politica, 51. årg. nr. 2 2019, 168-186

Alexander Bach, Jesper Svejgaard og Frederik Hjorth

Maskinlæring som politologisk værktøj

Maskinlæring er en metodisk tilgang til databehandling, som vinder indpas i den politologiske forskning og offentlige forvaltning. Her har tilgangen et lovende potentiale til at lave forudsigelser om eksempelvis brugeres og borgeres senere adfærd, hvilket blandt andet kan bruges til målretning af tidlige indsatser. Men hvad er maskinlæring mere konkret, og hvordan anvender man maskinlæring i praksis? I artiklen introducerer vi kernebegreber i relation til maskinlæring. Vi introducerer maskinlæringsalgoritmer i form af klassifikationstræer. Artiklens pointer illustrerer vi undervejs med et konkret eksempel på anvendelse af maskinlæring i dansk offentlig forvaltning, hvor maskinlæring bliver brugt til at forudsige uddannelsesfrafald på Københavns Professionshøjskole. Afslutningsvist diskuterer vi metodiske styrker og svagheder ved maskinlæring i en samfundsvidenskabelig kontekst.

Den empiriske samfundsvidenskab har historisk eftersøgt forklaringer frem for forudsigelser (Breiman, 2001; Hofman, Sharma og Watts, 2017). Som politologer er vi således vant til at estimere, hvilken kausal effekt en given variabel har på et outcome. Det er derimod sjældent, at vi i stedet vender undersøgelsens fokus og forsøger at forudsige et outcome ud fra de variable, vi har til rådighed. I denne artikel vil vi give en introduktion til, hvordan maskinlæring kan anvendes til prædiktion i en politologisk kontekst.

For beslutningstagere er prædiktion interessant, fordi det åbner mulighed for at målrette politiske tiltag og håndtere givne hændelser, allerede før de finder sted. Det findes der eksempler på inden for mange grene af den offentlige forvaltning. På socialområdet er indsatser forsøgt målrettet særligt udsatte unge (Chandler, Levitt og List, 2011), og på sundhedsområdet kan forudsigelse af komplikationer anvendes til at prioritere dyre operationer (Obermeyer og Emanuel, 2016; Kleinberg et al., 2015). Et eksempel fra det retslige område er det amerikanske politis forudsigelse af kriminalitet for at kunne målrette proaktiv patruljering (Perry et al., 2013), ligesom det amerikanske retsvæsen prædikterer indsattes risiko for at begå ny kriminalitet for at støtte dommeres beslutninger om varetægtsfængsling og prøveløsladelse (Berk, 2012; Kleinberg et al., 2017).

På uddannelsesområdet har der i mange år været fokus på at mindske frafaldet blandt studerende på de videregående uddannelser. Det gennemgående eksempel i denne artikel er fra Københavns Professionshøjskole, som har ekspe-

rimenteret med at forudsige frafaldsrisiko for at kunne målrette fastholdende indsatser (Back og Svejgaard, 2017).

Når det analytiske formål ændres fra estimation af effekter til prædiktion af outcomes, har det implikationer for håndteringen af data – herunder hvilke værktøjer analytikeren har til rådighed. Mens der findes en veludbygget metodisk litteratur om kausalestimation, er litteraturen om prædiktion og redskaber til prædiktion anderledes sparsom. Det er denne knaphed på politologisk tilgængelig litteratur, som vi ønsker at imødegå med artiklen. Med støtte i vores gennemgående eksempel vil vi beskrive de metodiske implikationer af at beskæftige sig med prædiktion frem for estimation. I særdeleshed vil vi fokusere på, hvordan et prædiktionsproblem kan håndteres med værktøjer, der falder inden for betegnelsen *maskinlæring*.

Begrebet maskinlæring stammer oprindeligt fra datalogien, men anvendes i dag i mange forskellige sammenhænge og med en vis definitorisk flertydighed. Maskinlæring betyder grundlæggende, at en computer benytter en algoritme til at finde mønstre i data – og dermed selv når frem til en model, der beskriver, hvordan variablene i et datasæt hænger sammen (Athey og Imbens, 2016; Samuel, 1959; Varian, 2014).

I samfundsvidenskaben er der en voksende interesse for maskinlæring til prædiktionsformål (Mullainathan og Spiess, 2017). De førnævnte case-studier er alle eksempler på, at maskinlæring bliver anvendt til prædiktion og målretning af tiltag i den offentlige forvaltning. I vores eksempel fra Københavns Professionshøjskole går maskinlæring konkret ud på at identificere frafaldsmønstre i data om tidligere studerende med det formål at prædiktere frafaldsrisiko blandt nuværende og kommende studerende.

Denne anvendelse af maskinlæring kaldes *superviseret* maskinlæring. I superviseret maskinlæring kender vi en afhængig variabel af interesse, et *outcome*, typisk noteret y, som vi ønsker at prædiktere. I vores case er outcome frafaldsrisiko. Med superviseret maskinlæring ønsker vi at finde frem til den funktion f, der bedst kan prædiktere y på baggrund af en række variable. Formelt er vores model derfor: $\hat{y} = f(\mathbf{X})$. Her er $\hat{y}$ vores forudsagte y, og målet er, at $\hat{y}$ skal være så god en tilnærmelse af y som muligt (James et al., 2013). Læringselementet i maskinlæring består i, at modellerne selv "lærer" mønstrene i et datasæt. Det vil sige, at vi ikke på forhånd definerer funktionen f ved at udvælge variable og teoretisk specificere, hvordan vi antager, at variablene i et datasæt hænger sammen. I stedet lader vi en algoritme finde frem til den funktion f, hvor $\hat{y}$ er den bedst mulige tilnærmelse af y. Læringselementet består endvidere i, at modellerne over tid selv kan tilpasse sig nye data.

Der findes et utal af forskellige algoritmer, som anvendes til maskinlæring – herunder velkendte redskaber fra den politologiske værktøjskasse såsom lineær og logistisk regression. Rent teknisk ligger der således ikke nødvendigvis et nybrud i begrebet maskinlæring. Når vi bedriver prædiktion frem for estimation er andre og mere komplicerede algoritmer imidlertid ofte nyttige, såsom neurale netværk og klassifikationstræer. De er særligt egnede til at finde komplekse og ikke-lineære mønstre i højdimensionale data.

