalt bedre forudsigelser end den, hvor variablene ikke er medtaget.

Figur 5.10: ROC-kurver for datasæt med og uden logdata



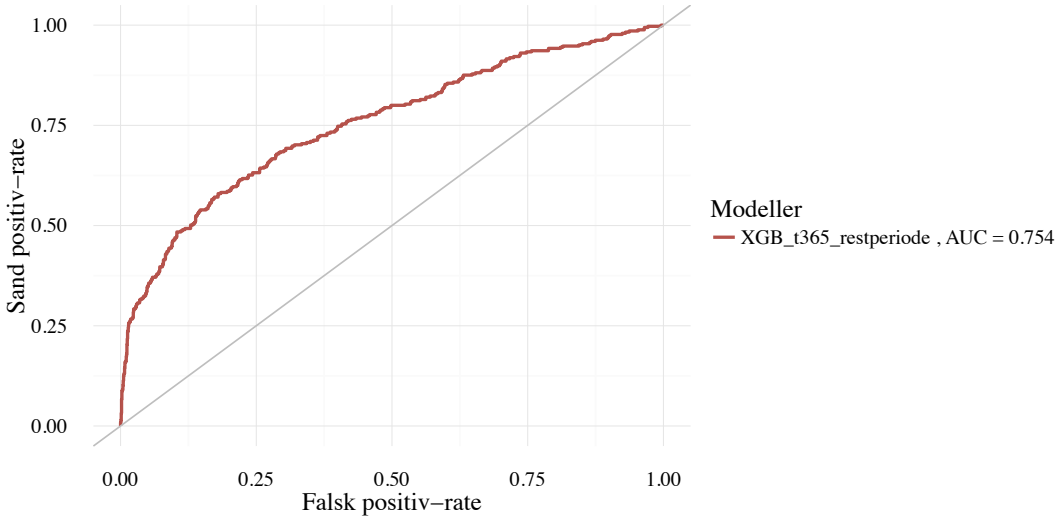
På baggrund af ovenstående er det konklusionen, at logdata i vores case ikke bidrager med nævneværdig forudsigelseskraft. Det kan ikke sluttes på baggrund af denne case, at logdata aldrig vil have betydning, eller at data med big data-karakter ikke kan have en prædiktiv værdi i en anden policy-kontekst. Fundet i casen Metropol rækker dog ud over netop denne case i det omfang, at det taler for en vis skepsis omkring at opstille store forventninger til datas prædiktive værdi, før data rent faktisk er blevet anvendt i praksis.

5.3.3 Modeller senere i studieforløbet

Indtil nu har vi fokuseret på modeller til forudsigelse af frafald allerede ved studiestart. Der kan dog være en værdi i at følge op på de studerendes risiko for frafald undervejs i studieforløbet. Det viser vi et eksempel på i figur 5.11 på næste side, hvor vi træner og tester en prædiktionsmodel, der medtager data for det første studieår og prædikterer frafald inden for det resterende studieforløb.

Modellens AUC er 0,754, hvilket er højere end den bedste model til at forudsige frafald ved studiestart, hvor AUC'en er 0,727 (se tabel 5.1 på side 64). Det er med andre ord nemmere at forudsige frafald, når vi står på dagen, hvor ét studieår allerede er passeret, sammenlignet med dagen for studiestart. Det kan fx skyldes, at der er mere data til rådighed efter ét studieår, såsom karakterdata. Den hypotese sandsynliggøres blandt andet, når vi kaster et blik på modellens mest indflydelsesrige variable, som fremgår af tabel 5.4 på side 81. Her ses det, at variablene med den 1., 3., 4. og 7. største importance alle vedrører data om de studerendes karakterer på Metropol, som ikke var tilgængelige ved studiestart. En øvrig forklaring på den højere AUC kan være, at frafaldet inden for

Figur 5.11: ROC-kurve for forudsigelse af frafald efter første studieår



det første studieår er mere tilfældigt og mindre forudsigeligt i sig selv sammenlignet med frafaldet senere i studieforløbet.

Det kan endvidere bemærkes i figur 5.11, at modellens ROC-kurve er meget stejl på det første stræk, hvilket er en positiv egenskab. Det betyder, at vi kan få en relativt høj sand positiv-rate, før vi begynder at få mange falske positive med i købet. Helt konkret har vi med vores model mulighed for at forudsige 25 pct. af de faktisk frafaldstruede studerende næsten uden nogen falske positive.

Det er endvidere værd at bemærke, at antallet af observationer i denne model (N = 9825) er markant mindre end i modellen, der forudsiger frafald inden for det første studieår. Det skyldes, at der er flere studerende, som falder uden for den afgrænsning, at de skal have gennemført et helt studieforløb, hvilket er en betingelse for at indgå i denne model. Implikationen er, at det altså ikke er et større N, der kan forklare, at denne model leverer bedre prædiktioner. I fald den bedre performance skyldes et mere rigt datasæt er det en konklusion, som vi vil forvente rækker ud over casen Metropol: bedre data giver bedre forudsigelser. Det er ikke så overraskende, men det er ikke desto mindre relevant at have in mente, hvor gode data man kan levere, hvis man som beslutningstager overvejer at implementere en prædiktionsmodel.

Tabel 5.4: De mest indflydelsesrige variable i model senere i studieforløbet

Variabel	**Importance**
1. Andel beståede eksamener (på Metropol)	0,326
2. Gns. holdstørrelse	0,232
3. Karaktergennemsnit (på Metropol)	0,212
4. Udvikling i karakterer (på Metropol)	0,200
5. Gns. kønsfordeling på hold	0,196
6. Karaktersnit fra ungdomsudd. fratrukket adgangskvotient	0,175
7. Tid på studiet før første bestået eksamen (på Metropol)	0,167
8. Gns. alder på hold	0,153
9. Karaktersnit fra ungdomsuddannelse	0,138
10. Karaktersnit fra ungdomsudd. fratrukket årgangens gns.	0,137
11. Alder	0,136
12. Antal år fra eksamen på ungdomsuddannelse	0,111
13. Adgangskvotient på uddannelsen	0,070
14. Antal prioriteter ved ansøgning	0,068
15. Andelen af praktik på første år	0,057
16. Længde af første praktik	0,040

5.4 Anvendelse af modellen i praksis

I dette afsnit vil vi opstille et framework for, hvordan en prædiktionsmodel som den nærværende kan tages i anvendelse. Det gør vi ved at sætte tal på og formalisere de praktiske omstændigheder, som udgør Metropols kontekst. På denne måde viser vi, hvordan frafaldsmodellen kan anvendes af en beslutningstager, som ønsker at målrette tiltag mod uddannelsesfrafald.

5.4.1 Tærskelværdiens betydning

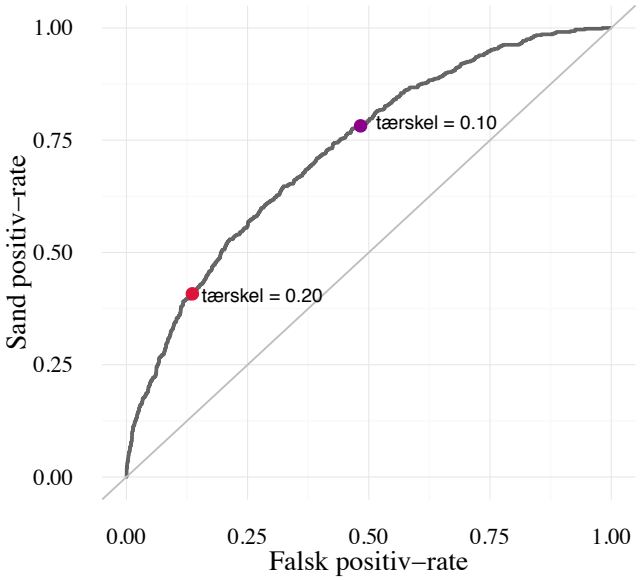
Indtil nu har vi evalueret modellerne på baggrund af ROC-kurver og AUC. Målene viser som bekendt et samlet tradeoff mellem sand positiv-raten og falsk positiv-raten for en given model på tværs af tærskelværdier. Når prædiktionsmodellen skal anvendes i praksis, er det imidlertid nødvendigt at fastsætte én tærskelværdi. Eksempelvis kunne beslutningen være, at alle studerende med en prædikteret sandsynlighed for frafald på 30 pct. eller derover skal klassificeres som frafaldstruede. I denne gruppe ville en andel af de studerende rent faktisk frafalde (sande positive), mens andre ikke ville (falske positive). Vi er imidlertid ude af stand til at skille de to grupper ad ex ante. Det har den konsekvens, at vi må rette et tiltag mod alle, vi har prædikteret som frafaldstruede, vel vidende at en andel af dem aldrig vil falde fra, uanset om de får tiltaget eller ej.

Når en tærskelværdi fastsættes, er konteksten afgørende, fordi en given tærskelværdi afspejler et givent punkt på modellens ROC-kurve, hvilket igen afspejler et givent tradeoff mellem de fire prædiktionstyper (SP, FP, SN, FN). Hvordan tærsklen fastsættes, hænger derfor uløseligt sammen med, hvor vigtigt det i konteksten er at få identificeret sande positive, vis-a-vis hvor problematisk det er, at præcisionen falder, fordi der kommer flere falske positive i gruppen.

Som eksempel kan man forestille sig, at det i casen Metropol ikke er altafgørende, om et tiltag mod frafald (såsom et tilbud om ekstra studievejledning) bliver rettet mod flere end dem, der egentlig har behov for det. Det vil være en udgift, men det er mindre problematisk, end hvis casen fx havde været fra sundhedsområdet, og tiltaget, der blev målrettet, var en medicin med store bivirkninger. Her ville høj præcision med et lavt antal falske positive, dvs. et lavt antal unødigt behandlede patienter, formentlig være højt prioriteret. Det ville endvidere kunne afhænge af sygdommens grad af alvor – en akut livstruende sygdom vil formentlig tilsige prioritering af en høj sand positiv-rate, selvom flere falske positive dermed også modtager behandlingen. Hvordan en models tærskelværdi fastsættes, er således stærkt afhængig af den kontekst, som modellen implementeres i.

I nedenstående figur 5.12 gengives ROC-kurven for den bedste model fra analysen, der er baseret på GBT-algoritmen og forudsiger frafald inden for et helt studieår (se evt. tabel 5.1 på side 64).

Figur 5.12: ROC-kurve for 1 års-frafaldsmodellen med to potentielle tærskelværdier

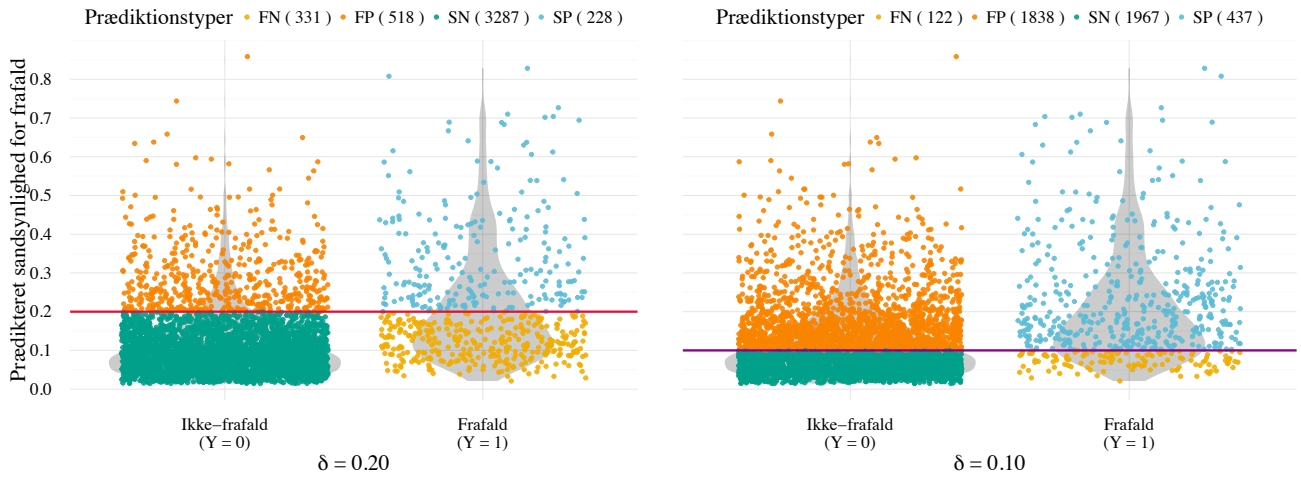


I figuren er angivet to punkter, der afspejler to potentielle tærskelværdier hhv. δ = 0,20 og δ = 0,10. Tærsklerne tolkes sådan, at alle studerende, som har en prædikteret sandsynlighed for at frafalde på hhv. mindst 20 pct. og 10 pct. bliver klassificeret som frafaldstruede. Når tærsklen sænkes fra 20 pct. til 10 pct., kan klassifikationen siges at blive mere lempelig, idet der herved "skal mindre til" for at blive klassificeret som et kommende frafald. Rent grafisk vil en lavere og mere lempelig tærskelværdi betyde, at punktet for tærskelværdien vil bevæge sig op ad ROC-kurven mod punktet (1,1), mens en højere og mindre lempelig tærskelværdi omvendt vil bevæge punktet mod origo i (0,0).

Som det ses er hældningen for ROC-kurven forskellig ved de to tærskelværdier. Omkring den højeste tærskelværdi, δ = 0,20, er forholdet mellem SPR og FPR større end ved punktet for den laveste tærskelværdi, δ = 0,10. Det svarer til en højere præcision. Til gengæld er SPR naturligt nok højere for den lave tærskelværdi – her klassificerer vi en større andel af de positive cases som positive. Sammenhængen mellem de fire prædiktionstyper ved ændringer i tærskelværdien kan grafisk illustreres som i figur 5.13.

I hver side af figuren er der to søjler med prikker. Prikkerne i højre søjle er studerende, der frafalder, og prikkerne i venstre søjle er studerende, der ikke gør. I venstre side af figuren er tærskelværdien sat til 0,20. Det ses, at der er forholdsvist mange studerende, der falder fra, som ligger under tærskelværdien (de gule prikker). Når tærsklen sænkes til 0,10 som i højre side af figuren, ses det, at næsten alle prikker i højre søjle ligger over tærskelværdien og altså korrekt forudsiges at ville falde fra. Til gengæld er der også markant flere prikker i venstre søjle, der nu ligger over tærskelværdien og fejlagtigt forudsiges at ville falde fra (de orange prikker).

Figur 5.13: Ændringer af tærskelværdien og antallet af prædiktionstyper



Vi opsamler tradeoff'et mellem prædiktionstyperne for de to tærskelværdier i confusion-matricerne i tabel 5.5. Tabellens farver modsvarer figurens. Som allerede anført stiger antallet af sande positive, når tærskelværdien sænkes. Det samme gør antallet af falske positive, endda ganske markant. Selvom antallet af sande positive stiger med næsten 100 pct. fra 228 til 437, stiger antallet af falske positive med mere end 300 pct. fra 518 til 1838.

Tabel 5.5: Confusion-matricer for to forskellige tærskelværdier

	$\delta = \mathbf{0.20}$				$\delta = \mathbf{0.10}$			
	$\hat{y}_i = 1$	$\hat{y}_i = 0$			$\hat{y}_i = 1$	$\hat{y}_i = 0$	
$y_i = 1$	**228**	**331**	SPR = 0,408	$y_i = 1$	**437**	**122**	SPR = 0,782
$y_i = 0$	**518**	**3287**	FPR = 0,136	$y_i = 0$	**1838**	**1967**	FPR = 0,483

Antallet af de respektive prædiktionstyper har stor betydning for målretningens potentiale. Hvis Metropol eksempelvis sætter tærskelværdien til 0,20 som ovenfor og retter et tiltag mod de studerende, der her klassificeres som frafaldstruede, så vil Metropol rette et tiltag mod 746 studerende. Hvis tærsklen sænkes til 0,10 skal Metropol i stedet rette tiltag mod 2275 studerende – altså mere end 3 gange så mange. Det kan gøre en stor forskel i praksis afhængigt af tiltaget.

5.4.2 Et framework for anvendelse af prædiktionsmodellen

Lad os nu sætte os i beslutningstagerens sted. Hvordan vil vi da bruge modellen til at målrette tiltag? Vi kan eksempelvis forestille os, at vi har tre mulige tiltag, som vi kan rette mod frafaldstruede studerende: en mail om gode studievaner, et visitationsmøde med studievejledningen og et tilbud om et personligt mentorforløb. Hvor mange studerende skal vi rette tiltagene imod – hvordan fastsætter vi en tærskelværdi? I det følgende viser vi, hvordan problemet kan gribes an.

Vi kan som beslutningstagere tage udgangspunkt i den forventede fortjeneste og omkostning ved at målrette et tiltag. Det er klart, at omkostningerne og fortjenesterne er en simplificeret tilgang til at beskrive en policy-kontekst. Simplificeringen tjener dog det formål at kunne illustrere et framework til, hvordan en tærskelværdi kan sættes og tiltag målrettes i praksis. Herunder opstiller vi derfor et eksempel, hvor vi illustrerer tankegangen ved at sætte den forventede fortjeneste og omkostning på formel:

Forventet fortjeneste:

$$F = SP \times \alpha \times indtægtssats$$

Forventet omkostning:

$$O = (SP + FP) \times udgiftssats$$

Her angiver SP og FP henholdsvis antallet af sande og falske positive. α angiver effekten af et givent tiltag, dvs. andelen af kommende frafald (sande positive), som tiltaget kan forhindre. Det kan i øvrigt bemærkes, at α er den *resulterende* effekt, dvs. andelen, som rent faktisk bliver fastholdt. α svarer derfor til *intent to treat effect* (ITT) snarere end *complier average causal effect* (CACE) eller *average treatment effect* (ATE), som vi kender det fra kausal estimation (Gerber & Green 2012: 141–143). Det er væsentligt at holde in mente, når α erstattes med en given, empirisk estimeret effekt.

Indtægtssatsen er indtægterne forbundet med at fastholde én af de sande positive, hvilket vi for simpelhedens skyld kan antage er den gennemsnitlige taxametersats for et gennemført studieforløb på Metropol. Det er selvsagt forsimplet, fordi det kan tænkes, at nogle frafaldstruede studerende ganske vist forhindres i at droppe ud inden for det første studieår, men dropper ud senere i deres studieforløb. Her vil kun en del af fortjenesten blive realiseret, idet taxametersatsen udbetales i rater under uddannelsen. Derudover antages det generelt i opstillingen her, at Metropols marginale udgifter for en studerende er 0 kr. Det virker måske som en rimelig antagelse for en enkelt eller nogle få studerende, men hvis frafaldet fx drejer sig om flere hundrede personer, vil det ganske givet være en noget grov forsimpling. Her må flere frafald – udover lavere taxameterindtægter – også forventes at resultere i lavere udgifter forbundet med at udbyde uddannelserne.

Udgiftssatsen er udgifterne forbundet med at rette tiltag mod én af de studerende, der klassificeres som frafaldstruet. Det kan vi antage er de direkte udgifter ved tiltaget, hvilket fx vil sige omkostningerne ved at tilbyde én studerende et mentorforløb. Det er igen forsimplet, bl.a. fordi udgiften per tiltag, fx et mentorforløb, kan tænkes at være lavere, hvis det tilbydes til flere.

Med disse forbehold kan vi opstille tabel 5.6 på den følgende side, der viser tre potentielle tiltag, som vi kan rette mod de studerende, vi prædikterer som frafaldstruede.