I artiklens første hovedafsnit introducerer vi til kernebegreber inden for maskinlæring. Fokus er på, hvad der gør maskinlæring særligt velegnet til prædiktion, og hvordan prædiktion adskiller sig fra kausalestimation. I andet hovedafsnit giver vi en introduktion til klassifikationstræer, som udgør byggestenene i mange populære algoritmer og har en særlig appel i samfundsvidenskaben, idet de kan modellere data fleksibelt og alligevel underlægges en intuitiv fortolkning (Montgomery og Olivella, 2018). I det tredje afsnit vender vi os mod metodiske udfordringer ved prædiktion, før vi afslutter artiklen med nogle etiske perspektiver.

Kernebegreber i maskinlæring

Formålet med kvantitative empiriske studier i samfundsvidenskaben er ofte at etablere kausalsammenhænge og estimere kausale effekter (Angrist og Pischke, 2015; Hariri, 2012). Det gøres ved at undersøge kontrafakta gennem sammenligninger under en alt andet lige-betragtning. Denne estimation af kausale effekter vil vi her omtale *estimation* og stille over for *prædiktion*, der anderledes handler om at forudsige et outcome. I en klassisk regressionsanalyse forsøger vi typisk at estimere kausale effekter af de uafhængige variable på et outcome. Det kunne fx være ved at fitte en model med en specifik funktionel form til et datasæt i en situation med eksogen variation i de uafhængige variable (Samii, 2016). I prædiktion er fokus ikke længere på estimation af selve parametrene, men derimod forudsigelse af outcome. I vores eksempel fra Københavns Professionshøjskole er formålet ikke at estimere betydningen af fx køn og alder på frafald. Formålet er så præcist som muligt at kunne prædiktere den enkelte studerendes risiko for frafald, hvilket så igen har til formål at blive i stand til bedre at pege på, hvem der frafalder, for at kunne målrette tiltag mod de mest frafaldstruede studerende.

I en politologisk kontekst har fokusskiftet fra estimation til prædiktion først og fremmest betydning for den måde, vi behandler data på. I afsnittet her præsenterer vi en række kernebegreber i maskinlæring, som følger af dette fokusskifte. En del af afsnittet trækker på James et al. (2013), en glimrende lærebogsintroduktion til maskinlæring. Shalev-Schwartz og Ben-David (2014), Lantz

(2015) og Conway og White (2012) er udmærkede alternative lærebogsintroduktioner til maskinlæring. For en mere teknisk og dybdegående behandling henviser vi til Hastie, Tibshirani og Friedman (2009).

Træningssæt og testsæt

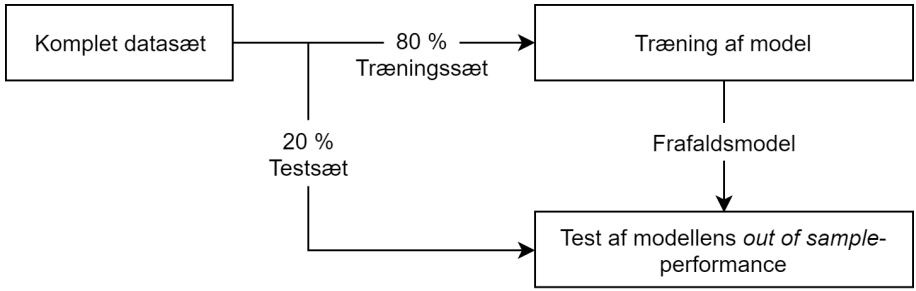
En central forskel mellem kausalestimation og prædiktion består i, at prædiktions-performance kan måles, mens estimations-performance ikke kan. Det fundamentale problem i kausal inferens består i, at vi aldrig får adgang til den kontrafaktiske situation og derfor ikke kan måle, hvor god en estimation vores model er af "virkeligheden" (Rosenbaum og Rubin, 1983). Kausal estimation hviler derfor til syvende og sidst på en række antagelser, der skal underbygge, at alt andet er lige, og at estimaterne af teoretisk interesse i modellen derfor er unbiased.

Prædiktionsmodeller er derimod baseret på korrelationer, som vi kan teste empirisk (James et al., 2013: 29-33). Fordi vi ikke er interesserede i kausalitet, behøver vi ikke gøre os videre antagelser om eventuelle årsagssammenhænge. Man kan teste en prædiktionsmodels forudsigelser ved at fitte – eller "træne", som det hedder inden for maskinlæring – en model på kendte, historiske data og teste den på nye data, der ikke indgik i træningen (James et al., 2013). I vores eksempel kan vi fx udvikle en model, som finder mønstre i data om tidligere studerendes frafald, og derpå bede modellen om at komme med forudsigelser om det fremtidige frafald hos en ny årgang af studerende. Efter et par år kan vi se, om forudsigelserne holdt stik.

Typisk giver det dog ikke praktisk mening at vente så længe med at teste modellens performance. Derfor deler vi i stedet det eksisterende datasæt op i to dele: et træningssæt og et testsæt. Opdelingen kan foregå ved en tilfældig opsplitning af det fulde datasæt, fx 80 pct. som træningssæt og de resterende 20 pct. som testsæt. Den kan også være intentionel – ved tidsserier kan det fx være meningsfuldt at bruge det nyeste år som testsæt og de forudgående år som træningssæt for at afprøve modellens robusthed over tid. Testsættet holder vi helt uden for den proces, hvorigennem vi fitter vores model. Dermed behandler vi testsættet som *out-of-sample*. Vi omtaler dette testsæt som et *holdout-sæt.* Vi kan nu træne modellen på de 80 pct. af data (træningssættet) og teste modellens prædiktions-performance out-of-sample på de resterende 20 pct. (holdout-sættet) (James et al., 2013: 176-178). Logikken er illustreret i figur 1.

I vores eksempel kender vi det faktiske frafald i holdout-sættet, og det udnytter vi til at sammenligne med modellens prædiktioner og beregne modellens prædiktions-performance. Den underliggende antagelse er, at modellens performance i holdout-sættet vil svare til den performance, vi kan forvente, når

Figur 1: Opsplitning af data i et trænings- og testsæt.



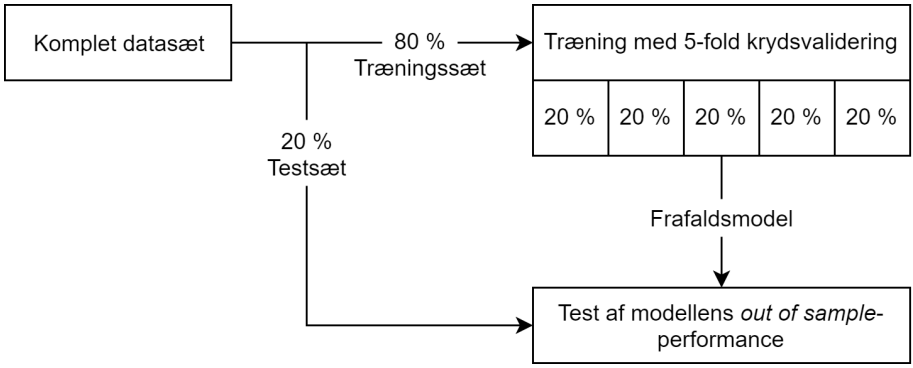
vi bruger modellen til at prædiktere på data for en ny årgang af studerende, som vi jo af gode grunde ikke kender udfaldene for endnu. Vi får altså et mål for modellens forventede prædiktions-performance out-of-sample ved at lave prædiktioner for holdout-sættet (James et al., 2013: 176-178).

k-fold krydsvalidering

Når vi validerer vores model ved at teste den på holdout-sættet, kan vi kun validere vores model én gang. Når vi først har brugt holdout-sættet til validering kan vi ikke gå tilbage og justere vores model for så at validere modellen på samme holdout-sæt igen. I så fald ville data ikke længere være holdt helt og aldeles ude af den proces, hvori vi træner vores model. Ofte vil vi imidlertid have behov for at træne, teste og justere vores model løbende (hvad vi vender tilbage til i afsnittet om *Tuning*). Det kunne vi gøre ved fra begyndelsen at opdele datasættet yderligere, så vi havde både et trænings-, test- og holdout-sæt. Ofte anvendes imidlertid en mere avanceret fremgangsmåde kaldet *krydsvalidering* (James et al., 2013: 175-183). Ved *k-fold krydsvalidering* opdeles træningssættet i folder. Typisk udføres krydsvalidering med $k = 5$ eller $k = 10$ (James et al., 2013: 181). Når modellen trænes, udelader man én fold, og træner modellen på de resterende $k - 1$ folder af data. Herefter kan den udeladte k'te fold bruges til at teste modellens performance, fordi denne er holdt *out-of-sample* i forhold til træningen af modellen. Proceduren gentages k gange, én for hver fold, hvorefter den gennemsnitlige performance kan beregnes. Vi har nu trænet modellen på træningssættet og har samtidig et estimat af modellens forventede performance out-of-sample alene ved at bruge forskellige dele af træningssættet på skift. Dermed kan vi altså sammenligne forskellige modeller eller forskellige konfigurationer af samme model for at finde frem til den model, der performer bedst. Krydsvalidering har derudover en anden fordel i forhold til blot at træne modellen op imod et testsæt. Med krydsvaliderings-proceduren opnår vi nem-

lig mere robuste estimater for den performance, som modellen kan ventes at have *out-of-sample* i holdout-sættet, fordi vi her træner og tester modellen flere gange på forskellige dele af træningssættet og beregner den gennemsnitlige performance (James et al., 2013: 181-183). Logikken illustreres i nedenstående figur.

Figur 2: Tuning af modellen med k-fold krydsvalidering



Funktionelle former og feature engineering

Forskellen mellem empirisk målbar og ikke-målbar performance gør på sin vis prædiktion simplere end estimation. Prædiktion stiller i mange henseender lavere krav til modellens teoretiske fundament. Opstiller man fx en lineær OLS-model til kausalestimation, antager man i udgangspunktet, at den datagenererende proces er lineær i parametrene, og evt. afvigelser fra linearitetsantagelsen skal modelleres eksplicit. I prædiktion er vi interesserede i $\hat{y}$ og ikke i de enkelte parameter-estimater, $\hat{\beta}_k$. Når vi opstiller en given prædiktionsmodel $\hat{y} = f(\mathbf{X})$, kan vi derfor også tillade os at betragte selve funktionen f som en *black box* (James et al., 2013: 17-20). Det kan vi, fordi vi ikke er interesserede i funktionen eller parametrene i sig selv – vi er blot interesserede i at lave gode prædiktioner. Når det ikke er nødvendigt at kunne forstå og forklare prædiktionerne, er det heller ikke afgørende at tilstræbe teoretisk parsimoni. Som følge heraf kan vi arbejde med en bredere palet af mere fleksible funktionelle former i prædiktionsmodeller, såsom modeller med meget komplekse interaktioner mellem parametrene (James et al., 2013: 19-25).

På samme måde har man ved kausalestimation brug for teori til at guide valget af, hvilke variable som inkluderes i modellen. Udeladelse af relevante variable vil give *omitted variable bias*, og medtagelse af for mange kan give problemer med multikollinearitet og høje standardfejl til følge (Wooldridge,

2009: 89-99). Når formålet er prædiktion, bekymrer vi os derimod ikke om kausalitet og derfor heller ikke om spuriøsitet, men blot om at finde variable med prædiktiv værdi.