Tabel 5.6: Effekt og omkostninger ved tre potentielle tiltag mod frafald

	Mail	**Visitationsmøde**	**Mentorforløb**
Omkostning i kr.	4200 kr. i alt	420 kr. per tilbud	4200 kr. per tilbud
Omkostning i timer	10 timer i alt	1 time per tilbud	10 timer per tilbud
Antaget gennemsnitlig fastholdelseseffekt	0,2 %	2 %	12,5 %

Som det ses af tabellen, er omkostningsprofilerne forskellige for de tre potentielle tiltag. Her antager vi, at omkostningerne forbundet med at sende en mail ud om gode studievaner er faste omkostninger. Det koster altså ikke mere at sende mailen ud til flere studerende. Anderledes antager vi, at omkostningerne forbundet med at tilbyde visitationsmøder og mentorforløb til de studerende er variable. Omkostningerne er desuden forskellige, fordi vi antager, at de kræver et forskelligt antal arbejdstimer. De specifikke tal, som vi har angivet i tabellen, har vi alene opstillet som regneeksempel, og de bør derfor tages med forbehold. Omkostningerne i kroner er baseret på antallet af timer for tiltaget ganget med Metropols omkostninger per studievejleder, der er 420 kr. i timen (løn, kontorfaciliteter, pension, mv.). Det er den samme timesats, som Metropol benytter for alle akademiske ansatte[1].

Derudover er de antagede fastholdelseseffekter også forskellige tiltagene imellem. En effekt på 2 pct. skal tolkes sådan, at når tiltaget målrettes en gruppe af studerende, der er blevet klassificeret som frafaldstruede, så bliver 2 pct. af frafaldene forhindret blandt de sande positive i denne gruppe. Det er vigtigt at understrege, at de antagede effekter ikke er estimeret empirisk – det er alene tænkte effekter, som vi opfinder og benytter i regneeksemplet her for at illustrere logikken. Antagelserne om effekternes størrelse er dog ikke kritiske antagelser for selve logikken – hvis senere forskning viser, at effekterne er nogle andre, vil det blot forskubbe kurverne op og ned i de to figurer 5.14 og 5.15 på side 89, som vi analyserer om lidt. Når kurverne forskubbes, har det dog betydning for det faktiske potentiale ved målretningen af tiltaget, og derfor har de faktiske værdier i analysen nedenfor en nævneværdig usikkerhed og bør alene tages som illustration. Vi vender tilbage til en bredere refleksion over behovet for effektestimater i diskussionen.

Slutteligt antager vi, at den forventede fortjeneste ved at forhindre ét frafald inden for det første år er den samlede, gennemsnitlige taxameterafregning for ét helt gennemført studieforløb på Metropol, hvilket er 147.486 kr. (Uddannelses- og Forskningsministeriet 2017). Den sats er som nævnt også en forsimpling, bl.a. fordi den studerende kan droppe

[1]Vi har indhentet satsen ved forespørgsel per mail til Kåre Degn, konsulent ved Metropol.

ud senere i studieforløbet, og fordi vi ser bort fra afledte effekter, såsom hvis en større fastholdelsesgrad er selvforstærkende.

Vi kan nu som beslutningstager i studievejledningen på Metropol – med forbeholdene in mente – komme med et overslag på den forventede nettofortjeneste ved et givent tiltag for en given tærskelværdi. Vi kan eksempelvis tage udgangspunkt i tiltaget *visitationsmøde* og en tærskelværdi på 0,20. Her vil beregningen se ud som følger:

Forventet nettofortjeneste:

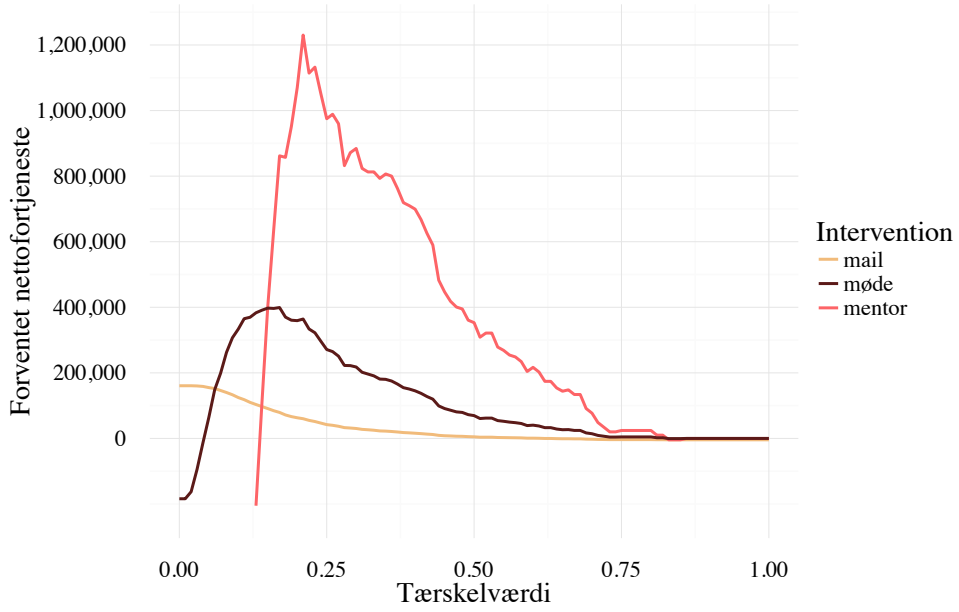
$$\begin{aligned} F - O &= SP \times \alpha \times indtægtssats - (SP + FP) \times udgiftssats \\ &= 228 \times 0,02 \times 147.486 - (228 + 518) \times 420 \\ &= 359.216,16 \end{aligned}$$

Det følger direkte af opstillingen, at den forventede nettofortjeneste alt andet lige bliver større, hvis tiltagets effekt er større, eller hvis indtægtssatsen bliver større. Omvendt bliver nettofortjenesten mindre, når tiltaget bliver dyrere. Derudover sker der to ting, når antallet af sande positive bliver større, fx fordi tærskelværdien sænkes. For det første vokser den forventede fortjeneste, fordi flere kan forhindres i at falde fra. For det andet vokser den forventede omkostning, fordi flere studerende tilbydes tiltaget.

Hvilken af disse to modsatrettede effekter, den voksende fortjeneste eller den voksende omkostning, der dominerer, vil afhænge af flere forhold. Først og fremmest afhænger det af prædiktionsmodellens performance. Jo mere præcis modellen er, jo bedre vil den være til at adskille positive og negative cases, hvormed FPR stiger relativt langsommere end SPR, når tærsklen sænkes. For den enkelte model vil det også have betydning, hvor på ROC-kurven, vi befinder os. Typisk er kurven konkav og stiger stejlt på det første stykke, hvorefter den flader ud. Det betyder, at det er forholdsvist nemt at adskille de første positive cases fra de negative, mens det vil være meget svært at få de sidste positive cases klassificeret som positive, uden at vi fejlagtigt klassificerer en masse negative cases som positive i samme ombæring.

Med udgangspunkt i denne logik kan vi som i figur 5.14 og figur 5.15 på de følgende sider illustrere sammenhængen mellem den forventede nettofortjeneste og de mulige tærskelværdier. Det kan vi som beslutningstagere i den skitserede kontekst anvende som grundlag for at fastsætte en tærskelværdi. De to figurer er baseret på henholdsvis analysens bedste model, GBT-modellen, og den logistiske regressionsmodel, der her tjener som en baseline fra politologens "klassiske" værktøjskasse.

Figur 5.14: Sammenhæng mellem tærskelværdi og forventet nettofortjeneste ved tre forskellige tiltag (GBT-model)

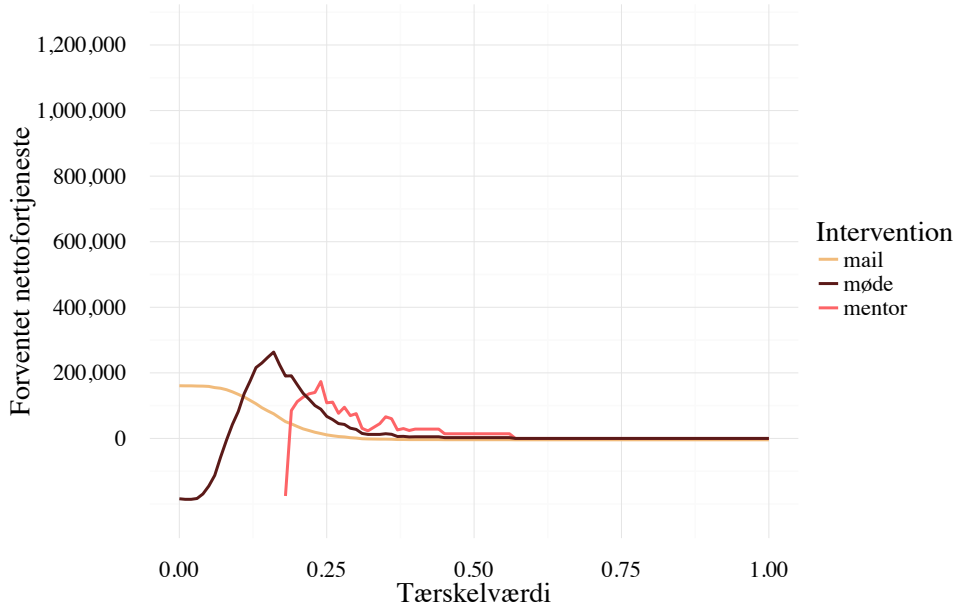


Hvis vi tager udgangspunkt i GBT-modellen i figur 5.14 og i tiltaget *visitationsmøde*, ses det, at den forventede nettofortjeneste topper omkring en tærskelværdi på 0,17. Hvis formålet er at maksimere Metropols profit, og vi som beslutningstagere skal afgøre, hvor mange studerende der skal tilbydes et møde med studievejledningen, vil det optimale valg være at tilbyde mødet til alle studerende, der har en prædikteret sandsynlighed for at falde fra på 17 pct. eller derover. Logikken er den samme for grafernes toppunkter for de øvrige tiltag, også for målretningen baseret på den logistiske regressionsmodel i figur 5.15.

Det ses, at toppunktet for tiltaget *mentorforløb* ligger til højre for toppunktet for tiltaget *visitationsmøde*. Det gælder for begge modeller i hhv. figur 5.14 og figur 5.15. Det betyder, at det for tiltaget *mentorforløb* er optimalt at sætte tærsklen højere end for tiltaget *visitationsmøde*, hvis formålet er at maksimere profitten. Den højere tærskelværdi vil betyde en mere præcis prædiktion. Det er ikke overraskende, at dette dyrere tiltag stiller strengere krav til modellens præcision – hvis modellen er for upræcis kan det ikke betale sig at tilbyde dette forholdsvist dyre forløb. Den enkelte studerende skal have en højere forudsagt frafaldsrisiko for at komme i betragtning til et dyrt mentorforløb.

Det ses endvidere af figurerne, at den forventede fortjeneste ved tiltaget *mail* kun falder, i takt med at tærskelværdien hæves. Det skyldes antagelsen om, at den marginale

Figur 5.15: Sammenhæng mellem tærskelværdi og forventet nettofortjeneste ved tre forskellige tiltag (logit-model)



omkostning ved en ekstra mail er 0. Det vil således være optimalt at rette dette tiltag mod samtlige studerende.

Tærskelværdien kan også fastsættes efter andre hensyn end maksimering af nettofortjenesten. Det kan fx tænkes, at vores primære hensyn er at rette tiltaget mod så mange frafaldstruede studerende som muligt for at maksimere fastholdelsen. I det tilfælde er vi som beslutningstagere interesserede i at sætte tærsklen så lavt som muligt. Her kan vi sætte tærsklen til det punkt, hvor grafen for det givne tiltag skærer x-aksen (gælder kun for *visitationsmøde* og *mentorforløb*)[2]. Her går den forventede fortjeneste ud med den forventede omkostning, og det er således en omkostningsneutral tærskel, hvor nettofortjenesten forventes at være 0. Hvis vi tager udgangspunkt i tiltaget *mentorforløb* er den omkostningsneutrale tærskel 0,13 i figur 5.14 på forrige side. Når tærsklen sættes til 0,13, er sand positiv-raten 65,3 pct. Det betyder, at 65,3 pct. af de faktisk kommende frafald indgår i denne gruppe af studerende, som vi klassificerer som frafaldstruede. Altså kan vi som beslutningstagere rette tiltaget *mentorforløb* mod op til 65,3 pct. af alle de reelt frafaldstruede, uden at der forventes at være en ekstra omkostning. Disse 65,3 pct. omtaler vi som den maksimale SPR nedenfor. Det er nemt at forestille sig en policy-kontekst, hvor formålet er at maksimere SPR, hvis fx institutionen Metropol afsætter eller modtager et givent beløb til at øge deres fastholdelse mest muligt som led i en national fastholdelsespolitik.

[2]Teknisk set tangerer graferne x-aksen i yderligere ét punkt, når tærsklen sættes til 1.

Samtidig kan SPR omsættes til et faktisk antal studerende, som vi kan fastholde ved målretningen. Det gøres ved at gange SPR med den antagede effekt af tiltaget og antallet af faktiske positive cases i datasættet.

I nedenstående tabel 5.7 opsamler vi sammenhængene mellem potentielle fortjenester og omkostninger i casen Metropol for de tre tiltag – både uden målretning og med målretning baseret på henholdsvis GBT-modellen og den logistiske regressionsmodel. Det kan bemærkes, at tallene i tabellen er baseret på et datasæt (testsæt) bestående af i alt 4364 studerende på Metropol, hvoraf 559 frafalder inden for et år. Til sammenligning starter der godt og vel 3.000 nye studerende på Metropol hvert studieår.

Tabel 5.7: Gevinster og omkostninger ved tre tiltag målrettet med forskellige modeller

		Mail	**Møde**	**Mentor**
Ingen model	Nettofortjeneste u. målretning	161.000	-184.000	-8.023.000
Logit-model	Maks. nettofortjeneste	161.000	263.000	174.000
	Maks. SPR	100 %	92,7 %	29,5 %
	Maks. antal fastholdte stud.	1	10	21
GBT-model	Maks. nettofortjeneste	161.000	399.000	1.230.000
	Maks. SPR	100 %	98,7 %	65,3 %
	Maks. antal fastholdte stud.	1	11	46

Af tabellen fremgår det, at der i casen Metropol ikke er et potentiale ved at bruge en prædiktionsmodel til at målrette udsendelsen af mails. Det kan bedst betale sig blot at rette tiltaget mod samtlige studerende, hvilket gør en prædiktionsmodel overflødig.

Anderledes forholder det sig med møder med studievejledningen og mentorforløb. Selv for et relativt billigt tiltag som et møde vil det fra Metropols synspunkt, rent økonomisk, ikke kunne betale sig at tilbyde tiltaget til alle studerende under de simplificede betingelser, som vi har opstillet. Det samme gør sig gældende for det mere omfangsrige tilbud om et mentorforløb. Fordi mentorforløbet er markant dyrere end mødet, stiller mentorforløbet endog større krav til præcis målretning. Det ses for det første ved, at den maksimale SPR er lavest for tiltaget mentorforløb. Det ses for det andet ved den store forskel på maksimal nettofortjeneste ved tiltaget mentorforløb, når målretningen baseres på GBT-modellen frem for logistisk regression.

Som vi så tidligere i analysen havde den logistiske regressionsmodel en AUC på 0,646, mens GBT-modellens AUC var 0,727. Den forbedrede prædiktionsevne kommer tydeligt til udslag i tabellen ovenfor. Der er ganske vist kun en lille gevinst at hente, hvis det er

tiltaget *visitationsmøde*, som ønskes målrettet. Med det dyrere tiltag *mentorforløb*, er det derimod forventningen, at GBT-modellen kan fastholde dobbelt så mange studerende som baseline-modellen. Samlet kan GBT-modellen fastholde 46 studerende, som ellers ville være droppet ud, hvis tiltaget og målretningen ikke var blevet iværksat. For dette tiltag har det altså i praksis en ganske stor betydning, om tiltag målrettes med en model, som har en AUC på 0,646 eller 0,727. Derfor har det også betydning i praksis, hvilken algoritme man anvender.

Set i lyset af de store forbehold ved skitseringen af konteksten skal de absolutte tal ikke tages for pålydende – tallene bygger som nævnt på en række meget simplificerede antagelser, og tallene er meget sensitive over for specifikationen af tiltagenes antagede effekt, omkostninger med videre. De faktiske tal kan derfor ikke ligge til grund for en konklusionen om, hvor stor eller lille en nettofortjeneste Metropol eller en anden uddannelsesinstitution kan forvente, hvis de ruller et mentorforløb eller et andet tiltag ud til deres studerende på baggrund af en GBT-baseret prædiktionsmodel.

Sammenhængen mellem tallene og den overordnede logik kan til gengæld godt ligge til grund for en samlet konklusion om, at der er en gevinst ved at målrette tiltag mod uddannelsesfrafald, og at gevinsten bliver større jo bedre en prædiktionsmodel, der anvendes. Alt efter hvordan tærskelværdien fastsættes, kan modellen optimeres til enten bedst mulig allokering af ressourcer eller flest muligt gennemførte uddannelser.

Kapitel 6

Diskussion

Med udgangspunkt i casen Metropol har vi udviklet en prædiktionsmodel og vist, hvordan den kan bruges til at målrette tiltag mod uddannelsesfrafald ved at inddrage viden om den givne policy-kontekst. Ud fra rent metodiske kriterier rummer maskinlæring altså et potentiale til målretning af tiltag mod frafald. Vi vil bruge det første afsnit i diskussionen, afsnit 6.1, på at opsamle analysens resultater og implikationerne heraf samt forholde os til analysens interne og eksterne validitet. Vi vil diskutere, i hvilket omfang vores resultater er generaliserbare til andre cases.

Indtil nu har vi primært behandlet vores problemstilling med metodiske briller og vurderet modellen på dens egne præmisser. I kapitlets resterende tre afsnit kommer der imidlertid grus i maskineriet. Der viser sig nemlig en række dilemmaer, når modellen skal anvendes i praksis. Den tekniske vurdering kan ikke stå alene, og vi vil derfor inddrage perspektiver fra litteratur-reviewet, som på forskellig vis problematiserer anvendelsen af maskinlæring i samfundsvidenskaben. Vi vil diskutere disse perspektiver med udgangspunkt i vores eget case, men de vedrører generelle aspekter ved maskinlæring som tilgang og vil således kunne sættes i relation til målretning af tiltag med prædiktionsmodeller mere generelt, også uden for casen Metropol og uddannelsesområdet.

Vi har ordnet refleksionerne i tre afsnit, der hænger sammen med forskellige aspekter af maskinlæring som tilgang til at målrette policy-tiltag. Sammenhængen fremgår af tabel 6.1 på den følgende side. Prædiktionsmodeller løber ind i udfordringer, når de skal anvendes på sociale fænomener, og i afsnit 6.2 diskuterer vi de metodologiske problematikker, som hidrører fra brugen af prædiktion frem for kausalestimation. De sidste to afsnit behandler henholdsvis epistemiske og etiske aspekter af vores tilgang.

Afsnit 6.3 indeholder epistemiske overvejelser omkring algoritmisk vidensproduktion, herunder manglende transparens og modellernes usikkerhed. I afsnit 6.4 følger en række normative, etiske refleksioner omkring de effekter, det kan have at anvende modellen til målretning af policy-tiltag.

Tabel 6.1: Problematikker knyttet til maskinlæring i samfundsvidenskaben

Aspekt af maskinlæring som tilgang	Afledte problematikker
Anvendelse af **prædiktion** i samfundsvidenskaben	Metodologiske
Anvendelse af **algoritmer** til at omsætte data til viden	Epistemiske
Anvendelse af denne viden til **målretning** af policy-tiltag	Etiske

6.1 Validiteten af analysens resultater

Vi vil i dette afsnit samle op på og diskutere analysens implikationer og validitet. Vi ordner diskussionen efter en distinktion mellem intern og ekstern validitet. I det første afsnit om intern validitet diskuterer vi analysens styrker og svagheder og således validiteten af vores konklusion om, at maskinlæring har potentiale til at målrette tiltag mod uddannelsesfrafald i casen Metropol. I det andet afsnit om ekstern validitet diskuterer vi, i hvilket omfang dette fund og analysens øvrige resultater kan generaliseres ud over casen Metropol.

6.1.1 Analysens interne validitet

Hovedfundet i analysen er, at maskinlæring har potentiale til at målrette tiltag mod uddannelsesfrafald i casen Metropol. Vi drager denne konklusion på baggrund af frafaldsmodellens prædiktions-performance, udtrykt ved en AUC på 0,727. Modellens performance er tilstrækkeligt høj til, at vi kan anvende den i et simpelt framework til målretning, der inddrager omkostninger og gevinster i den konkrete kontekst. Vi kan anvende frameworket til at målrette tiltagene sådan, at vi maksimerer enten antallet af fastholdte studerende eller den økonomiske fortjeneste ved øget fastholdelse. Kan vi regne med validiteten af disse resultater?

Spørgsmål om en analyses interne validitet handler i estimationssammenhæng om, hvorvidt vi håndterer vores case på en sådan måde, at kausalslutningen er troværdig (Andersen et al. 2010: 103–104). Validiteten er dermed tæt knyttet til databehandlingen: Burde vi have specificeret vores model anderledes? Burde vi have imputeret vores manglende data? Har vi udeladt relevante variable, vi burde have inkluderet? Imidlertid er betydningen af disse overvejelser anderledes ved prædiktion end ved estimation. I

prædiktionssammenhæng drager vi ikke kausale slutninger og er derfor ikke bekymrede for modellens korrekthed på samme måde som ved estimation. Vores databehandling har ikke til formål at sikre unbiased resultater, men alene at øge modellens prædiktions-performance. Hvis vi fx bekymrer os om bias fra udeladte variable, er det altså ikke, fordi det risikerer at underminere gyldigheden af vores resultater, men fordi det fører til dårligere prædiktions-performance. Yderligere og grundigere databehandling kunne alene have hævet modellens performance. Vi anskuer med andre ord den bedste models AUC som et *lower bound*, dvs. et konservativt estimat af potentialet til målretning i casen Metropol.

En analyses validitet afhænger også af selve datagrundlaget, hvor vi typisk bekymrer os om målefejl. Systematiske målefejl kan give bias, tilfældige målefejl på den uafhængige variabel kan give bias og øget varians, og tilfældige målefejl på den afhængige variabel kan give øget varians (King et al. 1994: 151–168). I vores case kunne en kilde til systematiske målefejl fx være, hvis varierende administrative praksisser resulterede i systematiske forskelle i dataindsamlingen studierne imellem. Det kunne fx skyldes, at sekretæren med ansvar for sygeplejerskernes orlov og holdsætning registrerer anderledes end sine kollegaer på resten af uddannelserne. En kilde til tilfældige målefejl kunne være, hvis alle sektretærer af og til rapporterer orlov og holdsætning forkert. Sådanne målefejl vil give bias i en estimationssammenhæng (King et al. 1994: 151–168). Det er en udbredt opfattelse, at sådanne målefejl ikke er et problem i prædiktion (Fuller 2009: 74). Denne påstand er dog betinget af, at observationerne i træningssættet og testsættet begge er et random sample af observationer fra den samme underliggende, datagenererende proces (Fuller 2009: 74–75). Det virker som en rimelig antagelse i vores analyse af casen Metropol, fordi vores træningssæt er et tilfældigt sample fra det fulde datasæt, og testsættet består af det resterende data. Vi kan forstå dette sådan, at målefejlene er tilfældigt fordelt mellem trænings- og testsættet, og outcome derfor genereres ens i de to.

Der er dog en væsentlig indvending imod denne antagelse, når frafaldsmodellen tages i anvendelse i praksis. Her hviler troværdigheden af frafaldsmodellens prædiktioner på, at vores testsæt er et retvisende billede på det "virkelige" testsæt, som vi interesserer os for, dvs. de fremtidige årgange af studerende. Her løber vi imod problemer, hvis der er systematiske målefejl i vores nuværende data, som ikke genfindes i det virkelige testsæt. Lad os fx forestille os, at førnævnte sekretær rettede sin registreringspraksis ind, sådan at den systematiske målefejl forsvandt. I så fald er vi ikke garanteret, at modellens AUC i vores testsæt er en god tilnærmelse af den faktiske AUC, når vi anvender modellen på de kommende årgange af studerende. Vores analyse hviler med andre ord på en implicit antagelse om, at miljøet er stabilt. Vi kan derfor spørge, om det virker plausibelt, at vores testsæt i analysen er en god tilnærmelse af det virkelige

testsæt. Vi har ingen anledning til at tro, at eventuelle målefejl i vores testsæt ikke også skulle gøre sig gældende i det virkelige testsæt. Det kan dog ændre sig over tid, eksempelvis når personalet i Metropols administration skiftes ud eller administrative praksisser ændres. Det er dog ikke muligt at efterprøve, om vores testsæt modsvarer det virkelige testsæt, da sidstnævnte i sagens natur er utilgængeligt.

Ud over målefejl kan der være andre årsager til, at mønstrene i det virkelige testsæt vil være forskellige fra mønstrene i vores testsæt, eksempelvis grundet ændringer i studieordninger, SU-regler eller reformer af national uddannelsespolitik. Derfor forventer vi ikke, at vores testsæt er et 1:1-billede på det virkelige testsæt. Dén indvending er imidlertid ikke særlig for lige vores analyse eller for prædiktion, men går under betegnelsen Lucas-kritikken og gælder mere generelt: statistiske modeller er baseret på historiske data og afspejler derfor historiske mønstre, der ikke nødvendigvis vil gælde fremadrettet (Lucas 1976). Denne indvending vender vi tilbage til i vores metodologiske diskussion. Vi forventer dog, at sådanne ændringer i frafaldsmønstre vil være træge, og at vores testsæt derfor i tilstrækkelig grad kan antages at modsvare det virkelige testsæt.

6.1.2 Analysens eksterne validitet

Samlende finder vi ikke grund til at underkende validiteten af vores konklusion om maskinlærings potentiale til målretning på Metropol. Det næste oplagte spørgsmål er, i hvilket omfang fundet rækker ud over casen Metropol. I hvilken grad er vores analyse generaliserbar? Det potentiale, vi finder i casen Metropol, afhænger både af casens datagrundlag, de anvendte algoritmer, mønstrene i outcome og den kontekst, som modellen udvikles og anvendes i. Vi vil derfor strukturere diskussionen af analysens eksterne validitet efter disse fire punkter.

Datas beskaffenhed

Vi finder i analysen, at frafaldsmodellens performance og dermed potentialet til målretning i vidt omfang afhænger af tilgængeligheden og kvaliteten af data. Jo flere forskellige oplysninger vi har om vores observationer jo bedre. I vores case har vi adgang til relativt mange og forskelligartede informationer om de studerende, som allerede opsamles i Metropols database. Spørgsmålet er, om det gør vores case til noget særligt, eller om vi vil kunne forvente et lige så godt datagrundlag uden for casen Metropol? Det er vores forventning, at vi kan genfinde samme datakvalitet på resten af uddannelsesområdet. Metropol deler administrativt system med andre professionshøjskoler og får

mange af sine data fra fælles offentlige databaser. Som vi så i litteraturgennemgangen, findes der endnu bedre data på området for ungdomsuddannelser (Şara 2014). De gode data på området stiller uddannelsesinstitutionerne godt i forhold til at finde effektive prædiktorer i deres datasæt. I denne henseende kan en uddannelsesinstitution måske ligefrem betragtes som en *most likely*-case sammenlignet med institutioner på andre policy-områder (Flyvbjerg 2006: 229–231). Hvis ikke vi havde fundet et potentiale for maskinlæring på uddannelsesområdet, havde det derfor varslet dårligt for andre policy-områder med mere sparsomme data.

Uddannelsesområdets gode data er dog ikke unikt i en dansk sammenhæng. Vi kan selvfølgelig ikke direkte udtale os om datagrundlaget på andre policy-områder på baggrund af vores case. Det er dog kendt, at der i Danmark er relativt gode og vidtfavnende data til rådighed om landets borgere, hvilket alt andet lige øger sandsynligheden for et potentiale til målretning hos offentlige serviceudbydere generelt. Generaliseringen til andre policy-områder er dog betinget af karakteren af data, da særligt personfølsomme data i praksis kan være utilgængelige for den enkelte myndighed (Retsinformation 2000). Derfor er maskinlærings potentiale til målretning også betinget af følsomheden af data på det enkelte område.

Vedrørende data finder vi slutteligt, at logdata fra de studerendes brug af intranettet har en meget begrænset prædiktiv værdi i vores case. Det er et overraskende resultat, fordi meget hype netop samler sig om potentialet af denne type data, der er forholdsvist ny i samfundsvidenskaben. Vi kan dog ikke generalisere bredt til andre policy-områder, at denne type data ikke har noget potentiale. Vores analyse peger blot på, at realiseringen af et eventuelt potentiale som minimum vil kræve, at data er indsamlet og bevaret for en længere tidsperiode. Det er en lavpraktisk indsigt omkring denne type af observationsdata, at de ikke nødvendigvis er tilgængelige for offentlige myndigheder i det omfang, man kunne forledes til at tro ud fra hypen omkring big data.

Valg af algoritme

Vi viser i analysen, at potentialet til målretning afhænger af den anvendte algoritmes prædiktions-performance. Vi finder endvidere, at algoritmerne performer forskelligt i vores case, og at deres potentiale til målretning derfor er forskelligt. Vi forventer, at det samme vil gøre sig gældende på andre policy-områder helt generelt.

De træbaserede algoritmer klarer sig væsentligt bedre end den logistiske regressionsmodel, som er vi er mere vant med i samfundsvidenskaben. Der kan altså være ræson i at afprøve alternative algoritmer fra datalogien. Samtidig vil vi dog understrege, at

vi i specialet her behandler den logistiske regressionsmodel som en baseline for vores resultater. Vi regulariserer den således ikke, hvilket ganske givet ville have øget dens performance. Andre casestudier har fundet, at regulariseret logistisk regression godt kan være med i opløbet om den højste forudsigelseskraft (Aulck et al. 2016; Harmsen & Enggaard 2016).

Træmodellernes fordel er dog, at de kan modellere komplekse sammenhænge mere frit og selv "opdage" interaktioner i data. Det må forventes at være en fordel ikke bare på uddannelsesområdet, men på alle policy-områder, hvor outcome bestemmes af et komplekst samspil af sociale faktorer. Jo flere faktorer, der ventes at indvirke på et outcome, og jo mere komplekse sammenhængene er, jo større forbedring vil vi også forvente fra de træbaserede modeller sammenlignet med den logistiske regressionsmodel. Omvendt gælder det i en policy-kontekst med relativt få faktorer, eller hvor den teoretiske årsagssammenhæng er veletableret, at en logistisk regressionsmodel måske klarer sig lige så godt som de mere avancerede algoritmer. I de tilfælde kan en GBT-model være unødigt kompleks – særligt hvis konteksten giver anledning til at foretrække en parsimonisk og mere fortolkbar model. Det er en pointe, som vi tager op i afsnit 6.3.

Mønstre i outcome

Målretningspotentialet er også afhængigt af det fænomen, vi forudsiger og målretter efter. Der er jf. afsnit 3.2.3 en grænse for, hvor gode prædiktioner, det er muligt at opnå, når en del af variationen i outcome er tilfældig – som vi fx forventer er tilfældet for et socialt outcome som frafald. Det er ikke muligt på baggrund af vores analyse at vurdere, hvor tæt vi er på denne teoretiske grænse for forudsigelsernes præcision. Med en AUC på 0,727 lykkes det i nogen grad at forudsige frafald, som derfor ikke er et komplet uforudsigeligt outcome. Der er dog et stykke vej op til at realisere en lige så præcis forudsigelse som på ungdomsuddannelserne (se fx Kristoffersen 2015), og en del af forklaringen kan skyldes forskellige grader af determinisme i frafald på ungdomuddannelser henholdsvis videregående uddannelser. Vores undersøgelse føjer sig til rækken at studier, som jævnfør litteraturgennemgangen alle finder målbare korrelationer i data, som prædikterer frafald – uanset om casen er et gymnasium, en professionshøjskole eller et universitet. Samlet set virker det derfor rimeligt at generalisere potentialet for forudsigelse og målretning til uddannelsesområdet generelt.

Vi kan dog ikke på baggrund af vores analyse vurdere, om frafald i større eller mindre grad er et deterministisk fænomen end andre fænomener på øvrige policy-områder. Vores AUC er omtrent på niveau med de nylige casestudier om forudsigelse af tilbagefald blandt hospitalspatienter og recidiv blandt prøveløsladte (Harmsen & Enggaard 2016;

Kleinberg et al. 2017). Det virker sandsynligt, at der findes fænomener, som både har mere og mindre forudsigelige mønstre. Skattesvindel kunne være et eksempel på det første – det virker sandsynligt, at omgåelsen af juridiske regler følger visse forudsigelige mønstre (Public Perspectives 2016). Udbruddet af voldelig konflikt kunne være et eksempel på det sidste, hvor de sociale situationers kompleksitet og kontingens gør mønstrene mere uforudsigelige (Cederman & Weidmann 2017). Jo mere et outcome afhænger af tilfældige ekstrinsiske faktorer, jo lavere performance vil den teoretisk bedste model kunne have – med et lavere målretnings-potentiale til følge (Hofman et al. 2017). Denne pointe rækker ud over casen Metropol og uddannelsesområdet.

I denne henseende forventer vi også, som vi fandt i analysen, at datasættets størrelse har betydning. I vores bedste model har vi omtrent 23.000 observationer, og dens performance er forholdsvist sensitiv over for reduktion i mængden af observationer. Hvis mønstrene bag outcome havde været simplere, må det forventes, at samme performance kunne være opnået med færre observationer. Omvendt vil mere komplekse mønstre kræve et større datagrundlag.

Policy-kontekst

Endelig har også policy-konteksten betydning for potentialet til at bruge maskinlæring til målretning af tiltag. For det første er potentialet afhængigt af, om det er politisk opportunt at målrette tiltag på området. At vi har fået adgang til Metropol som case, kan tages som indikation på, at uddannelsesområdet er en politisk kontekst med relativt stor velvilje omkring målretning af tiltag. Det vil ikke overraske os, hvis det er sværere at opnå en lignende adgang til en institution på retsområdet eller sundhedsområdet. At det er mere politisk følsomt at forskelsbehandle på retsområdet og sundhedsområdet sammenlignet med uddannelsområdet, hvor vi fx er vant til at gøre forskel på folk på baggrund af deres karakterer, taler alt andet lige for et mindre potentiale til at målrette tiltag på de to områder.

Vores framework til målretning er endvidere afhængigt af, hvor nemt policy-konteksten lader sig kvantificere. I casen Metropol inddrager vi fortjenester, omkostninger og effekter af mulige tiltag. Denne tilgang er i princippet generaliserbar på tværs af policy-kontekster, der lever op til følgende rammebetingelser:

- Et outcome, der kan forudsiges
- Et tiltag, man vil målrette, og som man kender effekten af
- Kvantificerbare fordele og ulemper ved tiltaget, fx fortjenester/omkostninger

Det er betingelser, som principielt kan opfyldes af cases, der er helt substantielt forskellige fra Metropol og uddannelsesområdet. Det kan være et jobcenter, som ønsker at målrette et aktiveringstilbud til ledige, som rent faktisk har gavn af det, eller det kan være i retssystemet, at man ønsker at målrette en mildere afsoningsform mod personer, der ikke er i risiko for at begå kriminalitet på ny. Mens prisen på et tiltag vil kunne beregnes i de fleste tilfælde, vil det til gengæld variere mellem policy-områder, om effekten af et tiltag er velkendt, og om værdien af at forhindre eller tilskynde et outcome er kendt og kvantificerbart. Hvad er for eksempel værdien af at afsone med fodlænke frem for i fængsel? En fængselsplads kan spares, men det virker urimeligt ikke at medregne værdien for den enkelte, som nok er sværere at gøre op i kroner og øre. Derfor vil det afhænge af den konkrete policy-kontekst, om vores stiliserede framework i praksis kan give et håndfast handlingsgrundlag for beslutningstagere inden for andre policy-områder.

I denne del af diskussionen har vi forholdt os til validiteten af analysens resultater, og hvorvidt maskinlærings potentiale til målretning kan generaliseres ud over casen Metropol. I de følgende tre afsnit vil vi diskutere, hvorvidt tilgangen også har et potentiale i praksis. Vi ordner diskussionen i metodologiske, epistemiske og etiske perspektiver fra litteraturen.

6.2 Metodologiske refleksioner

Vi ønsker med dette afsnit at pege på og diskutere nogle generelle metodologiske udfordringer, der følger af at anvende prædiktionsmodeller på sociale fænomener. Med skiftet fra estimation til prædiktion følger en række metodiske implikationer, som vi gennemgik i afsnittet om maskinlæring som tilgang. I afsnittet her vil vi genbesøge de teoretiske koncepter og centrale metodiske udfordringer, som følger i kølvandet på dem.

6.2.1 Estimation vs. prædiktion genbesøgt

I analysens afsluttende afsnit blev en svaghed ved frafaldsmodellen og prædiktionstilgangen synliggjort. Da vi i afsnittet tog frafaldsmodellen i anvendelse i praksis arbejdede vi med *antagede* effekter af de tiltag, som Metropol kunne tænkes at implementere. Eksempelvis antog vi, at effekten af at tilbyde en personlig samtale med studievejledningen til de prædikterede frafaldstruede studerende ville resultere i, at 2 pct. at de faktisk frafaldstruede ville blive fastholdt. Det er en antagelse, som vi gør os for at illustrere logikken i, hvordan frafaldsmodellen kan tages i anvendelse i praksis. Med det formål at illustrere logikken er antagelserne ikke kritiske. Om vi antager en fastholdende effekt

på 2 pct. eller 10 pct. er ikke afgørende – det forskubber blot kurverne i de to figurer 5.14 og 5.15 i forrige kapitel, der viser forventede nettofortjenester for frafaldmodellen baseret på hhv. logistisk regression og GBT. Men hvad sker der, når modellen rent faktisk bliver brugt i praksis til at målrette tiltag mod de studerende, som starter på Metropol ved næste studiestart? Her er det næppe tilstrækkeligt at forlade sig på antagede effekter af studievejleding eller mentorforløb eller et hvilken som helst andet tiltag. Vi er interesserede i – eller har ligefrem behov for – at kende effekten af de tiltag, som vi gerne vil målrette. Dermed er vi tilbage ved estimationstilgangen. Det er med andre ord svært at forestille sig, at prædiktionstilgangen kan stå alene.

Sådan vil det dog ikke være for alle prædiktionsproblemer, fx ikke den særlige gruppe af prædiktionsproblemer, som Kleinberg et al. (2015) omtaler *umbrella problems*. Her beder Kleinberg et al. (2015) sine læsere om at forestille sig en situation, hvor en beslutningstager skal beslutte, om vedkommende vil medbringe en paraply på vejen til arbejde, fordi det måske kommer til at regne, når vedkommende skal hjem igen. Her er tale om et rent prædiktionsproblem: Beslutningen om, hvorvidt paraplyen skal medbringes eller ej, afhænger alene af den prædikterede risiko for regn. Beslutningen afhænger eksempelvis ikke af effekten af paraplyen, fordi paraplyen ingen effekt har på, om det kommer til at regne eller ej. Nytten af tiltaget *paraply* afhænger derfor kun af riskikoen for regn. Det forholder sig anderledes i casen Metropol, hvor nytten af tiltaget både afhænger af den studerendes risiko for frafald og effekten af tiltaget, der målrettes.