Det betyder, at datasæt til maskinlæring kan indeholde et væld af forskelligartede data, som vi ikke kan give nogen kausal fortolkning, og som vi ikke på forhånd har teoretiske forventninger til. I datasæt til maskinlæring vil man typisk også have mange flere variable, end vi traditionelt arbejder med i samfundsvidenskaben. Der er således ikke noget til hinder for at inkludere mange beslægtede variable og forskellige specifikationer af den samme variabel. Typisk vil en stor del af arbejdet ved maskinlæring bestå i at indsamle alt tilgængeligt data og eksperimentere med at generere nye variable. Denne proces kaldes *feature engineering*, og her afprøves mange forskellige specifikationer og kombinationer af variable for at maksimere deres prædiktive værdi (Foster et al., 2016). 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, såkaldt "p-hacking" (Gelman og Loken, 2013). Dette fifleri er imidlertid ikke blot acceptabelt, men ofte fundamentalt når maskinlæring anvendes til prædiktion, da modellen først og fremmest bliver vurderet på sin prædiktive performance (James et al., 2013: 26).

Overfitting

For en politolog har den største metodiske forandring ved prædiktion at gøre med, hvordan man bedst fitter en model til data. Både i estimations- og prædiktionssammenhæng fitter man rent teknisk en model til data ved at minimere en tabsfunktion (James et al., 2013: 17-26). En tabsfunktion er et udtryk for, hvor stor forskel der er mellem de fittede værdier $\hat{y}$ og de sande værdier y. Det koncept kender vi fx fra OLS, hvor vi finder det bedste fit til data ved at minimere tabsfunktionen i form af de kvadrerede residualer (James et al., 2013: 59-68).

Ved klassisk kausalestimation minimerer vi tabet *in-sample*. Det vil sige, at vi fitter modellen til vores eksisterende data, vores træningssæt. Det skyldes vores interesse i at forstå og forklare sammenhænge, fx hvilke personlige karakteristika som hænger sammen med frafald. Her vil et højt in-sample fit (høj R^2) være et udtryk for en god performance for en model. Ved prædiktion er vores interesse i stedet at minimere tabet *out-of-sample*, dvs. på hidtil usete data (James et al., 2013: 29-33; Kleinberg et al., 2015; Varian, 2014: 6-7). I vores

eksempel ønsker vi, at modellen leverer de bedst mulige forudsigelser, når vi prædikterer frafaldet blandt nye 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: 30). Årsagen er, at når vi maksimerer fittet in-sample, så risikerer vi at fitte til mønstre i data, som i virkeligheden er tilfældig variation. Det er variation eller "støj", som kan tilskrives uobserverede variable. Hvis vi fitter til disse mønstre i træningssættet, vil det øge vores tab i testsættet, hvor de samme mønstre ikke kan forventes at genfindes, fordi de varierer. Her *overfitter* vi vores model, hvilket mindsker modellens performance out-of-sample (James et al., 2013: 29-33).

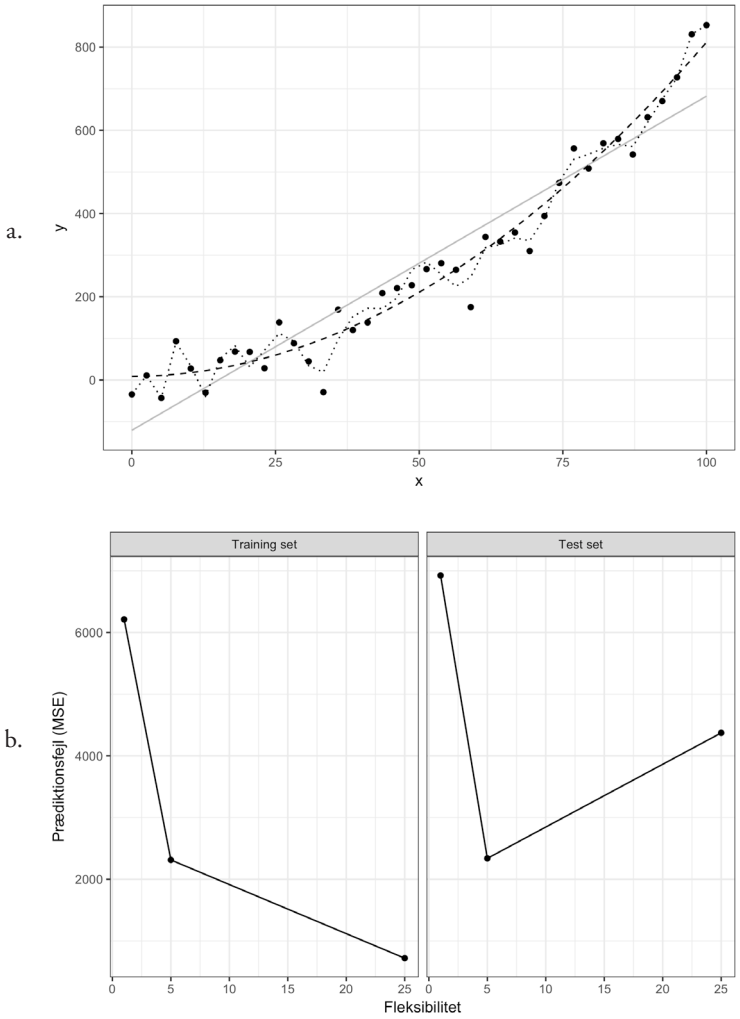
Sammenhængen mellem in-sample og out-of-sample performance er illustreret i figur 3. Her er MSE (mean squared error, dvs. den gennemsnitlige kvadrerede afvigelse mellem forudsagt og faktisk værdi) et ofte anvendt mål for en models tab eller fejlbarlighed (James et al., 2013: 17–19). En høj in-sample performance er ensbetydende med at modellen har lav MSE i træningssættet, mens en høj out-of-sample performance betyder lav MSE i testsættet.

I panel (a) øverst vises et træningssæt. De sorte prikker viser simulerede data generet fra en funktion $\hat{y} = f(\mathbf{X})$ plus tilfældig støj ϵ. Panelet viser ydermere tre forskellige modeller, som er fittet til datapunkterne. Den mest restriktive model er den lineære model, vist ved den grå linje, mens den prikkede linje viser den mest fleksible model (et 25.-gradspolynomium). Ind imellem de to ligger den stiplede linje, som viser y modelleret som et femtegradspolynomium af x.