Det er muligt at forestille sig andre policy-cases, som i større omfang minder om et umbrella problem. Det kunne eksempelvis være målretning af psykologhjælp til veteraner for at bedre tilværelsen for hjemvendte soldater. Det er dog ikke svært at forestille sig, at et målretning af et sådant tiltag også i nogen grad ville have et mere instrumentelt formål for øje som eksempelvis at hindre PTSD eller andre uønskede hændelser. I det tilfælde er det ikke længere et rent prædiktionsproblem i form af et umbrella problem, fordi effekten af tiltaget er ukendt og kommer med i rampelyset: Har psykologhjælp overhovedet en effekt på PTSD? Det vil formentlig være tilfældet ved langt hovedparten at de samfundsvidenskabelige prædiktionsproblemer, hvor det politisk besluttes at målrette et givent tiltag. Med en sådan beslutning i en politisk kontekst vil der formentligt ofte medfølge en forventning om en effekt af tiltaget på et givent outcome, som man ønsker at ændre og ikke blot skærme sig imod som med en paraply mod regn.

Når vi stiller spørgsmål ved de forventede effekter af et givent tiltag følger imidlertid en anden stor udfordring med i kølvandet. Hvad nu hvis effekten af tiltagene varierer individerne imellem? Problemet ligger i, at det effektestimat, som vi bruger, når vi viser, hvordan modellen kan tages i anvendelse i praksis, er ITT, dvs. den gennemsnitlige

effekt af at blive tilskrevet et tiltag. Det er imidlertid ikke givet, at et tiltags estimerede ITT, fx estimeret i et kontrolleret forsøg på Metropol, modsvarer den gennemsnitlige effekt at tiltaget, hvis vi eksempelvis målretter det mod de 10 pct. med den højeste risiko for at frafalde. Det er ikke utænkeligt, at de mest frafaldstruede også er de sværeste at fastholde, eller at de mest frafaldstruede helt generelt reagerer anderledes på et givent tiltag end de mindst frafaldstruede gør. Hvis det er tilfældet, er der tale om, at tiltaget har *heterogene* kausale effekter, og det vil komplicere frafaldsmodellen yderligere (Athey 2017).

Problemet ligger ikke i sig selv i, at de studerende reagerer forskelligt på et givent tiltag – det kender vi fra estimationstilgangen, og det er derfor, vi estimerer en *gennemsnitlig* kausal effekt, lad det være ITT, ATE eller CACE (Gerber & Green 2012). Problemet ligger i, at vi *målretter* et tiltag mod subgruppen af studerende med højest frafaldsrisiko og antager, at der i denne gruppe er den samme forventede effekt af tiltaget. Hvis det ikke er tilfældet, underminerer det vores tillid til den beregnede nettofortjeneste. På samme måde kan vi ikke længere bruge tilgangen til at fastsætte den optimale tærskelværdi, fordi tiltangen er baseret på en antagelse om, at nytten eller fortjenesten af at målrette et tiltag mod en given studerende kun er en funktion af den studerendes risiko for at frafalde. Det er næppe en plausibel antagelse i casen Metropol. Her er det sandsynligt, at effekten og nytten af et tiltag også er en funktion af karakteristika ved den enkelte studerende og den enkelte studerendes situation på studiet i øvrigt.

Hvis der eksisterer heterogene kausale effekter af tiltag imod studiefrafald, vil et næste naturligt skridt være at åbne en mulighed for, at de tiltag, som målrettes, er individualiserede. På den måde kan man målrette det tiltag mod de mindst frafaldstruede, som virker bedst hos dem, og målrette et andet mod de mest frafaldstruede, som virker bedst hos netop dem. Men hvilke tiltag skal vi målrette? Her løber vi imod et problem, som følger af den måske mest centrale forskel på estimation og prædiktion: Med prædiktion identificerer vi ikke årsagssammenhænge, kun korrelationer. Det betyder, at frafaldsmodellen strengt taget ikke gør os klogere på, hvorfor de studerende dropper ud, kun hvem der gør det. Og det er svært at forestille sig, at vi skulle være i stand til at skrue et godt tiltag sammen, som kan fastholde de mest frafaldstruede studerende, hvis vi ikke har viden om årsagerne til deres frafald.

Man kan dog indvende, at vi i analysen med importance-målene trods alt identificerer, hvilke variable som i størst omfang driver frafaldsmodellens prædiktioner. Eksempelvis har variablen for gennemsnitlig holdstørrelse stor indflydelse på prædiktionsmodellen – denne variabel har både det højeste gain, cover og frequence. Kan vi ikke bruge den viden, når vi udformer tiltag mod frafald? Det vil være højst problematisk i forlængelse af argumentet fremført ovenfor: Importance-målene udtrykker alene korrelationer. Vi

kan derfor ikke vide, om holdstørrelsen har en effekt på risikoen for frafald, eller om holdstørrelsen af en anden årsag hænger sammen med frafald. Problemet illustreres også af andre variable med top-20 importance, såsom variablen om de studerendes karaktergennemsnit fra en ungdomsuddannelse. Forestiller vi os, at der er noget særegent og iboende ved karaktersnittet, der påvirker, om de studerende dropper ud? Det gør vi måske, måske ikke, men vi ved det ikke – og under alle omstændigheder forekommer det sandsynligt, at årsagen skal findes blandt den noget mere komplicerede mixtur af årsager til frafald, såsom social og akademisk integration, som vi beskrev i vores litteraturgennemgang på baggrund af bl.a. Tinto (2012). Det kan tænkes, at den akademiske integration alt andet lige er mindre blandt de studerende med de laveste karaktergennemsnit, og i så få fald ville dette være mere nyttigt at have i tankerne, når et tiltag mod uddannelsesfrafald udformes.

Diskussionen her peger samtidig tilbage mod debatten om teoriens rolle i den empiriske samfundsvidenskab, som vi også stiftede bekendtskab med i litteraturgennemgangen. Her lød et argument, at maskinlærings indtog i samfundsvidenskaben gør vores bekymringer om spuriøsitet, kollinearitet og selektionsbias forældede. I forlængelse af ovenstående er dét synspunkt ikke gangbart i casen Metropol – brugen af komplicerede algoritmer baseret på maskinlæring fraskriver ikke teori, fagspecifik viden og kausalestimation deres roller, når formålet er at udvikle og målrette tiltag mod uddannelsesfrafald. På samme måde kan den store datamængde heller ikke kompensere for overvejelser om validiteten og datakvaliteten bag de sammenhænge mellem variable og frafald, som frafaldsmodellen peger på. Lazer et al. (2014) kalder det for *data-hybris*, når volumen skal kompensere for mangelfuldheder i data.

Man kan med rette spørge, om det havde været mere frugtbart at gribe vores analyse anderledes an. En subtil, men betydningsfuld, omformulering af problemstillingen kunne fx have lydt, om maskinlæring har potentiale til at *mindske frafaldet* på uddannelsesinstitutioner? Med den problemstilling ville vi skulle undersøge, om implementeringen af en prædiktionsmodel samlet set – på trods af forventningen om heterogene kausale effekter og andre metodiske udfordringer ved prædiktion – kan indfri en policymålsætning om lavere frafald. Den problemstilling kunne oplagt forsøges besvaret med et kontrolleret forsøg eller et naturligt eksperiment med kausalestimation for øje. Det kunne fx foregå ved at implementere en frafaldsmodel på én af to sammenlignelige uddannelsesinstitutioner. Sådan en undersøgelse kunne bidrage med værdifuld viden om potentialet ved prædiktions-modeller.

Det er dog et helt andet fokus end det metodiske fokus, vi har anlagt i opgaven her. Vi har stillet skarpt på de mekanismer, der udgør væsensforskellen mellem estimation og prædiktion, og på maskinlæring som en metodisk tilgang, der er ny i samfundsviden-

skaben. Det har kastet lys over en række udfordringer, som følger af tilgangen, og som rækker ud over casen Metropol. De indsigter havde vi formentlig ikke fået, hvis vi havde undersøgt potentialet ved maskinlæring som et mere klassisk estimationsproblem.

Vi har med andre ord undersøgt muligheden for at anvende maskinlæring til målretning af tiltag mod uddannelsesfrafald – ikke målretningens effekt på frafaldet. Estimation af modellens effekt på frafaldet ville dog være et oplagt næste skridt for forskning i maskinlærings potentiale i forvaltningen.

6.2.2 Prædiktioner og målretning over tid

At implementere en prædiktionsmodel i praksis giver også anledning til metodiske udfordringer, der vedrører en tidslig dimension. Det skyldes, at formålet ikke er at estimere en strukturel datagenererende proces, som vi er vant til fra estimation. I stedet er formålet at finde sammenhænge i data og målrette tiltag for netop at *ændre* disse sammenhænge. Det giver udfordringer over tid. Der er to mekanismer på spil.

For det første kan vi forestille os en situation, hvor Metropol målretter et succesfuldt tiltag mod de 10 pct. mest frafaldstruede studerende. Lad os sige, at tiltaget i stort omfang eliminerer frafaldet blandt denne gruppe af studerende. Det vil betyde, at de særlige karakteristika ved denne gruppe, som gjorde, at de blev identificeret som meget frafaldstruede, ikke længere vil hænge sammen med frafald. Lad os videre sige, at der var et stort overtal af mænd blandt disse 10 pct. af de studerende. Hvis de bliver fastholdt som følge af et tiltag rettet mod dem, så vil det at være mand i mindre omfang hænge sammen med en større frafaldsrisiko fremover. Men hvad sker der så, når frafaldsmodellen bruges til at målrette et tiltag næste gang? Så vil tiltaget i mindre omfang blive rettet mod denne gruppe mænd og i stedet blive rettet mod en ny gruppe, som nu udgør de 10 pct. mest frafaldstruede studerende. Det er ikke hensigtsmæssigt, hvis der er noget iboende ved det at være mand, som øger frafaldsrisikoen[1]. Vi ændrer med andre ord den datagenererende proces, som frafaldsmodellen er fittet til. Det ændrer vores prædiktioner, fordi modellen ikke er robust mod ændrede korrelationer. Udfordringen minder om et efterhånden klassisk eksempel, Google Flu (Lazer et al. 2014). Google Flu blev bygget til at forudsige influenza-epidemier på baggrund af søgninger på Google, som indeholdt symptomer på sygdommen. Undertiden har Google Flu dog i stort omfang overestimeret udbruddene af influenza. Som det beskrives af Lazer et al. (2014), skal årsagen findes i Googles egne algoritmer: Når man søger på "feber" eller "hoste" bliver man anbefalet søgninger på influenza-symptomer og

[1] Lad det være noget iboende genetisk, hvis mænd fx har sværere ved at sidde stille og koncentrere sig igennem en uddannelse, eller noget semi-iboende, hvis noget strukturelt i samfundet fx i større omfang tilskynder mænd end kvinder til at skifte uddannelse.

influenza-behandling, der er de selvsamme ord, som Google brugte til at forudsige influenza med. Konsekvensen er en stigning i søgninger på influenza-symptomerne. På den måde har algoritmernes klassifikationer i casen Google Flu, såvel som i casen Metropol, indflydelse på den datagenererende proces, som de fitter til.

For det andet, og relateret til ovenstående, kan det tænkes ikke kun at have betydning for frafaldsmønstrene, når nogle *modtager* tiltagene. Idéen med målretning er at målrette tiltaget mod nogle, og således ikke mod andre. Derfor er det en selvstændig problematik, at det kan ændre den datagenererende proces, når nogle studerende *fratages* et tiltag, de ellers ville have modtaget. Det kan tænkes, at et tiltag flyttes fra en lavrisiko-gruppe til en højrisiko-gruppe med den følge, at frafaldet stiger i førstnævnte.

Ovenstående er to mekanismer, der over tid kan ændre den datagenererende proces, som frafaldsmodellen er fittet til. I vores diskussion af analysens interne validitet stødte vi på en yderligere problematik vedrørende frafaldsmodellens tidslige dimension i form af Lucas-kritikken. Kritikken går på, at statistiske modeller som frafaldsmodellen er trænede på historiske data og således afspejler tidligere mønstre i data og dermed også tidligere policies (Lucas 1976). Hvis frafaldsmønstrene ændres over tid, vil modellen derfor kun langsomt tilpasse sig ændringerne, fordi de nye data med de nye mønstre kun vil udgøre en lille del af den samlede volumen af data, som modellen trænes på. Det er problematisk, når en model som frafaldsmodellen bruges til at planlægge fremtidige policies, såsom en fastholdelsespolitik, fordi det ikke er givet, at frafaldsmønstrene vil være de samme fremadrettet.

Det er en nærliggende tanke, at en model baseret på kausalestimation kunne imødekomme denne kritik, idet formålet her som bekendt er at estimere en strukturel, datagenererende proces. En estimationstilgang ville dog ikke løse problemet, for selvom vi her ville estimere en relativt uforanderlig datagenererende proces, så er denne ikke uafhængig af den fysiske og sociale verden. Hvis det at være mand eksempelvis hænger sammen med højere frafald, vil en del af denne effekt formentlig være betinget af samfundsnormer om mænds karrieredrømme, opdragelse med videre, hvilket også kan ændre sig. Sådanne strukturelle sammenhænge er ganske vist typisk træge at ændre, men de er ikke universelle og naturgivne.

De to nævnte mekanismer og Lucas-kritikken trækker i samme retning for frafaldsmodellens potentiale: Vi bør forvente et aftagende potentiale over tid, når vi bruger en given prædiktionsmodel til at målrette tiltag. Det er en hård kritik af maskinlæring anvendt til målretning. Det er på sin vis paradoksalt, at tilgangen spænder ben for sig selv, hvis den virker efter hensigten. I det omfang, at tiltagene har succes med at

fastholde de mest frafaldstruede, vil frafaldet i denne gruppe ikke kunne forudsiges af modellen fremadrettet.

I denne sidste formulering ligger imidlertid også en erkendelse, som tager brodden af kritikken: Sammenhængene i data ændres kun i det omfang, tiltagene har en effekt. I vores case er det langt fra sandsynligt, at tiltagene fuldstændig kan forhindre frafald[2]. Der er tale om tiltag såsom visitationsmøder, som måske kan mindske frafaldet, men næppe udviske de underliggende sammenhænge helt. Modellen vil med andre ord kun rykke en smule ved den datagenererende proces. Endnu bedre stillede er vi i situationer, hvor casen kan betragtes som et *umbrella problem*. I denne type problemer påvirker det målrettede tiltag slet ikke selve outcomet, men afbøder blot effekten af det.

Det er ligeledes muligt at opbløde Lucas-kritikken om, at modellen kun langsomt tilpasser sig ændrede frafaldsmønstre. Denne modelkonservatisme kan ikke helt undgås, men vil dog kunne imødekommes ved at sætte en nyere tidsgrænse for, hvilke dele af datasættet som medtages. Dertil kommer, at en del af styrken ved maskinlæring jo netop ligger i, at modellerne løbende kan tilpasse sig data – i denne henseende er vi bedre stillede end med den mere statiske tilgang, vi kender fra estimation, hvor Lucas-kritikken også gør sig gældende.

6.3 Epistemiske refleksioner

I dette afsnit vil vi diskutere forskellige problemstillinger, der alle relaterer sig til, hvordan algoritmer omsætter data til viden. Vi har allerede konstateret, at maskinlæring kun finder korrelationer i data, og at korrelationer ofte er et utilstrækkeligt grundlag at handle på. Men selv i de policy-kontekster, hvor betingelserne for handling menes at være opfyldt, er det på sin plads at reflektere over karakteren af algoritmisk vidensfrembringelse.

6.3.1 Beslutningsgrundlagets neutralitet

Automatiseringen af beslutningstagning bliver ofte retfærdiggjort med, at algoritmer reducerer risikoen for menneskelig vilkårlighed i sagsbehandlingen (Barocas & Selbst 2016). Den behaviorale samfundsvidenskab skorter ikke med eksempler på, hvordan menneskelig beslutningstagning er påvirket af kognitive biases (Taber & Lodge 2006; Tversky & Kahneman 1975; Kahneman 2011). Algoritmer træffer derimod beslutninger

[2]Den opmærksomme læser vil huske, at vi i analysen som regneeksempel satte forventningerne til tiltagenes effekt relativt lavt på mellem 0,2 og 12,5 pct. fastholdelse.

på et rent datainformeret grundlag. Den eksplicitte maskinisering af en beslutningsproces, dvs. fjernelsen af sagsbehandleren og dermed det menneskelige element, bestyrker indtrykket af datas objektivitet. En sådan "fetichering" af data er imidlertid uheldig (Mittelstadt et al. 2016). Vi vil her opridse tre grunde til, hvorfor algoritmers beslutningsgrundlag ikke uden videre kan regnes for neutralt.

For det første genereres data ikke "af sig selv". Data findes ikke bare. Som redegjort for i litteraturgennemgangen er det derfor en tvivlsom ontologisk antagelse, at data kan have en objektiv betydning frigjort fra teori (Kitchin 2014a). Data skal i første omgang *forestilles* som data, og dermed bygger de allerede på en base af fortolkning. Enhver disciplin har egne normer og standarder for, hvad der kan forestilles som data, hvad der bliver målt, og hvordan det bliver målt (boyd & Crawford 2012). Der er altid en forudgående teoretisk forståelse af et emne, som har betydning for, hvorhen vi retter søgelyset. Det sætter nogle meget konkrete begrænsninger for, hvilke data som er til rådighed, og dermed hvilke mønstre en algoritme kan afdække. Ligeledes er vores feature engineering begrænset af vores egne forestillinger om relevante data – og ikke mindst af institutionelle forestillinger, der afgør hvilke karakteristika, som bliver målt og registreret (Kitchin 2014a).

Det relaterer sig til vores anden pointe. En algoritmisk beslutningsproces giver kvantificerbare data forrang. Mål såsom karakterer og holdstørrelser vejer tungt i modellen, mens et mål som fx undervisningskvalitet slet ikke indgår[3]. Pointen her er ikke blot, at modellen går glip af et prædiktivt potentiale. Pointen er, at de faktorer, som nemt lader sig kvantificere og indarbejde i modeller, også bliver overrepræsenterede i resultaterne og institutionaliseres som betydningsfulde. Det er endnu en pointe, som understreger risikoen ved at handle på baggrund af korrelation. Det er ikke kun, fordi betydningen af fx holdstørrelse måske dækker over en spuriøs sammenhæng – det er også, fordi holdstørrelse har haft lettere ved at komme til udtryk i modellen end potentielt mere betydningsfulde, men svært kvantificerbare variable. Derfor kan det ureflekteret blive de kvantificerbare mål, som en institution orienterer sig efter (Dahler-Larsen 2014; Espeland & Sauder 2007).

Endelig kan der være tilfældige målefejl i data, hvilket er mere en metodisk end en epistemisk pointe. I forrige afsnit berørte vi, hvordan det kunne give problemer over tid. Det er imidlertid en anden konsekvens, vi her vil fremdrage. Vi pointerede tidligere, at tilfældige målefejl i data øger variansen. En konsekvens er, at prædiktionen af den enkelte studerendes frafaldsrisiko bliver mindre pålidelig. Det kan være af stor betydning

[3]Ofte søger institutioner naturligvis at evaluere undervisningskvalitet blandt de studerende og reducere det til et kvantificerbart mål. På Metropol er online surveys af undervisningen dog først opsamlet i databasen fra det nuværende skoleår og lider under meget lav deltagelse.

for den enkelte studerende, hvor prædiktionen er afgørende for, om man falder på den ene eller den anden side af en tærskelværdi og således, om man bliver målrettet et tiltag eller ej.

Data udgør altså ikke et neutralt beslutningsgrundlag. Et gammelt mundheld om databehandling lyder "garbage in, garbage out" og refererer til, at en models konklusioner kun kan være så pålidelige og neutrale som de data, den er baseret på (Barocas & Selbst 2016: 683–684). Denne berettigede kritik må dog holdes op imod alternativet. Lad os i vores case forestille os, at vi i stedet for en algoritme lod det være fx studievejledere, der skulle forudsige frafaldsrisiko med henblik på at målrette fastholdende tiltag. En studievejleder kan i modsætning til en algoritme komme med en holistisk vurdering af den enkelte studerende. Mange års erfaring har formentlig givet fagpersoner en indsigt og intuition, der ikke kan sættes på formel. Er det ikke et bedre beslutningsgrundlag for at målrette tiltag mod frafald? Der er to indsigelser mod dette alternativ.

For det første er der et pragmatisk spørgsmål om kapacitet og ressourcer – der er en grund til, at man i første omgang forsøger at automatisere processen med maskinlæring. Der er ikke noget fagpersonale, som har en direkte kontakt med alle de studerende og løbende ville kunne opdage risiko for frafald. Det ville næppe heller være logistisk muligt at skulle have alle nye studerende forbi en studievejleder inden for de første uger, hvor vi er særligt interesserede i at kende frafaldsrisikoen. For det andet kan man spørge, om holistiske beslutninger trods deres intuitive appel er at foretrække frem for en algoritmes rigorisme. Når data ikke er neutrale, skyldes det i vidt omfang, at de mennesker, som producerer og fortolker data, heller ikke er det. En del af rationalet bag algoritmisk beslutningstagning er jo netop at undgå menneskelige kognitive biases. Det er ligeledes tvivlsomt, om en studievejleder vil kunne processere alle relevante data og kun tage hensyn til forhold, som har en prædiktiv værdi for frafald. Kleinberg et al. (2017) demonstrerer denne pointe i en analyse af dommeres beslutningstagning. Casen er, at dommere skal træffe beslutning om varetægtsfængsling alene baseret på deres vurdering af, om de sigtede vil begå kriminalitet igen eller ej. Dommerne skal med andre ord prædiktere et outcome. Kleinberg et al. (2017) argumenterer for, at dommernes holistiske beslutninger kan sammenlignes med "overfitting". Argumentet er, at dommerne inddrager en række individuelle hensyn fra sag til sag, som egentlig er "støj" og ikke burde gøres til en del af ligningen, fordi de ikke hænger sammen med forudsigelsen af, om de sigtede vil begå kriminalitet på ny.

Hvis vi således definerer formålet snævert som præcise prædiktioner, vil maskinen nok trumfe mennesket. Der kan dog være andre hensyn at tage end prædiktionsperformance. Algoritmers rigoristiske, regelbaserede beslutningstagning går også under en betegnelse med negative konnotationer: teknokrati. Når holistisk sagsbehandling

er intuitivt appellerende, er det ikke på grund af, at beslutningsgrundlaget er mere neutralt, men fordi det gør forvaltningen af regler mindre rigid. Meget lidt offentlig forvaltning kan reduceres helt til algoritmer. Der vil altid være behov for et element af skøn i sagsbehandling; det er sagsbehandlingens raison d'etre (Andersen 2014). Det er en tråd, vi følger op på i afsnit 6.4.1 om accountability.

6.3.2 Prædiktioners usikkerhed

Et andet epistemisk aspekt ved maskinlæring vedrører den forståelse, vi har af resultaternes usikkerhed (Mittelstadt et al. 2016). Vi er vokset op med at forstå verden igennem en linse af årsager og effekter (Mayer-Schönberger & Cukier 2013: 163). Vi leder efter sikker viden at forstå verden ud fra og handle på, og vi forventer derfor også, at politisk forvaltning er baseret på sikker viden og evidens (Power 1997).

De prædiktioner, som frafaldsmodellen i casen Metropol leverer, er dog fejlbarlige. Det indfanges bl.a. af det berømte citat af George Box: "*All models are wrong, but some are useful.*" (bl.a. i Box 1979). Vi stiftede også bekendtskab med dén observation i afsnit 3.2.3 om grænser for prædiktions-performance: Uanset hvor god en frafaldsmodel, vi opstiller, vil den være fejlbarlig, fordi frafald ikke er en deterministisk proces, men også resultatet af tilfældig variation, hvilket introducerer en ikke-reducerbar fejl i modellen. Derudover kan det også stride mod vores forståelse af sikker viden, at modellen er probabilistisk. Hvis vi eksempelvis prædikterer 70 pct. sandsynlighed for, at en studerende frafalder, prædikterer vi samtidig 30 pct. sandsynlighed for, at den studerende ikke frafalder. Hvis den studerende ikke frafalder, var vores prædiktion så forkert? Ikke nødvendigvis. Det kan vi strengt taget ikke vurdere på baggrund af én enkelt prædiktion, men kræver teoretisk set, at den samme studerende gennemlever sit studieforløb 100 gange og kun frafalder i 70 af dem. En sådan probabilistisk tankegang kan stride mod vores intuitive forståelse af sociale fænomener, fordi den studerende aldrig kan gennemleve det præcist samme studieforløb mere end én gang.

I forlængelse heraf kan vi spørge, hvor sikre vi skal være på en prædiktion, for at denne kan danne grundlag for handling. Hvor stor en fejlmargin er vi villige til acceptere? To aspekter er relevante at nævne her. Det første drejer sig om modellens reducerbare fejl, og om hvorvidt modellen er præcis nok set i lyset af, hvor præcis den teoretisk set kan blive. Her løber vi ind i den udfordring, at vi ikke kender den øvre grænser for modellens prædiktions-performance. Vi kan konkludere, at en model med en AUC = 0,7 er bedre end en model med AUC = 0,6, men hvis mere data og bedre tuning kunne hæve AUC til 0,8, så ville vi næppe stille os tilfredse med modellen. Hvis en AUC på 0,7 omvendt er den øvre grænse for forudsigelsen af et givent fænomen grundet stor tilfældig variation

i fænomenet, ville vi formentlig være mere tilbøjelige til at stille os tilfredse med modellen, selvom den er lige så fejlbarlig som før. Udfordringen her udstiller desuden den ulempe ved prædiktionstilgangen, at der ikke findes udbredte praksisser for at evaluere en prædiktionsmodels robusthed og resultaternes usikkerhed (som vi også så i vores egen analyse), sådan som vi eksempelvis kender det fra estimationssammenhæng, hvor vi estimerer standardfejl og konfidensintervaller (Hofman et al. 2017). Det er nemt at problematisere den mere eller mindre arbitrære grænsedragning mellem p-værdier på henholdsvis over eller under 0,05, men i estimationssammenhæng eksisterer trods alt konventioner for, hvad man videnskabeligt anser for statistisk signifikante resultater. Det gør der ikke i prædiktionssammenhæng, hvilket alt andet lige øger det, man kan kalde *researcher degrees of freedom*. Det hæmmer troværdigheden af en analyse (Simmons et al. 2011).

Fraværet af konventioner om resultaternes usikkerhed leder frem til det andet aspekt ved, om vi er villige til at acceptere en given fejlmargin. Det er et aspekt, som drejer sig om, at vurderingen af usikkerheden for en prædiktionsmodel også har en etisk dimension, som er svær at adskille fra den konkrete handling, der målrettes. Jo mere indgribende et tiltag er over for den enkelte, jo sværere vil det også være at retfærdiggøre fejl. Vi er ikke så bekymrede for at målrette et tilbud om et frivilligt mentorforløb, men hvor høj skal en frafaldsrisiko være, før vi eksempelvis pænt tør bede en nystartet studerende om at genoverveje sin studieplads? Hvor sikre skal vi være på, at en fængslet ikke begår ny kriminalitet, før vi prøveløslader?

Hidtil har vi behandlet dette etiske element i modellens præcision relativt ubekymret. I analysen viste vi eksempelvis, hvordan tærskelværdien for prædikteret sandsynlighed kan justeres op og ned for at øge enten modellens præcision eller mængden af sande positive. Vi viste videre, hvordan en kalkule over Metropols mulige gevinster ved forskellige tærskelværdier kan opstilles. Vi forholdt os altså kun til de tekniske elementer i fastsættelsen af tærskelværdien, og ikke de etiske dilemmaer, som uløseligt hænger sammen hermed. Denne ubekymrethed afspejler i nogen grad vores konkrete case, hvor fejlklassifikationer i forhold til andre områder, fx sundhedsområdet, har en relativt lille betydning. For det første er det ikke katastrofalt med falske negative, dvs. frafald som vi ikke opdager. Her er omkostningen en forpasset mulighed for at rette tiltag mod en frafaldstruet studerende, men dette ville også være tilfældet helt uden modellen. For det andet er vi ikke så bekymrede for falske positive, dvs. studerende som vi fejlagtigt klassificerer som frafaldstruede. De påtænkte tiltag er relativt harmløse sammenlignet med eksempelvis omfattende medicinsk behandling. Der skal dog ikke meget fantasi til at forestille sig, hvordan policy-konteksten kunne ændre sig, sådan at man fx lagde mere direkte pres på de frafaldstruede studerende om at vige deres plads for en anden.

Konteksten behøver dog ikke være åbenlyst indgribende, før disse implikationer af falske negative eller positive egentlig er større, end vi har regnet dem for at være. Lad os eksempelvis overveje muligheden for, at et visitationsmøde med studievejledningen kan skubbe til en lille tvivl om ens studievalg eller studieegnethed. Lad os antage, at denne tvivl i nogle få tilfælde bliver udslagsgivende for, at en studerende falder fra, som ellers ville have gennemført. På den måde kan der være en lille risiko for, at en forkert prædiktion får en høj omkostning. For at kaste et blik ud over casen Metropol kan vi tage et eksempel fra sundhedsområdet: Skal vi operere hoften for en ældre svækket person? Opvejer den høje sandsynlighed for en lille forbedring i livskvalitet, den lille risiko for store komplikationer (Kleinberg et al. 2015)? Denne type af overvejelser er nødvendige at have med, når vi målretter tiltag i praksis. Opvejer den høje sandsynlighed for en lille forbedring af frafaldet, den lille risiko for en meget negativ effekt?

I et probabilistisk verdenssyn vil alle valg indebære en sådan afvejning af mulige gevinster og risici. Det er ikke noget særligt for maskinlæring som tilgang. Som Mayer-Schönberger & Cukier (2013) argumenterer for, har vi altid levet i en verden af probabiliteter, bare uden at være bevidste om det. Derfor forhindrer usikkerhed os da heller ikke i forsøg på at handle ud fra prædiktioner. Resultaternes iboende usikkerhed er ikke nogen afvæbnende kritik af algoritmisk beslutningstagning. Algoritmer er sjældent ment til at være ufejlbarlige (Mittelstadt et al. 2016). De bliver implementeret i kontekster, hvor der mangler et bedre alternativ. Eksempelvis er der på Metropol foreløbig heller ikke lagt op til, at algoritmer helt skal erstatte individuel sagsbehandling. De skal målrette studievejledernes indsats, ikke erstatte den fuldstændig.

Indsigten fra dette afsnit er således opsamlende, at en velovervejet implementering af en prædiktionsmodel forudsætter en anerkendelse af modellens fejlbarlighed og prædiktionernes usikkerhed. I et fravær af konventioner er det endvidere nødvendigt fra sag til sag at vurdere, om modellens usikkerhed er acceptabel. Denne vurdering kan ikke foretages i et vakuum, men er afhængig af den konkrete policy-kontekst, og vurderingen foretages samtidig i et samfund, som i en given grad har en aversion mod risici (Power 1997). Implementeringen af frafaldsmodellen hviler på epistemiske overvejelser som disse, og derfor hviler potentialet ved målretning af tiltag mod uddannelsesfrafald ligeledes på overvejelserne.

6.3.3 Manglende transparens i beslutningstagningen

Et sidste epistemisk aspekt, vi vil berøre, vedrører selve processen, hvormed data omsættes til outcomes. Når der træffes en beslutning i forvaltningen, er det en demokratisk

forventning, at sammenhængen mellem beslutningsgrundlag og afgørelse, mellem data, outcome og tiltag, skal være transparent. Det vil sige, at den skal være forståelig og åben for kontrol og kritik (Mittelstadt et al. 2016: 4). I algoritmisk beslutningstagning kan det være svært at leve op til, idet algoritmernes beslutningsregler ikke nødvendigvis er intuitivt forståelige. Tværtimod bliver maskinlæring ofte karakteriseret som en *black box*, dvs. en utilgængelig form for vidensproduktion (Pasquale 2015). Hvor samfundsvidenskaben traditionelt er optaget af parsimoniske teorier og modeller, der på en nyttig måde forsimpler komplekse fænomener, er maskinlæring nyttigt til at tilpasse sig efter al kompleksiteten i data.

Vi kommer dermed ind på et praktisk tradeoff mellem modelpræcision og fortolkbarhed, som vi beskrev i afsnit 3.3 om algoritmevalg. Vores GBT-model opnår ganske vist nævneværdigt mere præcise forudsigelser end det simple klassifikationstræ, men den er også meget sværere at forstå. Boosting-processen, hvor på hinanden følgende træer fittes til de forriges residualer, resulterer i en model med komplet utilgængelige beslutningsregler.

De beslutningsregler, som et klassifikationstræ følger, kan derimod fremstilles grafisk og relativt intuitivt. Figur 3.8 på side 45 viste stubben af et klassifikationstræ, som vi trænede på frafaldsdata fra Metropol. En sådan lavpraktisk mulighed for fortolkning kan meget vel være en stor fordel, hvis forvaltningen står over for at skulle forklare og retfærdiggøre algoritmiske beslutninger. En studievejleder vil måske gerne kunne begrunde over for en studerende, hvorfor vedkommende er blevet indstillet til et særligt mentorforløb. En sådan forklaring vil være praktisk taget umulig med vores GBT-model. Det vil ikke være muligt at redegøre for præcis, hvilke af den studerendes variable som indgår i afgørelsen eller med hvilken vægt. Valget af det simple træ vil til gengæld resultere i et betragteligt tab af nøjagtighed (med en AUC-værdi på 0,656 mod GBT-modellens 0,727). Om det kan opvejes af modellens øgede fortolkbarhed må ikke desto mindre hvile på en praktisk afvejning i den konkrete forvaltningspraksis. En mulig mellemvej er en model baseret på fx Random Forest, som performer næsten lige så godt som GBT-modellen. Den er ikke så simpel som et enkelt træ, men følger en beslutningslogik, der er lettere at redegøre for end GBT, omend det stadig ikke er muligt at illustrere beslutningens grundlag i form af et flowchart af betingelser.

Der findes forskellige bud på, hvordan den manglende transparens kan imødekommes ved at mindske modelkompleksiteten og forsøge at gøre outcome fortolkbart (se fx Ribeiro et al. 2016). En del af uigennemsigtigheden kan dog ikke fjernes ved modelforsimpling, da den er et iboende træk ved beslutningstagning baseret på maskinlæring. Det er således ikke blot datas og algoritmernes kompleksitet, som bidrager til uigennemsigtigheden. Det er også selve den foranderlige beslutningslogik (Burrell 2016).

Maskinlæring er jo netop karakteriseret ved, at modellen løbende tilpasser sig nye data. Enhver grafisk fremstilling af beslutningslogikken vil derfor være et øjebliksbillede, som ændrer sig næste gang, modellen trænes med de nyeste informationer. En offentlig myndighed ville derfor ikke kunne sikre fuld transparens, om de så lagde algoritmerne offentligt frem.

Konsekvensen af den manglende transparens er, at legitimiteten af en beslutning bliver svær at udfordre (Kitchin 2014b: 172-174). Det sætter et betydeligt spørgsmålstegn ved maskinlærings potentiale i den offentlige forvaltning. Her vil vi som borgere i et demokratisk samfund oftest have krav på at forstå det konkrete rationale bag en afgørelse (Andersen 2014). Der kan være visse kritiske anvendelsesområder, hvor man dermed helt er nødt til at afholde sig fra maskinlæring. Det vil formentlig afhænge af, hvor indgribende karakter en beslutning har. Det er tråde, som vi følger op på i næste afsnit.

6.4 Etiske refleksioner

I dette afsnit vil vi diskutere de etiske problemstillinger, som følger af anvendelsen af maskinlæring til prædiktion og målretning. Hvor de epistemiske refleksioner hidtil har angået karakteren af den viden, som ligger til grund for en handling, vil vi i dette afsnit fokusere på selve handlingen. Det gør vi med en vifte af normative overvejelser, der angår handlingernes *retfærdighed* eller *fairness*, og som derfor er relevante selv i det tilfælde, at vores vidensgrundlag havde været neutralt, sikkert og transparent. Det er problematikker, som handler om forskelsbehandling, accountability, retten til privatliv og konstitutive effekter af målretning.

6.4.1 Accountability

Med udviklingen af en prædiktionsmodel bliver ansvaret for at målrette Metropols tiltag mod frafald delegeret til en algoritme. Det er stadigvæk en menneskelig beslutning, hvilke tiltag der skal iværksættes, og hvor bredt de skal rulles ud. Som vi så med frameworket i analysen, kan disse beslutninger til en vis grad formaliseres, sådan at der kan blive sat en optimal tærskelværdi for prædikteret frafaldsrisiko. Derefter er det imidlertid en algoritmisk beslutning, om den enkelte studerende falder over eller under tærskelværdien. Spørgsmålet er, hvordan det stiller studerende, som måtte være uenige i afgørelsen. Hvem står til ansvar for beslutningen, når det er en algoritme, som tager den?

En algoritme kan næppe i sig selv have et etisk ansvar for sine afgørelser[4], men det kan til gengæld de, som skriver den. Det må være en rimelig forventning, at epistemiske refleksioner om modellens fejlbarlighed indgår i udviklingsprocessen, fordi dårligt design kan føre til fejlagtige prædiktioner (Mittelstadt et al. 2016). I noget omfang må en programmør altså kunne stå på mål for en algoritmes udformning. Ansvarstilskrivelsen er imidlertid ikke så ligetil, når der er tale om selvlærende algoritmer. En programmør kan godt forsøge at håndtere data efter alle kunstens forskrifter, men bestemmer ikke de konkrete beslutningsregler. De er i sagens natur datadrevne og foranderlige. Dermed står vi potentielt med et *accountability gap*, hvor ingen har kontrol over beslutningen, og hvor ingen derfor påtager sig ansvaret (Mittelstadt et al. 2016: 9).

Måske er netop dette aspekt af maskinlæring endda med til at gøre tilgangen tiltalende for beslutningstagere. I litteratur-reviewet påpegede vi, at hypen omkring maskinlæring og big data kan ses i lyset af en samfundsmæssig udvikling mod stadig større usikkerhed. Når beslutningstagere er så villige til at afprøve nye datadrevne governance-værktøjer, kan det forklares med, at usikkerhed skaber en enorm efterspørgsel på evidens (Dahler-Larsen 2011). I den henseende ligger maskinlæring i naturlig forlængelse af en længere historisk bevægelse mod en centraliseret forvaltning, hvor metodisk strenghed, standardisering og kvantitativ beskrivelse træder i stedet for tidligere tiders autoritetstro og tillid til offentlige myndigheder (Porter 1996). Beslutninger bliver forandret fra at være handlinger baseret på eksperters dømmekraft til at være resultatet af standardiseret regelefterfølgelse, hvormed "*... impersonality rather than status, wisdom or experience becomes the measure of truth.*" (Rose 1991: 678).