I panel (b) nederst vises modellernes MSE'er. X-aksen angiver fleksibilitet i modellernes funktionelle form. Den grå linje i venstre side af panelet viser modellernes MSE'er i træningssættet. Det ses, at trænings-MSE'en kun bliver mindre, jo mere fleksible modellerne bliver. Den prikkede linje har således den laveste trænings-MSE, fordi den nærmest perfekt fitter alle punkterne i træningssættet. Linjen i højre side af panelet viser modellernes test-MSE. Her er det anderledes den stiplede linje, som har den laveste test-MSE. Den lineære grå linje har en relativt høj test-MSE, fordi den lineære model er for restriktiv og ikke er en specielt god tilnærmelse af formen på den datagenererende model. Vi siger, at modellen *underfitter*. Den meget fleksible prikkede linje har også en høj 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*.

Mens linjen til venstre blot er aftagende og viser dalende trænings-MSE for mere fleksible modeller, viser linjen til højre anderledes en U-form i modellernes test-MSE. Modellernes out-of-sample performance er dermed hverken optimal for en model, der er for restriktiv (den lineære model) eller for fleksibel

Figur 3. Illustration af overfitting



Figur 3a øverst viser et simuleret datasæt hvor y er modelleret som hhv. en lineær funktion af x (den grå linje), med en mere fleksibel form i form af et femtegradspolynomium (stiplet linje), og en endnu mere fleksibel form i form af et 25.-gradspolynomium (prikket linje). Figur 3b nederst viser linjernes prædiktionsfejl (målt ved mean squared error) i hhv. training set (dvs. in-sample) og test set (dvs. out-of-sample). Den mest fleksible form har mindst prædiktionsfejl in-sample, men som følge af af overfitting relativt større prædiktionsfejl out-of-sample. Inspireret af et lignende eksempel i James et al. (2013: 31). Linjerne i panel b er medtaget for at understøtte den visuelle sammenligning og repræsenterer ikke den sande sammenhæng mellem fleksibilitet og prædiktionsfejl.

(25.-gradspolynomiumet) i forhold til den datagenererende proces. Sammenhængen 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 modellen med den bedste out-of-sample performance, er det afgørende at ramme den rigtige balance mellem risikoen for underfitting og risikoen for overfitting. Særligt står begrebet overfitting centralt i maskinlæringslitteraturen. Det skyldes, at det med meget fleksible modeller – som fx klassifikationstræer – er meget nemt at fitte data meget præcist. Her kan en mindre fleksibel model potentielt give en lavere test-MSE og højere out-of-sample performance (Friedman, 2001).

Tradeoff'et mellem bias og varians

Den U-formede sammenhæng mellem en models fleksibilitet og dens performance out-of-sample skyldes to konkurrerende hensyn, når vi fitter en model til data. Der er tale om et tradeoff mellem bias og varians (James et al., 2013: 33). Varians kan vi forstå som et udtryk for, at der er tilfældig variation i data, og at de resulterende modeller derfor varierer, alt efter hvilket datasæt vi træner dem på. Generelt har mere fleksible modeller en højere varians (James et al., 2013: 33). Vi kan her kaste et blik tilbage på figur 2. Den prikkede kurve er meget fleksibel og har en høj varians. Det betyder, at hvis bare få datapunkter ændrer sig, så vil det betyde en stor ændring i funktionen for den prikkede kurve. Og det er meget sandsynligt, at datapunkterne vil ændre sig i et nyt datasæt, eksempelvis testsættet, fordi punkterne i figuren til dels afspejler tilfældig støj ϵ.

Bias kan forstås som en fejl, der følger af, at vi forsøger at modellere en kompleks sammenhæng fra virkeligheden med en meget enklere model (James et al., 2013: 35). Den grå linje i figuren viser fx en lineær model, som er meget lidt fleksibel, og det ses, at den funktionelle form er en dårlig tilnærmelse af den funktionelle form på den datagenererende proces. Det betyder, at uanset hvor mange flere observationer, vi tilføjer i træningssættet, så vil den lineære model 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).

Den bedste out-of-sample performance opnår vi ved samtidig at minimere både en models 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 (James et al., 2013: 33-36). I takt med at en model gøres mere fleksibel, vil det typisk være sådan, at bias indled-

ningsvist mindskes relativt hurtigere, end variansen øges. Derfor vil den forventede test-MSE samlet set blive mindre. Det vil forsætte indtil et vist punkt, hvor nytten af mere fleksibilitet ikke er specielt gavnlig, fordi formen på den estimerede funktion 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 i figur 2 giver U-formen på den forventede test-MSE som funktion af en models fleksibilitet (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 så lille bias som muligt. 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*.

Regularisering

Regularisering går ud på at begrænse en models fleksibilitet og kompleksitet (James et al., 2013: 203-204). En måde at gøre det på er ved at kontrollere fleksibiliteten direkte med parametre, der begrænser, hvor fleksibelt modellen kan fitte data. *Shrinkage* er en udbredt teknik til regularisering, der "straffer" en model for kompleksitet ved at introducere en såkaldt regulariseringsterm i modellens tabsfunktion, som bliver større, jo mere kompleks en model bliver (James et al., 2013: 203-204, 214-215). Der findes flere former for shrinkage. Én af dem fungerer rent matematisk ved, at der i modellens tabsfunktion minimeres kvadrerede residualer plus en regulariseringsterm. Regulariseringstermen består af en parameterværdi ganget med summen af modellens kvadrerede koefficienter. Den frie parameter kan antage værdier fra 0 til uendelig. Når den sættes til 0, svarer regressionen til almindelig OLS, fordi hele regulariseringstermen bliver 0. Når parameterværdien sættes større end 0 bliver modellen "straffet" for at have større koefficienter, fordi det giver et større tab. Regulariseringen tilskynder derfor til mindre koefficienter, og tilskyndelsen vokser i takt med parameterværdien. Med andre ord biaser regularisering modellen mod mindre koefficienter, men det giver samtidig en simplere model med mindre varians, da større koefficienter giver mere variable prædiktioner. Således får vi en regressionsmodel, hvor tradeoff'et mellem bias og varians modelleres via parameteren – jo større parameterværdi des mindre varians og mere bias, og omvendt (James et al., 2013: 214-224).