En del af denne afpersonalisering af beslutninger anser vi typisk som ønskværdig, eksemplificeret ved et weberiansk idealbureaukrati kendetegnet ved saglig sagsbehandling fri for menneskelig vilkårlighed. Problemet opstår, når bureaukratiets beslutninger ikke kan udfordres. Når en algoritme fx forudsiger, at der er 50 procents sandsynlighed for, at en studerende dropper ud, kan det give et tilsyneladende objektivt grundlag for at iværksætte et policy-tiltag. Det giver en måde at affeje beslutningens usikkerhed på ved at rationalisere problemet til et spørgsmål om numerisk information. Det bliver med andre ord en måde at træffe beslutninger på, uden at et konkret menneske skal beslutte sig. Hermed eksternaliseres beslutningen fra de menneskelige beslutningstagere og gør det uklart, hvem der har det endelige ansvar.

Faren ved denne automatisering af beslutninger er endvidere, at forvaltningen bliver meget teknokratisk. En forvaltningsbeslutning må altid kunne ankes (Andersen 2014). Som vi tidligere har argumenteret for, gør manglende algoritmisk transparens det ekstremt

[4]Muligheden for etiske maskiner er dog et genstand for en del forskning i maskinlæring knyttet til kunstig intelligens (Mittelstadt et al. 2016)

svært at appellere til, at grundlaget for en beslutning er fejlagtigt. Det algoritmiske beslutningsgrundlag for en afgørelse er simpelthen til en vis grad en *black box*. I værste fald kan de studerende ende i et kafkask system, der kan underlægge dem individuelle beslutninger uden at yde anden begrundelse, end at det er algoritmisk bestemt (Kitchin 2014a).

For at undgå dette scenarie er offentlige myndigheder nødt til at påtage sig ansvaret for konsekvenserne af en prædiktionsmodel. Hvis de ikke kan begrunde indholdet af den enkelte afgørelse, er de i det mindste nødt til overordnet at kunne retfærdiggøre anvendelsen af en automatiseret beslutningsmodel. Er det overordnede formål – i vores tilfælde at begrænse frafald – tilstrækkelig vigtigt til at retfærdiggøre algoritmisk beslutningstagning? Det indebærer blandt andet en grundig overvejelse af de etiske konsekvenser af fejlklassifikationer, som vi berørte med forrige afsnit. Hvor store er de mulige negative effekter af falske positive og negative? (Johnson 2014).

Det er endvidere nødvendigt politisk at anerkende den iboende usikkerhed ved alle beslutninger, som vedrører komplekse sociale sammenhænge. Det problem kan ikke overkommes ved brug af datadreven governance. Det er således utilstrækkeligt at håndtere fastsættelsen af en tærskelværdi for vores prædiktioner som et rent teknisk problem. Selvom en prædikteret sandsynlighed kan fremstå som et objektivt mål, der kan retfærdiggøre forskelsbehandling af studerende, er prædiktion forbundet med grundlæggende usikkerhed. Det er et maskiniseret, datadrevet skøn, som må være åbent for kritik, så den enkelte studerende har reel mulighed for at få omgjort en afgørelse (Andersen 2014).

I praksis vil maskinlæring derfor skulle implementeres fleksibelt og sensitivt over for policy-konteksten. I det omfang, at der er tale om indgribende policy-tiltag, vil den sikreste vej nok være at lade en menneskelig sagsbehandler træffe den endelige afgørelse, sådan at det prædikterede outcome kun bliver ét blandt flere input til beslutningen. I vidt omfang er det også sådan en anvendelse, der er lagt op til på Metropol. Foreløbigt har alle de tiltag, som er udtænkt, en frivillig karakter. Det er ikke sådan, at der træffes algoritmisk beslutning om implementering af tiltag, kun om målretningen. Ikke desto mindre vil det nok være klogt at forberede sig på, hvordan man skal forholde sig den dag en studerende over tærskelværdien opponerer imod at blive betragtet som frafaldstruet – eller den dag en studerende under tærskelværdien gør krav på et mentorforløb på lige fod med sine frafaldstruede medstuderende.

6.4.2 Forskelsbehandling

Helt centralt i debatten står sådanne spørgsmål om forskelsbehandling. Præmissen for hele vores problemformulering er, at det i et eller andet omfang kan retfærdiggøres "at målrette policy-tiltag" og dermed gøre forskel på folk. Hvis enhver form for forskelsbehandling i den offentlige sektor var etisk uforsvarlig, ville det ikke være relevant med en prædiktionsmodel. En sådan position, der afviser enhver form for diskrimination, er dog heller ikke udbredt. Den offentlige forvaltning benytter sig allerede i vidt omfang af diskrimination baseret på gruppetilhørsforhold (Mayer-Schönberger & Cukier 2013: 160–161). Vi tilbyder fx screeninger og vacciner på baggrund af aldersgruppe, økonomisk støtte på baggrund af boligform, adgang til universitetet på baggrund af studenterbevis osv. Der er dog en fortsat etisk og juridisk debat om, hvornår diskrimination er i orden – med mange kuriøse eksempler på gråzoner i samfundet. Det er fx dømt ulovligt at diskriminere efter køn, når prisen fastsættes på en bilforsikring eller for en klipning hos frisøren (Hansen 2013).

Som oftest vedrører debatten en række særlige gruppekategorier som fx køn, etnicitet og religion, der typisk er blevet juridisk beskyttet på baggrund af historiske tilfælde af diskrimination (Barocas 2014). Eksempelvis indgår data om etnicitet ikke i vores model. Det er et bevidst valg, fordi Metropol ikke ville forsvare, at etnicitet kom til at indgå som faktor i en beslutningsproces[5]. Det er et valg, som ganske givet forringer prædiktionsmodellen jævnfør tidligere casestudier, hvor samme hensyn ikke blev taget (Kovačić 2010; Şara 2014).

Denne form for normative hensyn til særlige gruppekategorier nyder bred opbakning. Argumentet bliver ofte framet som et spørgsmål om *statistisk diskrimination*, hvormed der menes, at det er urimeligt at blive diskrimineret på baggrund af sit gruppetilhørsforhold (Barocas & Selbst 2016; Lippert-Rasmussen 2011). Det ville ud fra denne forståelse være uretfærdigt, at en studerende med anden etnisk baggrund end dansk blev behandlet som frafaldstruet, alene fordi gruppen af studerende med anden etnisk baggrund *gennemsnitligt* havde et højere frafald. Diskrimination handler altså ikke om, hvorvidt der eksisterer en statistisk sammenhæng eller ej, eller om prædiktionerne er sande eller falske. Det handler om, at det kan anses for uretfærdigt, at man bliver behandlet ud fra ens gruppetilhørsforhold frem for at blive behandlet som et selvstændigt individ (Kitchin 2014a).

Det er i denne sammenhæng almindeligt at skelne mellem positiv og negativ forskelsbehandling, hvorfor vi fx på nogen områder laver kvoter for dårligere stillede

[5]At køn ikke anses som lige så problematisk i denne konkrete policy-kontekst er selvfølgelig interessant, men ikke noget vi diskuterer yderligere

grupper. Hvis frafaldsmodellen på Metropol bliver anvendt til at hjælpe en udsat gruppe af studerende igennem studiet, kan det ses som positiv forskelsbehandling, der skal hjælpe med at bryde en historisk sammenhæng i data, vi som samfund finder socialt uretfærdig. Det vil mange finde mere acceptabelt, end hvis informationen skulle bruges til at smide studerende ud. Fordi de studerende har forskellige udgangspunkter, er det nødvendigt at behandle dem ulige for at stille dem lige. Dermed evaluerer vi etikken i forskelsbehandling ud fra vores forståelse af social retfærdighed (Lippert-Rasmussen 2011). Andre mener dog, at formålet med diskriminationen er irrelevant og ikke retfærdiggør forskelsbehandling (Barocas & Selbst 2016: 695). Der er stadig tale om, at den enkelte bliver behandlet ud fra sin gruppe og ikke som et individ (Lippert-Rasmussen 2011). Derfor bliver rimeligheden af fx kvote-politikker da også jævnligt debatteret. Som vi skal diskutere i sidste afsnit, er det heller ikke altid ligetil at afgøre, om anvendelsen af en prædiktionsmodel kun har den ønskede effekt og dermed alene kan regnes som positiv diskrimination.

I forlængelse heraf må vi også overveje, om det overhovedet er et tilstrækkeligt værn med statistisk diskrimination, at vi fjerner etnicitet fra frafaldsmodellen. Litteraturen viser, at det kan være svært at gardere sig mod statistisk diskrimination, fordi den beskyttede kategori ofte korrelerer med en lang række andre variable, som stadig er med i datasættet (Barocas & Selbst 2016). Modellen risikerer dermed at finde proxyvariable for etnicitet. Hvis fx postnumre eller bestemte gymnasier korrelerer med etnicitet, og etnicitet korrelerer med frafald, så vil modellerne implicit forskelsbehandle på baggrund af etnicitet. Det kan lyde uskyldigt, men det ville nok være temmelig kontroversielt, hvis det blev kendt, at Metropol behandlede studerende fra Albertslund og Brøndby Strand anderledes end øvrige postnumre.

Denne form for statistisk diskrimination gennem en proxyvariabel ville dog også være en risiko ved menneskelig sagsbehandling (Agan & Starr 2016). Vi vil argumentere for, at maskinlæring – i hvert fald i teorien – tilbyder et bedre værn mod diskrimination end menneskelige beslutningstagere. Det er ikke, som man måske kunne tro, fordi man med maskinlæring bedre kan sikre sig, at bestemte variable udelukkes fra at indgå i modellen. Det er tværtimod, fordi man med maskinlæring kan forsøge at mætte modellen med al tilgængelig information, som er langt mere end noget menneske ville kunne overskue. En fordel ved maskinlæring er jo netop, at man kan modellere så komplekse sammenhænge i data, at vi nærmer os *individualiseret* prædiktion. Ved traditionel statistisk beskrivelse ser vi på gennemsnitlige tendenser, og dermed er vi henvist til at behandle alle i en gruppe ens, dvs. som *typiske* studerende. Med maskinlæring kan vi i videre udstrækning modellere mangfoldigheden af subjekter ved at tage alle karakteristika i betragtning (Johnson 2014). Hvis vi har tilstrækkeligt med rige og granulære informationer om det

enkelte individ, vil vi i princippet nærme os en idealiseret situation, hvor statistisk diskrimination ophører som fænomen.

Det er en vidtgående konklusion, som næppe ville nyde udelt opbakning. Kritikere kunne med en vis ret indvende, at argumentet bygger på en forsimplet, deterministisk forståelse af den sociale verden (Kitchin 2014a: 8), og at maskinlæring ikke kvalificerer sig som individuel behandling, men blot som en disaggregering af individet til en sum af karakteristika (Johnson 2014: 5). Her er det dog på sin plads at reflektere over, hvad det egentlig vil sig at behandle nogen som et individ i forvaltningssammenhæng. Lad os forestille os en sagsbehandler, der sidder over for en studerende og skal komme med en individuel prædiktion af vedkommendes frafaldsrisiko. Uanset hvad den studerende må komme med af oplysninger, kan sagsbehandleren ikke meningsfuldt udtale sig om den studerendes risiko uden et kendskab til, hvilke karakteristika der sandsynliggør frafald. Med andre ord er det nødvendigt at forholde sig til den studerenes lighed med andre grupper af studerende. Individuel behandling implicerer altså viden om karakteristika på gruppeniveau (Lippert-Rasmussen 2011). Derfor må vi kunne betragte det som individuel behandling, så længe sagsbehandlingen tager højde for alle relevante informationer, der med rimelighed er tilgængelige i en situation (Lippert-Rasmussen 2011: 54).

Svagheden ved dette argument er, at vi er nødt til at konkretisere, hvilke informationer som er relevante og tilgængelige. I forrige afsnit fremførte vi den epistemiske pointe, at det aldrig ville kunne lade sig gøre at udtømme mulighedsrummet af relevant information. Dertil kunne man føje en lavpraktisk indsigelse om, at mere granulære og detaljerede data ofte er besværlige og dyre at tilvejebringe. Metropol kunne både købe sig til ekstra socioøkonomiske data og foretage hyppige obligatoriske surveys af de studerende, men her kommer de praktiske omstændigheder i vejen. Vi vil altså aldrig få nok information til perfekt individualiseret behandling. Den vigtige, pragmatiske erkendelse er imidlertid, at statistisk diskrimination er en forsømmelse, som kommer i grader (Lippert-Rasmussen 2011). Jo mere relevant information, vi har med i vores model, jo mindre diskriminerer vi på baggrund af simple gruppetilhørsforhold. Der er stor forskel mellem en kompleks model, som ganske vist mangler visse relevante dimensioner, og en model, som *alene* diskriminerer på baggrund af etnicitet.

I praksis er vi langt fra idealet om præcis, individualiseret målretning i vores case. Sådan vil det nok være for alle offentlige institutioner. Vi kan imdilertid også stille spørgsmålstegn ved, om mere præcise forudsigelser altid er at foretrække (Barocas 2014). Diskussionen i dette afsnit har taget afsæt i det udbredte synspunkt, at det er ønskeligt at blive behandlet som individ. Vi vil her afrundingsvis lufte to argumenter imod, at det altid er tilfældet.

Det første er, at det kan være mere inkluderende og solidarisk at behandle alle ens og ignorere visse karakteristika, selvom det muligvis har stor prædiktiv værdi. Mange studerende vil sikkert foretrække, at deres høje alkoholforbrug, sygefravær osv. ikke indgår i en model. Vi vil med andre ord hellere anonymisere visse karakteristika, end behandles som individer (Lippert-Rasmussen 2011). Det andet er, at målretning af tiltag jo også betyder, at nogle ikke får behandlingen. Det er den hyppigste indvending mod positiv diskrimination. Fordi ligebehandling er så central en forventning i et demokratisk samfund, kan Metropol få svært ved at retfærdiggøre over for deres studerende, at nogle tilbydes eller tvinges til et mentorforløb, mens andre ikke gør. Også selvom de appellerer til en forestilling om positiv diskrimination. Hvis en studerende oplever sin behandling som uretmæssig, er det måske trods alt mere acceptabelt i et system, hvor alle behandles ens, end i et system, hvor man oplever at blive udpeget eller fravalgt ud fra personlige karakteristika.

6.4.3 Retten til privatliv

Med de afsluttende argumenter i forrige afsnit berører vi et andet dillemma om brugen af data i forvaltningen: afvejningen mellem hensynet til den potentielle samfundsmæssige gevinst og hensynet til den enkeltes ret til privatliv (Kitchin 2014b: 165; Foster et al. 2016: 299). Kitchin (2014b) indfanger flerheden af dilemmaer med formuleringen: "*Indeed, there are many fundamental normative questions that need reflexive consideration concerning who can generate, access, share, analyse datasets, for what purposes, in what contexts, and with what constraints.*" (Kitchin 2014b: 183).

Retten til privatliv er i Danmark blandt andet beskyttet med persondataloven, der sætter rammerne for, hvornår og hvordan virksomheder, organisationer og myndigheder kan opbevare og behandle personoplysninger (Retsinformation 2000; Datatilsynet 2015). Selve hensynet til retten til privatliv vedrører derfor helt generelt administrationer, som opbevarer personoplysninger. To forhold gør dog overvejelserne om retten til privatliv særligt relevante, når en administration som Metropol ikke bare opbevarer, men også påtænker at bruge maskinlæring til at målrette tiltag på baggrund af personoplysningerne. For det første giver maskinlæring Metropol mulighed for at monitorere samtlige studerendes data i takt med, at data genereres, fordi modellerne selv kan tilpasse sig disse. For det andet er der ikke bare tale om monitorering, men om monitorering som danner grundlag for at iværksætte målrettede tiltag. Med udgangspunkt i disse to forhold vil vi herunder kaste lys over hensynet til retten til privatliv i casen Metropol ved at forholde os til dels følsomheden af data og brugen af data (Foster et al. 2016).

Hensynet til følsomheden af data vedrører især tre forhold. Det første vedrører et ønske om anonymitet og retten til at definere skellet mellem, hvad der er offentligt, og hvad der er privat (Kitchin 2014b: 172–173; Foster et al. 2016: 299–300). Noget data opleves mere privat og følsomt end andet. Eksempelvis deler mange af os gladeligt information og data om os selv på sociale medier som Facebook, Instragram og LinkedIn. Samtidig ville de fleste af os nok være mere betænkelige ved at dele vores journaler fra lægen åbent på nettet – og måske også ved, at andre offentlige myndigheder end lægen frit skulle have adgang til vores journaler. I casen Metropol kunne sygdomshistorik ganske givet have en høj prædiktiv værdi, når vi forudsiger frafald, men det er ikke sikkert, at de studerende synes, at deres sygdomme og skavanker er et anliggende, der vedører Metropol. På samme måde er det ikke sikkert, at de studerende synes, at deres karakterer fra ungdomsuddannelsen bør have betydning for, hvordan de bliver behandlet af Metropol, når de først har fået en plads på studiet.

Det andet forhold drejer sig om datasikkerhed. Her er omdrejningspunktet at sikre kontrol med, hvem der har adgang til data, hvilket både gælder når data genereres, opbevares og udveksles. Når mange datasæt fra forskellige aktører udveksles og kobles, øges risikoen for brud på datasikkerheden (Foster et al. 2016: 300; Kitchin 2014b: 176).

Det tredje forhold drejer sig om personhenførbarheden (Foster et al. 2016: 306–308). Der er forskel på, om vi deler data, fx studerendes karakterer, fordi de indgår i et gennemsnit for deres årgang, eller om vi deler de enkelte studerendes karakterblad. Gevinsten ved maskinlæring som tilgang er dog størst, når vi har adgang til meget granulære data. Problemet er her, at selv hvis navn, adresse, personnummer mv. fjernes i data, vil sådanne data i større omfang være personhenførbare, fordi det er nemmere at identificere den enkelte studerende i datasættet på grund af detaljegraden, hvis man kender til forhold om den studerende (Foster et al. 2016: 302).

De tre forhold vedrørende følsomheden af data – anonymitet, datasikkerhed og personhenførbarhed – vedrører alle individets ret til at have indflydelse på, hvem der skal vide hvad om vedkommende, og det er således forskellige aspekter af at bestemme, hvad der er privat. Denne ret kan blive udfordret, når maskinlæring anvendes sammen med big data, fordi det bliver sværere og sværere at deltage i hverdagslivet uden at efterlade sig spor som følge af digitale teknologiers medierende rolle (Kitchin 2014b: 167–168). Det kan være spor om forbrug, arbejde, rejser, kommunikation, motion, online-adfærd, interaktioner med myndigheder med videre. Langt mere data er derfor til rådighed om den enkelte, som i fremtiden kan blive gransket og monitoreret og indgå i den offentlige forvaltning. Det udfordrer grænsedragningen mellem den private og offentlige sfære. I forlængelse af denne udvikling er det relevant at diskutere, hvad data bruges til.

Selvom data om et individ bliver genereret i én kontekst, kan samme data blive genbrugt i en anden, potentielt uden individets vidende. Det udfordrer individets ret til at kontrollere, hvad data bruges til (Kitchin 2014b: 171). Det er frafaldsmodellen et godt eksempel på. Modellen er baseret på alt fra data om de studerendes karakterer fra en ungdomsuddannelse, deres adfærd på intranettet, hvor højt de prioriterede deres uddannelse, da de ansøgte samt karakteristika som køn og alder – data, som ikke i sin tid blev genereret til at forudsige frafald og målrette tiltag. Det kan Metropol forsøge at håndtere ved at bede de studerende om samtykke til at bruge deres data på denne måde. Dermed ville data kun blive brugt med de studerendes samtykke. Udfordringen her ligger i betingelserne for, at samtykket rent faktisk kan siges at være informeret, idet der kan være en diskrepans mellem formelt at have accepteret nogle vilkår og at have indgået et faktisk informeret samtykke. En mindste-forudsætning er, at de studerende rent faktisk har læst vilkårene. Dertil kommer, om de også har forstået vilkårene og måske tilmed implikationerne af dem (Foster et al. 2016: 306–308; Kitchin 2014b: 174). Hvis det er forudsætningen for, at et samtykke kan siges at være informeret, er det meget krævende for ikke at sige praktisk umuligt at opnå et sådant (Kitchin 2014b: 174; Nissenbaum 2011; Foster et al. 2016: 307).

Samlende giver brugen af data i frafaldsmodellen som led i Metropols administration anledning til en række refleksioner over, hvordan de studerendes ret til privatliv bevares. Det drejer sig om deres ret til at definere *det private* og dermed have kontrol over, hvem der bruger hvilket data hvornår. Det drejer sig samtidig om at sikre sig, at de studerende er indforståede med, hvordan deres data og personoplysninger behandles og bruges. Det er en hårfin balancegang mellem på den ene side at få mest ud af data i forvaltningsøjemed, hvilket her vil sige at bruge så rige og detaljerede data som muligt for at opnå de bedste prædiktioner – og på den anden side samtidig respektere de studerendes ret til privatliv. Det er en problemstilling i sig selv. Dertil kommer, at monitoreringen af de studerendes data kan have en effekt på de studerendes adfærd. Selve idéen om, at monitorering – eller overvågning, om man vil – har en effekt på adfærden hos de monitorerede subjekter er ikke ny (Foucault 2012; Kitchin 2014b: 179–183; Johnson 2014: 5–7). Det er ikke utænkeligt, at maskinlærings potentiale til omfangsrig monitorering af alle de enkelte studerende kan have forholdsvist vidtgående effekter på de studerendes adfærd. Det er et af emnerne for næste afsnit.

6.4.4 Konstitutive effekter af frafaldsmodellen

Formålet med frafaldsmodellen er som bekendt at målrette tiltag mod frafald, og i den forstand håber og forventer vi, at modellen får en effekt på frafaldet. Anvendelsen af

modellen kan dog også have en række utilsigtede sideeffekter, som vi kan forstå med begrebet *konstitutive effekter* (Dahler-Larsen 2014).

Begrebet dækker over effekter, der følger af selve frafaldsmodellens *klassifikationer* – det forhold at vi sætter et mærkat på de studerende. Herved ordner vi verden i klasser eller typifikationer, som vi bruger til at forstå og skabe mening af verden og til at handle efter (Schutz 2005: 37–51; Dahler-Larsen 2014: 976). På den måde tjener klassifikationerne til at ordne en ellers kompleks og ustruktureret virkelighed ved at aggregere, simplificere og dekontekstualisere et udvalg af kvantificerbare aspekter ved tilværelsen (Dahler-Larsen 2014; Espeland & Sauder 2007; House & Howe 1999). Fordi klasserne ikke er naturgivne, men konstruerede, og fordi det samme gælder for algoritmen, som klassificerer observationerne i et datasæt, følger det, at de resulterende klassifikationer heller ikke er naturgivne, men repræsenterer et partikulært verdenssyn (House & Howe 1999). Det er kontroversielt, fordi klassifikationerne har virkelige og betydningsfulde konsekvenser for de studerende.

Når frafaldsmodellen tages i anvendelse vil en andel af de studerende få mærkatet *frafaldstruet*. Det, at få sådan et mærkat på sig, kan påvirke selvforståelsen hos de studerende (Dahler-Larsen 2014). Eksempelvis kan vi forestille os en situation, hvor en studerende bliver bedt om at møde op hos studievejledningen. I forlængelse af et sådant påbud, fx udsendt per mail, vil det være naturligt at begrunde påbuddet, eksempelvis ved at fortælle den studerende, at vedkommende er identificeret til at være i en særlig risiko for at frafalde, og Metropol derfor gerne vil hjælpe. Men hvad nu hvis den studerende aldrig selv har tænkt den tanke, at der er en risiko for, at vedkommende dropper ud? Måske den studerende aldrig ville være droppet ud, før denne mail blev sendt, men den studerende nu kommer til at overveje muligheden. Herved påvirkes den studerende af selve klassifikationen. Samtidig kan tiltaget tænkes at have en effekt, der ligefrem er modsatrettet intentionen om at nedbringe frafaldet, hvis det får studerende til at frafalde, som ellers ikke ville have gjort det. I så fald bliver klassifikationen som frafaldstruet i et givent omfang en selvopfyldende profeti.

Klassifikationen som frafaldstruet kan videre påvirke den studerendes sociale relationer (Dahler-Larsen 2014). Måske den studerende i større omfang vil føle sig distanceret fra sine medstuderende eller fra studiet i det hele taget, fordi vedkommende har fået et mærkat på sig, som i kraft af at være simpelt og dekontekstualiseret kan fremstå objektivt og autoritativt (Espeland & Sauder 2007: 16–19; Mittelstadt et al. 2016: 5, 9–10). Det er en betydningsfuld konsekvens i sig selv, hvis den studerendes sociale relationer påvirkes af klassifikationen. Derudover vil det ikke være overraskende, hvis eksemplerne nævnt overfor vil trække i retning af en lavere social integration af den

studerende – hvilket kan tænkes at øge frafaldet frem for at mindske det, som vi så i litteraturgennemgangen om årsager til uddannelsesfrafald.

Derudover er det også tænkeligt, at implementeringen af frafaldsmodellen kan ændre de studerendes adfærd. Det kan rent teknisk have konsekvenser for frafaldsmodellens præcision (fordi det ændrer den datagenererende proces), men det er derudover ikke uden etiske implikationer. Det kan være adfærdsændringer i det små, såsom ændret adfærd på intranettet, fordi man føler sig overvåget i en slags digitalt panoptikon og gerne vil at undgå at blive klassificeret som frafaldstruet.

Vi kan også forestille os andre potentielt mere yderliggående adfærdsændringer. Vi kan i denne anledning forestille os et mere radikalt eksempel, hvor det ikke bare er en frafaldsmodel, der implementeres, men snarere en art "uddannelses-Netflix", hvor studerende på ungdomsuddannelser anbefales et mindre udvalg af videregående uddannelser, som de prædikteres at være særligt egnede til baseret på baggrundsinformation om de studerende. Det er ikke ren science fiction. På det amerikanske universitet Austin Peay State University foreslås de studerende eksempelvis, hvilke kurser de bør vælge baseret på de studerendes akademiske meritter (Johnson 2014; Romero & Ventura 2010).

Formålet med en sådan anbefalings-tankegang er at målrette individualiseret information for at hjælpe individet med at nå frem til information, som prædikteres at være relevant. Men beslutningen om, hvilken information der er relevant, er subjektiv, og når individets valgmuligheder på denne måde forsøges påvirket og indskrænket, kan det ses som et udtryk for paternalisme (Johnson 2014). Det undergraver det enkelte individs autonomi, hvis individets valg i større omfang kommer til at afspejle institutionelle præferencer end individets egne (Johnson 2014; Mittelstadt et al. 2016). Det er paradoksalt, når individualiseringen og målretningen af tiltag jo netop har til formål at støtte individet.

En sådan paternalistisk tendens kan også læses ind i frafaldsmodellen i casen Metropol. Sat på spidsen synes rationalet at være, at alle studerende kan passe ind i én af to mulige kategorier, *frafaldstruet* eller *ikke-frafaldstruet*, og at alle de førstnævnte bør fastholdes. Men er det altid ønskværdigt at minimere frafald? Måske nogle studerende bare har valgt en forkert uddannelse. Situationen set fra den studerendes perspektiv inddrages ikke af algoritmen. I stedet målretter denne et tiltag i overensstemmelse med institutionelle præferencer, som også afspejles i udformningen af tiltaget, samt hvor tærskelværdien bliver sat. Set i det lys undermineres individets autonomi og frie vilje – hvad der måske i mindre omfang ville være tilfældet, hvis den studerende selv havde benyttet sig af en åben-dør-politik hos studievejledningen og søgt deres råd.

Frafaldsmodellen kan endvidere siges at underminere det, vi kan kalde individets individualitet. Når vi eksempelvis klassificerer en studerende med en træbaseret algoritme, profilerer vi alle de studerende med ja/nej-spørgsmål for at finde frem til den ene af to mulige frafalds-kategorier, de passer bedst i. Ét er, at denne datadrevne profilering er baseret på et udvalg af kvantificerbare træk ved de studerende. Noget andet er, at når de studerende først er klassificeret, kollapser vi alle deres individuelle profiler og identiteter til ét mærkat, henholdsvis frafaldstruet eller ikke-frafaldstruet. Herefter behandler vi hele denne potentielt vidt forskellige gruppe af studerende ens på basis af deres fælles mærkat (Van Wel & Royakkers 2004: 133). Den individualiserede prædiktion omsættes altså ikke i en individualiseret behandling.

Ved anvendelse af frafaldsmodellen reducerer vi samtidig et komplekst socialt fænomen til et rent teknisk problem, målretning, som vi kan løse og optimere gennem computerisering. Frafaldsmodellen kan derved gøre os blinde for at adressere frafaldsudfordringer, som har dybere strukturelle rødder. Det kunne fx være kønsdebatten centreret om, hvordan folkeskolens indretning er tilpasset henholdsvis drenge og piger. I stedet muliggør frafaldsmodellen blot forvaltning af frafaldsudfordringens mere umiddelbare manifestationer (Johnson 2014). Det kunne fx være sammenhængen mellem holdstørrelse og frafald, som vi fandt i analysen. Derved strukturerer modellen også, hvilke policy-tiltag der præsenterer sig som relevante løsninger på problemet (Dahler-Larsen 2014). Individuelle prædiktioner af frafaldsrisiko kalder fx på individuelle tiltag. Kun individuelle tiltag, såsom et visitationsmøde eller et mentorforløb, passer ind i vores framework til anvendelse af frafaldsmodellen. Vi risikerer dermed at orientere os væk fra årsager og løsninger på kollektivt niveau, fordi de ikke passer ind i modellen. Litteraturen har peget på social og akademisk integration som afgørende for fastholdelse. Måske havde det været mere nyttigt at håndtere frafald som et kollektivt fænomen ved at fokusere på fx en styrkelse af det sociale studiemiljø og flere undervisningstimer.

Med frafaldsmodellen ordner vi altså populationen af studerende i to kategorier, der kan bruges til at strukturere tiltag efter. Selve implementeringen af frafaldsmodellen er derfor politisk, fordi den rationaliserer et problem, frafald, ved at ordne det i klasser, hvilket har betydning for, hvilke aspekter af frafaldsproblematikken, vi ser, hvordan vi fortolker dem, hvordan vi taler om dem, og hvorhen vi dirigerer vores handlinger og ressourcer for at imødekomme dem. Implementeringen af frafaldsmodellen er politisk, fordi modellen bidrager til at konstruere de kategorier, som er kollektivt betydningsfulde i samfundet (Dahler-Larsen 2014).

Kapitel 7

Konklusion

Har maskinlæring potentiale til at målrette tiltag mod uddannelsesfrafald? Sådan spurgte vi i indledningen til denne opgave. Vi har undersøgt problemstillingen med udgangspunkt i et eksplorativt casestudie af Professionshøjskolen Metropol, hvor vi har udviklet en prædiktionsmodel til forudsigelse af frafald. I undersøgelsen har vi overordnet haft to formål for øje. For det første har vi undersøgt potentialet til målretning ud fra metodiske kriterier. Vi fandt, at vores bedste model leverede tilstrækkeligt gode forudsigelser til at danne grundlag for målretning. Det viste vi ved at opstille et framework til målretning, som inddrog stiliserede træk ved Metropols kontekst. Med udgangspunkt i dette framework konkluderede vi, at maskinlæring i vores case har et metodisk potentiale til at målrette tiltag mod uddannelsesfrafald.

For det andet har vi undersøgt potentialet til målretning under hensyn til de dilemmaer, som uvægerligt følger, når en prædiktionsmodel tages i anvendelse i en konkret, administrativ praksis. Her diskuterede vi metodologiske, epistemiske og etiske problematikker, hvormed vi tog maskinlærings praktiske potentiale op til kritisk revision. Med disse refleksioner er vi gået et skridt videre end de eksisterende casestudier om maskinlæring og frafald, der alene forholder sig til det metodiske potentiale i tilgangen. Samtidig har vi bidraget til den spæde politologiske litteratur om emnet ved at gå i dybden med en konkret forvaltningscase, som har givet os mulighed for at konkretisere og nuancere flere af litteraturens kritiske perspektiver.

Maskinlæring som tilgang indebærer et fokusskifte fra estimation til prædiktion. Hvor formålet med estimation er at levere unbiased estimater og maksimere en models in-sample fit, er formålet med prædiktionsmodeller at maksimere performance out-of-sample. Vi diskuterede, hvordan prædiktions-performance kan måles og introducerede

confusion-matricer, ROC-kurver og AUC-værdier som velegnede værktøjer. Opsplitningen af vores datasæt i et trænings- og testsæt tillod os at måle prædiktions-performance direkte. Dermed kunne vi datadrevet frem for teoridrevet vælge funktionel form og modelspecifikation og lægge bånd på modelkompleksiteten med regularisering. Det rette niveau af regularisering kunne vi fastsætte empirisk ved tuning med krydsvalidering. Disse trin gjorde det muligt empirisk at finde frem til en model, som undgår både under- og overfitting til data og dermed maksimerer out-of-sample prædiktions-performance.

Mens vores analyse i store træk var datadrevet, gav den forudgående databehandling anledning til en skepsis over for de røster, der har proklameret teoriens endeligt. Vores analyse er baseret på feature engineering, hvor teori fx har guidet bearbejdningen af variable, herunder håndtering af manglende data. Vi så i analysens afsnit om importance-mål, at mange af vores konstruerede variable havde stor prædiktiv værdi i frafaldsmodellen. Selvom feature engineering hviler på forudsætningen om at maksimere prædiktions-performance frem for at levere kausale estimater, er databehandlingen stadig guidet af teori. Som vi senere diskuterede, er data ikke neutrale eller værdifri, og idet kun numerisk information kan indgå i frafaldsmodellen, har kvantificerbare variable forrang. Det er relevant at huske på for en beslutningstager, der ønsker at målrette et tiltag, at det kun er den viden, som passer ind i modellen, der kan indgå.

I analysen fandt vi, at de træbaserede algoritmer klarede sig bedre end vores baseline-model baseret på logistisk regression. Vores bedste model var baseret på ensemble-metoden Gradient Boosted Trees. Ved at fitte klassifikationstræer sekventielt på fejlen fra det foregående træ i en proces kaldet boosting, opnåede vi en AUC på 0,727 med denne model. Til sammenligning opnåede baseline-modellen en AUC på 0,646. Af to grunde skal den præcise numeriske forskel tages med et gran salt. For det første fordi modellernes AUC er behæftet med en vis usikkerhed, da der indgår sampling undervejs – blandt andet når data splittes i et trænings- og testsæt. For det andet fordi vi har behandlet baseline-modellen som netop en baseline og derfor bevidst undladt at regularisere og tune den, selvom vi forventer, at det kunne have forbedret modellen. Da vi tunede GBT-modellen med bayesiansk optimering hævede vi dens AUC fra 0,702 til 0,727.

Da vi tog Gradient Boosted Trees-modellen og baseline-modellen i anvendelse i vores framework til målretning, illustrerede vi den praktiske betydning af forskellen i deres performance. Her viste førstnævnte et markant større potentiale til at målrette tiltag, hvad enten formålet var at maksimere Metropols nettofortjeneste eller minimere frafaldet ved at maksimere antallet af studerende, som kunne tilbydes tiltaget. I samme del af analysen viste vi, hvordan viden om konteksten kan benyttes til at fastsætte en tærskelværdi for modellens forudsigelser med det formål at indfri en given politisk

målsætning, såsom at minimere frafaldet. Vores framework til at benytte en prædiktionsmodel til målretning forudsætter imidlertid, at vi dels har kendskab til effekten af tiltaget, som målrettes, dels har kendskab til kvantitative mål for fordele og ulemper ved målretningen. Det er nogle forholdsvist krævende rammebetingelser, som illustrer, at der kan være ganske langt fra at omsætte prædiktioner til et håndfast handlingsgrundlag. Det taler for et vist mådehold i forventningen om potentialet til målretning, hvilket er en indsigt, der rækker ud over både Metropol og uddannelsesområdet.

I forlængelse heraf diskuterede vi, om vores konklusioner er valide, i den forstand at vi kan forvente samme performance rent metodisk, når frafaldsmodellen anvendes i praksis. På den ene side giver vores databehandling grund til at se potentialet som et konservativt estimat. På den anden side giver risikoen for målefejl anledning til skepsis. Problemet er, at modellen bygger på en antagelse om et stabilt miljø, hvis vores testsæt i analysen skal være en troværdig tilnærmelse af det "virkelige" testsæt, dvs. de fremtidige studerende. Hvis frafaldsmønstrene i de to testsæt ikke er de samme vil vores prædiktions-performance være misvisende. Her kan ændrede målefejl over tid være en kilde til ændrede mønstre, hvilket andre eksterne årsager også kan – såsom ændringer i studieordninger, SU-regler og reformer af national uddannelsespolitik. Det er en udfordring at fundere fremtidige policies på fortidige og foranderlige mønstre. Det er dog en kritik, som vedrører statistiske modeller generelt. Ved at anvende en model baseret på maskinlæring, som tilpasser sig nye mønstre, er vi sådan set bedre stillet end med en klassisk statisk model.

En relateret problematik opstår dog, fordi modellens forudsigelser bruges til at målrette policy-tiltag, hvorved den *selv* kan påvirke mønstrene i data. Det er paradoksalt, at modellen kan undergrave sit eget fundament og potentiale, hvis den virker efter hensigten. Her bør vi dog huske på, at vi kun forventer en relativt begrænset effekt på frafaldsmønstrene af de tiltag, som vi målretter. Vi forventer, at frafaldsmønstrene er relativt træge over tid. Og hvis tiltagene viser sig at være mere effektive til at hindre frafald, end vi havde forventet, vil studievejledningen trods alt næppe modtage det som dårligt nyt.

Vi diskuterede endvidere selve præmissen for vores problemstilling om målretning: at forskelsbehandling i forvaltningen kan retfærdiggøres. Vi argumenterede for, at statistisk diskrimination er en forsømmelse, der kommer i grader, og at en prædiktionsmodel baseret på maskinlæring i denne henseende stiller os bedre end menneskelig sagsbehandling. Hvor vi med vanlige statistiske modeller kollektiviserer og behandler alle ens ud fra gennemsnitlige effekter, kan vi med maskinlæring i teorien inddrage så meget information om den enkelte, at beslutningen ikke baseres på simple gruppetilhørsforhold.

Forskningen er rig med eksempler på, hvordan den menneskelige kognition i denne henseende må se sig slået af maskinen.

Der kan dog være andre grunde til at foretrække menneskelig sagsbehandling. Vi argumenterede blandt andet for, at maskinlæring med en *black box* af evigt foranderlige beslutningsregler gør forudsigelserne uigennemskuelige. Hensynet til transparens udspringer af borgeres legitime krav på indsigt i sammenhængen mellem beslutningsgrundlag og handling i en forvaltning baseret på demokratiske principper. Her støder vi på en afvejning mellem transparens og fortolkbarhed på den ene side, og modelkompleksitet og prædiktions-performance på den anden. Det er fx ikke sikkert, at GBT-modellens gode prædiktions-performance kan kompensere for uigennemsigtigheden af dens meget komplekse funktionsmåde. Behovet for transparens aktualiseres yderligere i den epistemiske anerkendelse af, at en prædiktionsmodel er fejlbarlig og ikke er baseret på et neutralt beslutningsgrundlag. Hvor sikre skal vi være på en prædiktion, før denne kan omsættes til handling? Eksternalisering af sådanne beslutninger fra den menneskelige sagsbehandling, kan ses som en reaktion på øget efterspørgsel på evidens. Denne udvikling udfordrer hensynet til accountability, når beslutninger om at iværksætte tiltag mod fx frafald som i vores case flyttes fra studieadministrationen til en algoritme. Disse overvejelser stiller betydeligt spørgsmålstegn ved maskinlærings potentiale som beslutningsgrundlag i forvaltningen i praksis.