Tuning

Selve idéen om regularisering er ikke et nyt bidrag fra maskinlæringslitteraturen. Nogle politologer vil endog være bekendte med shrinkage-metoder som ridge- og lasso-regression. Det, som derimod er et nybrud i politologien, er den tilgang, hvormed maskinlæring kan anvendes til at fastsætte værdien for frie parametre som i eksemplet ovenfor. 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 enkelthed 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: 227-228; Varian, 2014: 6-7).

I tuning-processen ligger en del af læringselementet i maskinlæring. En udbredt metode er *k-fold krydsvalidering* som beskrevet tidligere. Her trænes modellen med en given kombination af parametre på $k - 1$ folder, og modellens performance testes på den k'te fold. Processen gentages k gange, én for hver fold, hvorefter den gennemsnitlige MSE beregnes som mål for performance. Herefter kan parametrene justeres, hvorefter krydsvalidering gentages, og modellernes out-of-sample MSE kan sammenlignes. Denne proces fortsætter, og slutteligt vælges den model med den kombination af parametre, som har den laveste MSE. Til at afsøge mulighedsrummet af parameterkombinationer anvendes ofte *grid search* i tuningen. Her angives en værdimængde for hver af de frie parametre, hvorefter alle kombinationer af parameterværdierne afprøves én efter én (James et al., 2013: 227-228; Varian, 2014: 6-7).

Med afsæt i disse kernebegreber vender vi nu vores opmærksomhed mod en særligt udbredt tilgang til prædiktion i maskinlæringslitteraturen, prædiktion ved hjælp af klassifikationstræer.

Prædiktion med klassifikationstræer

Der anvendes en lang række forskellige algoritmer til maskinlæring (se fx Hastie, Tibshirani og Friedman, 2009; Murphy, 2012), hvoraf ingen er universelt bedst på tværs af undersøgelser og datasæt (James et al., 2013: 29). En algoritme er et sæt af regler for, hvordan et specifikt problem løses – fx regler for, hvordan observationer skal håndteres for at finde mønstre og sammenhænge i et datasæt. Betegnelsen algoritme kan i vid udstrækning læses synonymt med *model*. Begrebet refererer mere præcist til en models indre regelsæt for, hvordan en tabsfunktion minimeres og modellen fittes til data.

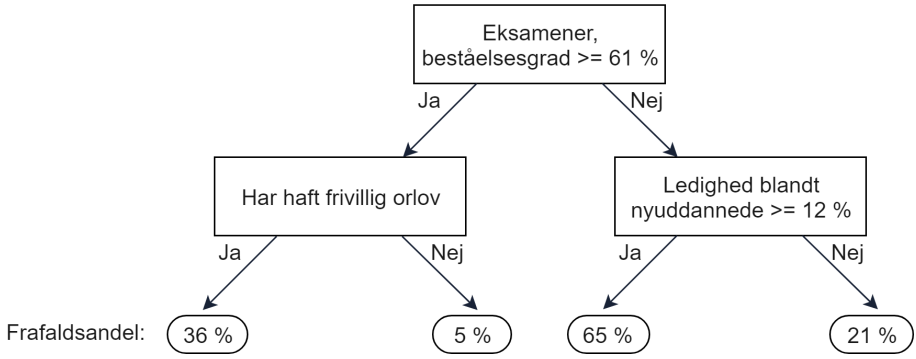
I vores eksempel, Københavns Professionshøjskole, anvendes en model baseret på klassifikationstræer. Klassifikationstræer har en særlig appel, idet de kan

modellere data meget komplekst og fleksibelt og alligevel bibeholde en intuitiv fortolkning. Klassifikationstræer fungerer ved at minimere en tabsfunktion igennem segmenteringen af data på baggrund af de variable, som algoritmen har til rådighed i datasættet. Tabsfunktionen kan tage forskellige former, der er mere eller mindre komplekse mål for, hvor "rene" outcome-grupper som kan dannes ud fra variablene i datasættet (Hastie, Tibshirani og Friedman, 2009: 356). I vores case om uddannelsesfrafald kan renhed forstås som et mål for, i hvor høj grad observationerne kan opdeles, sådan at de resulterende grupper kun indeholder hhv. frafaldne og gennemførte studerende.

Helt konkret segmenterer klassifikationstræer et datasæt ved at sortere observationerne i grupper på baggrund af en række binære spørgsmål om de uafhængige variable (James et al., 2013: 303-314). I vores case kunne det fx være et spørgsmål om, hvorvidt en given studerende i datasættet er en mand. Hvis ja, bliver den studerende placeret i den ene gruppe af observationer, og hvis nej i den anden gruppe. Efter algoritmen har foretaget et split, gennemløbes alle variable og alle deres mulige split igen for hver af de grupper, som er resultatet af tidligere split. Kontinuerte variable såsom alder kan splittes på alle de værdier, som variablen kan tage. Algoritmen leder som nævnt efter det split, der bedst segmenterer datasættet i grupper med et ensartet frafaldsmønster – dvs. at gruppen enten har en meget lav eller meget høj andel studerende, der frafalder. Algoritmen fortsætter med at segmentere datasættet indtil et givent stop-kriterium nås, fx at der kun er et præspecificeret antal studerende i hvert segment.

Figur 4 viser et eksempel på et klassifikationstræ til forudsigelse af frafald på Københavns Professionshøjskole.

Figur 4: Illustration af et klassifikationstræ fra Københavns Professionshøjskole



Træet viser frafaldsandelen for studerende i fire forskellige segmenter, der i overensstemmelse med træmetaforen kaldes *blade*. Risikoen for frafald er højest blandt den gruppe af studerende, som har en beståelsesgrad under 61 procent, og som går på en uddannelse med relativt høj dimittendledighed. Træet her er groet lavt for eksemplets skyld – i praksis opnås mere ensartede frafaldsmønstre i segmenterne ved at gro træet dybere og dermed splitte træet i flere segmenter (James et al., 2013: 303-306).

For politologer er en nyttig egenskab ved klassifikationstræer, at de som udgangspunkt kræver mindre præ-processering af data, dels fordi centrering og skalering af variable er unødvendigt, dels fordi outliers ikke har den ekstreme indflydelse, som vi kender det fra OLS (Hastie, Tibshirani og Friedman, 2009: 352). Det skyldes, at alle variable behandles som dikotome i træmodeller, hvor observationerne på en variabel splittes i to, over og under en given tærskel. Dermed er det underordnet, om observationerne ligger lige over eller meget over tærsklen – de har den samme indflydelse på modellen.

Et kendetegn ved klassifikationstræer er, at en segmentering betinger senere segmenteringer, hvilket fortsætter nedad i træet (James et al., 2013: 306-311). 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 (Montgomery og Olivella, 2018: 1-2). Det bidrager til gode prædiktioner, at vi ikke behøver at begrænse os til eksempelvis en lineær model.

Fordi klassifikationstræer kan modellere data meget fleksibelt, er de tilsvarende sensitive over for overfitting. Et træ kan uden videre gro, indtil det korrekt klassificerer samtlige observationer i et givent træningssæt. Men, som vi har set, vil det 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. Præcist, hvor meget træet skal regulariseres for at give den bedste performance out-of-sample, afgøres empirisk i tuningprocessen (James et al., 2013: 307-311).

I praksis giver et enkelt klassifikationstræ som regel ikke en særlig høj prædiktions-performance. Klassifikationtræer udgør i stedet byggestenen for en række mere komplekse algoritmer. Det gælder særligt såkaldte ensemble-modeller, som består af et stort antal klassifikationstræer. En populær ensemble-

model som allerede har vundet noget indpas i empirisk politologi er *Random Forest*, som i stedet for at fitte ét klassifikationstræ til data, fitter en hel skov af træer og tager gennemsnittet af deres prædiktioner. Til hvert træ anvender algoritmen kun en tilfældig subsample af træningssættets observationer og variable (James et al., 2013: 187-190). Det giver den fordel, at vi opnår en række forskelligartede og ukorrelerede træer, og når vi tager gennemsnittet af deres observationer bliver variansen af prædiktionerne mindre, ligesom når vi i andre sammenhænge tager gennemsnittet af en række uafhængige observationer (James et al., 2013: 316-323).

I et studie af statslige overgreb mod egne befolkninger bruger Hill og Jones (2014) fx Random Forest-modeller til at teste konkurrerende teoriers evne til at forudsige overgreb. Forfatterne finder blandt andet, at juridiske institutioner er stærkt prædiktive – en faktor som ikke figurerer prominent i gængse teorier på området.

En anden populær ensemble-model er *Gradient Boosted Trees*, som eksempelvis ligger til grund for frafaldsmodellen på Københavns Professionshøjskole. Gradient Boosted Trees fitter en række klassifikationstræer *sekventielt*, således at hvert nyt træ bygger videre på de forudgående ved at lægge størst vægt på de observationer, som blev klassificeret forkert ved de tidligere træer (James et al., 2013: 321-323).

Omkostningen ved ensemblemodellerne er, at de resulterende modeller er sværere at fortolke, men til gengæld giver de typisk en væsentligt forbedret prædiktions-performance (Athey og Imbens, 2016).

Metodiske udfordringer ved maskinlæring anvendt til målretning

Anvendelsen af maskinlæring har uden tvivl stort potentiale i den offentlige forvaltning, hvor prædiktion kan bruges til at målrette politiske tiltag. De nye træbaserede algoritmer har også et stort potentiale til at afdække komplekse interaktioner mellem variable.

Samtidig er det dog nødvendigt at huske, hvad maskinlæring ikke kan: afdække kausale sammenhænge. Maskinlæring er særlig relevant i tilfælde, hvor vi alene er interesserede i at prædiktere et outcome. I praksis er det dog de færreste politiske problemstillinger, som vil være rene prædiktions-problemer. I vores gennemgående eksempel ønsker Københavns Professionshøjskole fx at forudsige frafald blandt studerende for at kunne målrette en indsats mod de mest frafaldstruede. Det forudsætter, at vi er bekendte med indsatser, som uafhængigt af de bagvedliggende årsager kan benyttes til at afværge frafald. Forhindring af frafald er med andre ord ikke et rent prædiktions-problem. Vi

er ikke kun afhængige af præcise prædiktioner, men også af at kende effekten af de tiltag, som vi gerne vil målrette – og dermed er vi tilbage ved kausalestimation. Det er med andre ord svært at forestille sig, at prædiktion kan stå alene uden kausalestimation.

Anvendelsen af maskinlæring giver også anledning til en mere ejendommelig metodisk udfordring. Den følger af, at formålet ikke er at estimere en strukturel datagenererende proces, som vi er vant til som politologer. I stedet er formålet at finde sammenhænge i data og målrette tiltag for netop at *ændre* disse sammenhænge.

Lad os forestille os en situation, hvor Københavns Professionshøjskole målretter et succesfuldt tiltag mod de 10 pct. mest frafaldstruede studerende. Lad os sige, at tiltaget i stort omfang reducerer frafaldet blandt denne gruppe af studerende. Det vil betyde, at de særlige karakteristika ved denne gruppe, som gjorde, at den blev identificeret som meget frafaldstruet, 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. 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.

Konklusion og perspektiver

Inden for mange områder af den offentlige forvaltning er der en stigende interesse for at anvende prædiktion til målretning af policy-tiltag. Dermed følger også en stigende politologisk interesse for superviseret maskinlæring, som er særligt velegnet, når vi interesserer os for præcise prædiktioner frem for præcise parameterestimater. I artiklen har vi introduceret en række kernebegreber omkring anvendelsen af maskinlæring til prædiktion: out-of-sample performance, test- og træningssæt, under- og overfitting, feature engineering, regularisering, tuning og krydsvalidering. Maskinlæring muliggør endvidere mere fleksibel modellering af data, og i artiklen har vi introduceret klassifikationstræer og deres afledte ensemble-modeller, som er begyndt at finde anvendelse i samfundsvidenskaben.