Modellerne bygger endvidere på en relativt vidtrækkende brug af personoplysninger, som udfordrer hensynet til de studerendes ret til privatliv. Prædiktionsmodeller baseret på maskinlæring giver mulighed for omfangsrig monitorering, som også kan virke tilbage på de monitorerede og have en række utilsigtede, konstitutive effekter. Det var afsættet for de afsluttende etiske refleksioner. Her var vores argument, at selve klassifikationen som frafaldstruet kan føre til selvopfyldende profetier. Mere generelt problematiserede vi, at modellen strukturerer, hvilke problemer vi ser – og hvilke vi overser – og hvordan vi kan handle på dem. Desuden er det spørgsmålet, om det altid er ønskværdigt at fastholde de studerende. Vi argumenterede for, at frafaldsmodellen kan undergrave den enkeltes beslutningsautonomi, da modellen ikke inddrager de studerendes synspunkter, men alene udtrykker institutionelle præferencer. Det er paradoksalt, når modellen netop er sat i verden for at målrette et tiltag mod dem, der har mest behov for det. Det problematiserer endeligt, at vi med frafaldsmodellen ganske vist laver individuelle prædiktioner, men ikke formår at omsætte disse til individualiserede tiltag. Dertil mangler vi blandt andet viden om heterogene effekter af mulige tiltag.

Med opgaven har vi lagt os i forlængelse af en række tidligere studier, der har forsøgt at forudsige frafald med prædiktionsmodeller. Studierne kommer fra en datalogisk tradition, hvor fokus er på modellernes forudsigelseskraft. Der bliver ikke stillet spørgs-

målstegn ved, om forudsigelserne rent faktisk kan omsættes til handling med henblik på at indfri en politisk målsætning. Vores bidrag med opgaven har været at opstille et framework for, hvordan prædiktionsmodellerne kan tages i anvendelse i praksis, og reflektere over den slipstrøm af konsekvenser, en sådan model kan have. I en tid hvor maskinlæring som tilgang bevæger sig fra datalogien ind i samfundsvidenskaben, ser vi det som en god anledning til at forholde sig kritisk til potentialet såvel som faldgruberne. Det gælder det metodiske potentiale ved at fokusere på prædiktion frem for estimation og det praktiske potentiale til målretning af tiltag set i lyset af de konsekvenser, der kan følge, når en prædiktionsmodel anvendes i en konkret, administrativ praksis.

Vi finder samlende, at maskinlæring i vores case kan levere prædiktioner, som er tilstrækkeligt præcise til, at de kan danne grundlag for at målrette tiltag mod uddannelsesfrafald. Set i lyset af den efterfølgende diskussionen, har vi dog svært ved at se, at en prædiktionsmodel skulle kunne stå alene. Vores betænkelighed er tofoldig. For det første – og særligt i relation til den offentlige forvaltning – er det tvivlsomt, at beslutningen om at målrette et tiltag kan stå helt uden menneskelig bistand af hensyn til blandt andet transparens, accountability og modellers fejlbarlighed. For det andet – og i relation til samfundsvidenskaben mere generelt – har vi svært ved at se, at maskinlæring som tilgang kan stå alene.

Der er nemlig et stykke vej fra at lave en forudsigelse til at omsætte den til handling – et stykke vej, som er svært at tilbagelægge uden kausalestimation. Set i forhold til den *credibility revolution*, som er udbredt i den empiriske samfundsvidenskab, er det derfor ikke entydigt, at maskinlæring er et skridt fremad. Vi vil dog heller ikke sige, at tilgangen er et skridt tilbage, men finder det mere produktivt at se maskinlæring som et skridt til siden. Med et skridt til siden giver tilgangen en anledning til at kaste et blik på estimationstilgangen udefra og bidrage konstruktivt i en anerkendelse af tilgangenes komplementaritet. Nogle policy-problemer kan med fordel anskues som prædiktionsproblemer. Forudsætningen er blot, at man kender effekterne af de tiltag, man ønsker at målrette. Derfor levner vores opgave også rum til videre undersøgelser – eksempelvis et effektstudie af tiltagene, som Metropol håber, kan fastholde de frafaldstruede studerende. Sådan en undersøgelse ville være en oplagt følgesvend på den sidste strækning fra forudsigelse til fastholdelse.

Litteratur

Agan, A. Y., & Starr, S. B. (2016). Ban the box, criminal records, and statistical discrimination: A field experiment. *U of Michigan Law & Econ Research Paper No. 16-012*.

Agresti, A., & Finlay, B. (2009). *Statistical Methods for the Social Sciences*. Pearson International.

Andersen, J. (2014). *Forvaltningsret*. Karnov Group.

Andersen, L. B., Hansen, K. M., & Klemmensen, R. (2010). *Metoder i statskundskab*. Rosinante&Co.

Anderson, C. (2008). *The end of theory: The data deluge makes the scientific method obsolete*. Besøgt den 23-06-2017 på: https://www.wired.com/2008/06/pb-theory/.

Angrist, J. D., & Pischke, J.-S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *The Journal of economic perspectives*, *24*(2), 3–30.

Angrist, J. D., & Pischke, J.-S. (2015). *Mastering 'metrics: The path from cause to effect*. Princeton University Press.

Athey, S. (2017). Beyond prediction: Using big data for policy problems. *Science*, *355*(6324), 483–485.

Athey, S., & Imbens, G. (2016). The state of applied econometrics-causality and policy evaluation. *arXiv: 1607.00699*.

Aulck, L. S., Velagapudi, N., Blumenstock, J. E., & West, J. (2016). Predicting student dropout in higher education. *arXiv: 1606.06364*.

Barocas, S. (2014). Data mining and the discourse on discrimination. In *Data Ethics Workshop, Conference on Knowledge Discovery and Data Mining*.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, *104*(3), 671.

Bayer, J., Bydžovská, H., Géryk, J., Obšívač, T., & Popelínský, L. (2012). Predicting drop-out from social behaviour of students. In *Proceedings of the 5th International Conference on Educational Data Mining – EDM 2012*, (pp. 103–109).

Berk, R. (2012). *Criminal justice forecasts of risk: A machine learning approach*. Springer Science & Business Media.

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. *Robustness in statistics*, *1*, 201–236.

boyd, d., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, *15*(5), 662–679.

Braxton, J. M. (2000). *Reworking the student departure puzzle*. Vanderbilt University Press.

Braxton, J. M., Sullivan, A. V. S., & Johnson, R. M. (1997). Appraising Tinto's theory of college student departure. In J. C. Smart (Ed.) *Higher Education: Handbook of Theory and Research 12*, (pp. 107–164). Springer Science & Business Media.

Breiman, L., et al. (2001). Statistical modeling: The two cultures. *Statistical science*, *16*(3), 199–215.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1).

Burrows, R., & Savage, M. (2014). After the crisis? Big Data and the methodological challenges of empirical sociology. *Big Data & Society*, *1*(1).

Cederman, L.-E., & Weidmann, N. B. (2017). Predicting armed conflict: Time to adjust our expectations? *Science*, *355*(6324), 474–476.

Chandler, D., Levitt, S. D., & List, J. A. (2011). Predicting and preventing shootings among at-risk youth. *The American Economic Review*, *101*(3), 288–292.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Chen, T. (2014). *Introduction to Boosted Trees*. Besøgt den 03-07-2017: http://homes.cs.washington.edu/ tqchen/pdf/BoostedTree.pdf.

Chen, T., He, T., & Benesty, M. (2016). *xgboost: Extreme Gradient Boosting*. R package version 0.4-4. URL: https://CRAN.R-project.org/package=xgboost.

Clark, W. R., & Golder, M. (2015). Big data, causal inference, and formal theory: Contradictory trends in political science?: Introduction. *PS: Political Science & Politics*, *48*(1), 65–70.

Dahler-Larsen, P. (2011). *The evaluation society*. Stanford University Press.

Dahler-Larsen, P. (2014). Constitutive effects of performance indicators: getting beyond unintended consequences. *Public Management Review*, *16*(7), 969–986.

Datatilsynet (2015). *Kort om persondataloven*. Besøgt den 13-07-2017 på: https://www.datatilsynet.dk/offentlig/kort-om-persondataloven/.

De Mauro, A., Greco, M., & Grimaldi, M. (2016). A formal definition of Big Data based on its essential features. *Library Review*, *65*(3), 122–135.

Dekker, G., Pechenizkiy, M., & Vleeshouwers, J. (2009). Predicting students drop out: A case study. In *Proceedings of the 2nd International Conference on Educational Data Mining*, (pp. 41–50).

Desrosières, A. (1998). *The Politics of Large Numbers: A History of Statistical Reasoning*. Harvard University Press.

Ding, Y., & Simonoff, J. S. (2010). An investigation of missing data methods for classification trees applied to binary response data. *Journal of Machine Learning Research*, *11*(1), 131–170.

DMLC (2016a). *Introduction to Boosted Trees*. Besøgt den 22-06-2017 på: http://xgboost.readthedocs.io/en/latest/model.html.

DMLC (2016b). *XGBoost Parameters*. Besøgt den 22-06-2017 på: http://xgboost.readthedocs.io/en/latest/parameter.html.

Einav, L., & Levin, J. (2014). The data revolution and economic analysis. *Innovation Policy and the Economy*, *14*(1), 1–24.

Espeland, W. N., & Sauder, M. (2007). Rankings and reactivity: How public measures recreate social worlds. *American journal of sociology*, *113*(1), 1–40.

EVA (2017). *Studiestartens betydning for frafald på videregående uddannelser*. Danmarks Evalueringsinstitut.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, *27*(8), 861–874.

Flyvbjerg, B. (2006). Five misunderstandings about case-study research. *Qualitative inquiry*, *12*(2), 219–245.

Foster, I., Ghani, R., Jarmin, R. S., Kreuter, F., & Lane, J. (2016). *Big Data and Social Science*. Chapman and Hall/CRC.

Foucault, M. (2007). *Security, territory, population: lectures at the Collège de France, 1977-78*. Springer.

Foucault, M. (2012). *Discipline & punish: The birth of the prison*. Vintage.

Friedman, J., Hastie, T., & Tibshirani, R. (2009). *The elements of statistical learning*. Springer.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, (pp. 1189–1232).

Fuller, W. A. (2009). *Measurement error models*. John Wiley & Sons.

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*.

Gerber, A. S., & Green, D. P. (2012). *Field experiments: Design, analysis, and interpretation*. WW Norton.

Gerring, J. (2004). What is a case study and what is it good for? *American Political Science Review*, *98*(2), 341–354.

Goet, N. (2016). *What big data can teach political scientists*. Besøgt den 23-06-2017 på: https://blog.politics.ox.ac.uk/big-data-can-teach-political-scientists/.

Grimmer, J. (2015). We are all social scientists now: how big data, machine learning, and causal inference work together. *PS: Political Science & Politics*, *48*(1), 80–83.

Hansen, D. (2013). 'Dameklip' bliver nu forbudt hos frisøren. *Politiken*. Besøgt den 25-04-2017 på: http://politiken.dk/forbrugogliv/forbrug/forbrugersikkerhed/art5435482/Dameklip-bliver-nu-forbudt-hos-frisøren.

Harmsen, O., & Enggaard, T. R. (2016). *Using machine learning to target policy interventions*. (Kandidatspeciale). Tilgået på: https://diskurs.kb.dk/item/diskurs:96347:1.

Herzog, S. (2006). Estimating student retention and degree-completion time: Decision trees and neural networks vis-à-vis regression. *New directions for institutional research*, *2006*(131), 17–33.

Hofman, J. M., Sharma, A., & Watts, D. J. (2017). Prediction and explanation in social systems. *Science*, *355*(6324), 486–488.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, *81*(396), 945–960.

House, E., & Howe, K. R. (1999). *Values in evaluation and social research*. Sage Publications.

Imai, K. (2017). *Quantitative Social Science: An Introduction*. Princeton University Press.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

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

Johnson, J. A. (2014). The ethics of big data in higher education. *International Review of Information Ethics*, *21*(1), 3–10.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kang, J. S., Kuznetsova, P., Luca, M., & Choi, Y. (2013). Where *not* to eat? Improving public policy by predicting hygiene inspections using online reviews. In *EMNLP*, (pp. 1443–1448).

Kapelner, A., & Bleich, J. (2015). Prediction with missing data via bayesian additive regression trees. *Canadian Journal of Statistics*, *43*(2), 224–239.

King, G., Honaker, J., Joseph, A., & Scheve, K. (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American Political Science Review*, *95*(1), 49–69.

King, G., Keohane, R. O., & Verba, S. (1994). *Designing social inquiry: Scientific inference in qualitative research*. Princeton University Press.

Kitchin, R. (2014a). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1).

Kitchin, R. (2014b). *The data revolution: Big data, open data, data infrastructures and their consequences*. Sage Publications.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). *Human decisions and machine predictions*. Working paper no. 23180. National Bureau of Economic Research.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review*, *105*(5), 491–495.

Kotsiantis, S. B., Pierrakeas, C. J., & Pintelas, P. E. (2003). Preventing student dropout in distance learning using machine learning techniques. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, (pp. 267–274). Springer.

Kovačić, Z. J. (2010). Early prediction of student success: Mining student enrollment data. In *Proceedings of Informing Science & IT Education Conference*.

Kristoffersen, J. (2015). *School and University Dropout Prediction revisited*. (Kandidatspeciale). Tilgået via forespørgsel hos forfatteren.

Kulager, F. (2016). Politiet har opdaget, at data kan spå. nu kan de forudse forbrydelser. *Zetland*. Besøgt den 25-04-2017 på: https://www.zetland.dk/historie/s81PmbGv-aOZj67pz-50e53.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of google flu: traps in big data analysis. *Science*, *343*(6176), 1203–1205.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomforest. *R News*, *2*(3), 18–22.

Lippert-Rasmussen, K. (2011). "We are all different": Statistical discrimination and the right to be treated as an individual. *The Journal of ethics*, *15*(1-2), 47–59.

Little, R. J. A., & Rubin, D. B. (2014). *Statistical analysis with missing data*. John Wiley & Sons.

Lowe, W., & Benoit, K. (2013). Validating estimates of latent traits from textual data using human judgment as a benchmark. *Political Analysis*, *21*(3), 298–313.

Lucas, R. E. (1976). Econometric policy evaluation: A critique. In *Carnegie-Rochester conference series on public policy*, vol. 1, (pp. 19–46). Elsevier.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data, a revolution that will transform how we live, work and think*. John Murray.

Metropol (2017). *Årsrapport 2016*. Besøgt den 18-07-2017 på: https://www.phmetropol.dk/om+metropol/tal+og+fakta/aarsrapport.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2).

Mollineda, R., Alejo, R., & Sotoca, J. (2007). The class imbalance problem in pattern classification and learning. In *II Congreso Español de Informática (CEDI 2007). ISBN*, (pp. 978–84).

Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87–106.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Nissenbaum, H. (2011). A contextual approach to privacy online. *Daedalus*, *140*(4), 32–48.

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future – big data, machine learning, and clinical medicine. *The New England journal of medicine*, *375*(13), 1216–1219.

Pascarella, E. T., & Terenzini, P. T. (1980). Predicting freshman persistence and voluntary dropout decisions from a theoretical model. *The Journal of Higher Education*, *51*(1), 60–75.

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press.

Perry, W. L., McInnis, B., Price, C. C., Smith, S. C., & Hollywood, J. S. (2013). *Predictive policing: The role of crime forecasting in law enforcement operations*. Rand Corporation.

Porter, T. M. (1996). *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton University Press.

Power, M. (1997). From risk society to audit society. *Soziale systeme*, *3*(1), 3–21.

Public Perspectives (2016). *SKATs datastrategi styrker kundeoplevelsen og finder skatteunddragerne*. Besøgt den 06-07-2017 på: http://publicperspectives.dk/skats-datastrategi-styrker-kundeoplevelsen-og-finder-skatteunddragerne/.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. URL: https://www.R-project.org/.

Retsinformation (2000). *Lov om behandling af personoplysninger* (lov nr 429 af 31/05/2000). Besøgt den 14-07-2017 på: https://www.retsinformation.dk/forms/r0710.aspx?id=828.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 1135–1144).

Rieder, G., & Simon, J. (2016). Datatrust: Or, the political quest for numerical evidence and the epistemologies of Big Data. *Big Data & Society*, *3*(1).

Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *40*(6), 601–618.

Rose, N. (1991). Governing by numbers: Figuring out democracy. *Accounting, Organizations and Society*, *16*(7), 673–692.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

Rumberger, R. W., & Lim, S. A. (2008). Why students drop out of school: A review of 25 years of research. *California Dropout Research Project Report*.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, *3*(3), 210–229.

Şara, N.-B. (2014). *School drop out prediction*. (Kandidatspeciale). Tilgået via forespørgsel hos forfatteren.

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological methods*, *7*(2), 147–177.

Schutz, A. (2005). *Hverdagslivets sociologi*. Hans Reitzel.

Schwarz, D., Traber, D., & Benoit, K. (2015). Estimating intra-party preferences: comparing speeches to votes. *Political Science Research and Methods*, *5*(2), 1–18.

Seawright, J., & Gerring, J. (2008). Case selection techniques in case study research. *Political Research Quarterly*, *61*(2), 294–308.

Silver, N. (2012). *The signal and the noise: why so many predictions fail–but some don't*. Penguin.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, *22*(11), 1359–1366.

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.) *Advances in Neural Information Processing Systems 25*, (pp. 2951–2959). Curran Associates, Inc.

Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, *50*(3), 755–769.

Tetlock, P. E., & Gardner, D. (2016). *Superforecasting: The art and science of prediction*. Random House.

Therneau, T., Atkinson, B., & Ripley, B. (2015). *rpart: Recursive Partitioning and Regression Trees*. URL: https://CRAN.R-project.org/package=rpart.

Tinto, V. (1993). *Leaving college: Rethinking the causes and cures of student attrition.*. University of Chicago Press.

Tinto, V. (2012). *Completing College: Rethinking Institutional Action*. University of Chicago Press.

Troelsen, R. (2011). Frafald på de videregående uddannelser – hvad ved vi om årsagerne? *Dansk Universitetspædagogisk Tidsskrift*, *6*(10), 37–44.

Tversky, A., & Kahneman, D. (1975). Judgment under uncertainty: Heuristics and biases. In *Utility, probability, and human decision making*, (pp. 141–162). Springer.

Twala, B., Jones, M., & Hand, D. J. (2008). Good methods for coping with missing data in decision trees. *Pattern Recognition Letters*, *29*(7), 950–956.

Uddannelses- og Forskningsministeriet (2017). *Tilskud*. Besøgt den 20-07-2017 på: http://ufm.dk/uddannelse-og-institutioner/videregaende-uddannelse/professionshojskoler/okonomi/tilskud.

Van Wel, L., & Royakkers, L. (2004). Ethical issues in web data mining. *Ethics and Information Technology*, *6*(2), 129–140.

Varian, H. R. (2014). Big data: New tricks for econometrics. *The Journal of Economic Perspectives*, *28*(2), 3–27.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Wooldridge, J. M. (2009). *Introductory Econometrics: A Modern Approach*. South-Western Cengage Learning.

Yan, Y. (2016). *rBayesianOptimization: Bayesian Optimization of Hyperparameters*. R package version 1.1.0. URL: https://CRAN.R-project.org/package=rBayesianOptimization.