Vi skitserede også to metodiske udfordringer ved at anvende prædiktion til målretning. De pegede begge i retning af, at der nok vil være få cases i den offentlige forvaltning, hvor vi vil have glæde af prædiktion uden også at beskæftige os med kausalestimation.

Dertil kommer, at det også kan rejse etiske udfordringer at anvende maskinlæringsmodeller til prædiktion og målretning af tiltag. Algoritmer kan måske nok lave mere præcise forudsigelser end mennesker, men der kan være andre hensyn at tage end præcision alene. I offentlig sagsbehandling er der fx en forventning om, at beslutninger kan udfordres og ankes. Det harmonerer dårligt med beslutningstagning baseret på maskinlæringsmodeller, der ofte bliver kritiseret for at være uigennemskuelige *black boxes*, som er svære at bestride (Kitchin, 2014).

Det er også langt fra trivielt, hvordan man balancerer ønsket om præcise prædiktioner over for beskyttelse af den enkeltes privatliv. Vi kan fx forestille os en situation, hvor registerdata om alt fra børns tandlægebesøg til forældrenes indkomstniveau sammenkøres for at lave præcise forudsigelser af social udsathed for at kunne målrette en tidlig, præventiv indsats. Det er endvidere nødvendigt at forholde sig til, at forudsigelser nogle gange vil være forkerte. Jo mere indgribende et tiltag er over for den enkelte, jo mere nøje må man have forholdt sig til de etiske konsekvenser af fejlagtige prædiktioner. I tilfældet med socialt udsatte børn er det fx ikke blot det målrettede tiltag, som kan påvirke det enkelte barn. Også selve klassifikationen af et barn som socialt udsat kan påvirke omgivelsernes syn på barnet og barnets syn på sig selv. Samme typer overvejelser er relevante i vores gennemgående eksempel – hvordan påvirker det egentlig en studerende at blive prædikteret til at være i risiko for at frafalde sin uddannelse?

Etiske og metodiske udfordringer til trods er maskinlæring efter alt at dømme kommet for at blive som redskab i den politologiske værktøjskasse. Anvendelsesmulighederne er mange, og voksende udbredelse i den danske såvel som i udenlandske offentlige sektorer indikerer både udbud og efterspørgsel. Kunsten bliver at balancere mellem det teknisk mulige og det samfundsmæssigt ønskværdige.

Referencer

Angrist, Joshua D. og Jörn-Steffen Pischke (2015). *Mastering 'Metrics: The Path from Cause to Effect*. Princeton University Press.

Athey, Susan og Guido Imbens (2016). The state of applied econometrics-causality and policy evaluation, *ArXiv: 1607.00699*.

Bach, Alexander og Jesper Svejgaard (2017). Maskinlæring på skolebænken. Speciale, Institut for Statskundskab, Københavns Universitet.
Berk, Richard (2012). *Criminal Justice Forecasts of Risk: A Machine Learning Approach*. Springer Science & Business Media.
Breiman, Leo (2001). Statistical modelling: The two cultures. *Statistical science* 16 (3): 199-231.
Chandler, Dana, Steven D. Levitt og John A. List (2011). Predicting and preventing shootings among at-risk youth. *The American Economic Review* 101 (3): 288–292.
Conway, Drew og John Myles White (2012). *Machine Learning for Hackers*. Sebastopol: O'Reilly.
Foster, Ian, Rayid Ghani, Ron S. Jarmin, Frauke Kreuter og Julia Lane (2016). *Big Data and Social Science*. Chapman: Hall/CRC.
Friedman, Jerome H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics* 29 (5): 1189-1232.
Gelman, Andrew og Eric Loken (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.
Hariri, Jacob Gerner (2012). Kausal inferens i statskundskaben. *Politica* 44 (2): 184-201.
Hastie, Trevor, Robert Tibshirani og Jerome Friedman (2009). *The Elements of Statistical Learning*. Springer.
Hill, Daniel W. og Zachary M. Jones (2014). An empirical evaluation of explanations for state repression. *American Political Science Review* 108 (3): 661-687.
Hofman, Jake M., Amit Sharma og Duncan J. Watts (2017). Prediction and explanation in social systems. *Science* 355 (6324): 486-488.
James, Gareth, Daniela Witten, Trevor Hastie og Robert Tibshirani (2013). *An Introduction to Statistical Learning*. Springer.
Kitchin, Rob (2014). Big data, new epistemologies and paradigm shifts. *Big Data & Society* 1 (1). https://doi.org/10.1177/2053951714528481.
Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig og Sendhil Mullainathan (2017). *Human Decisions and Machine Predictions*.
Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan og Ziad Obermeyer (2015). Prediction policy problems. *American Economic Review* 105 (5): 491-495.
Lantz, Brett (2015). *Machine Learning with R.* Birmingham: Packt Publishing
Montgomery, Jacob M. og Santiago Olivella (2018). Tree-based models for political science data. *American Journal of Political Science* 62 (3): 729-744.
Mullainathan, Sendhil og Jann Spiess (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives* 31 (2): 87-106.

Murphy, Kevin P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT press.
Obermeyer, Ziad og Ezekiel J. Emanuel (2016). Predicting the future: Big data, machine learning, and clinical medicine. *The New England Journal of Medicine* 375 (13): 1216-1219.
Perry, Walt L., Brian McInnis, Carter C. Price, Susan C. Smith og John S. Hollywood (2013). *Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations*. Rand Corporation.
Rosenbaum, Paul R. og Donald B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* 70 (1): 41-55.
Samii, Cyrus (2016). Causal empiricism in quantitative research. *The Journal of Politics* 78 (3): 941-955.
Samuel, Arthur L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development* 3 (3): 210-229.
Shalev-Shwartz, Shai og Shai Ben-David (2014). *Understanding Machine Learning: From Theory to Algorithms*. New York: Cambridge University Press.
Varian, Hal R. (2014). Big data: New tricks for econometrics. *The Journal of Economic Perspectives* 28 (2): 3-27.
Wooldridge, Jeffrey M. (2009). *Introductory Econometrics: A Modern Approach*. South-Western Cengage Learning.