Eksamens projekt

Indholdsfortegnelse

Indledning

Investeringer i human kapital sker med forventninger til afkast i arbejdsmarkedet (Becker 1964). Grundet direkte samt indirekte diskrimination, i.e. lønforskelle, kønnene imellem, er kønnenes traditionelle specialisering, mænd i lønnet arbejde og kvinder i hjemmet, fremhævet som økonomisk optimal (Becker 1991; Becker 1985). Med forventning om denne kønsspecialisering forventes det yderligere, at kvinder investerer mindre i human kapital, i.e. løndiskrimination er cirkulær og selvforstærkende fra et human kapitalsynspunkt (Blackburn et al. 2002). På trods af den forudsete cirkularitet, er der gennem de seneste årtier sket dramatiske kønsforandringer i human kapital investeringer: kvinder udgør nu majoriteten af studerende på længere videregående uddannelser (se nedenfor). Igen, fra et human kapital perspektiv burde kvinders øgede investeringer i human kapital betyde mindre lønforskelle mellem kvinder og mænd. Markante lønforskelle kønnene imellem er dog fortsat observeret, endda også imellem kvinder og mænd med længerevarende videregående uddannelser, hvor mænd i 2014 havde en bruttoindkomst 36,54 % højere end kvinder , jf. Tabel 1.

Table 1: Bruttoløn fordelt på køn, 2014

	Mænd	Kvinder
Lange videregående uddannelser	670.133	490.782

Kilde: Danmarks Statistik

En mulig forklaring for disse lønforskelle er divergerende afkast på investeringer i forskellige længerevarende uddannelser. Uddannelsesvalg er præget af segregering: kun få videregående uddannelser har et optag af studerende med lige andele af mænd af kvinder. Kønsforskelle i uddannelsesvalg kunne skyldes divergerende præferencer kønnene imellem (Hakim 2000), men *'. . . individual preferences (and thus choices) are always socially embedded and constrained, and may be shaped by unjust background conditions, as well as by habit and engrained normative assumptions.'* (Crompton 2007: 234). Socialt determinerede kønspræferencer i forhold til uddannelse har potentiale til at cementere lønforskelle kønnene imellem på trods af samme resulterende niveau af human kapital. Derfor bør det undersøges, hvorvidt der foreligger en tendens blandt kvinder til at søge mod længerevarende videregående uddannelser, der leder til relativt lavere lønninger, og en tendens blandt mænd til at søge mod længerevarende videregående uddannelser, der resulterer i relativt højere lønninger. Hvis

uddannelser med overrepræsentation af kvinder generelt leder til lavere lønninger, kan det både skyldes kønsforskelle i socialt afgrænsede præferencer, i.e. kvinder søger aktivt mod disse lavere lønnet uddannelser, eller lønforskelle opstået grundet feminisering af bestemte faggrupper. Øget feminisering af faggrupper resulterer ofte i lavere lønninger for disse grupper, grundet for eksempel diskrimination eller over-crowding (Rubery 2015; Bergmann 1974). Sagt på en anden måde, kan lavere lønninger for kandidater fra uddannelser med en overrepræsentation af kvinder skyldes, at kvinder søger mod uddannelser med lavere lønninger, samt at kvinder generelt modtager lavere løn, og lønniveauet for uddannelser med overrepræsentation af kvinder derfor er faldet. Uanset forklaringen, er det klart at lønulighed kønnene imellem allerede initieres ved kvinder og mænds valg af uddannelse.

Som det bevises nedenfor, kan uddannelsesvalg alene forklare omkring en fjerdedel af lønforskellen blandt kvinder og mænd med længerevarende videregående uddannelser. Dette kan skyldes en række faktorer, blandt andet kønspræferencer for uddannelsesretninger. Specifikke uddannelsesretninger leder til arbejde i højtlønnede faggrupper, såsom mandsdominerede business og tekniske uddannelser. Uddannelsesretninger kan forklare en del af lønforskellen kvinder og mænd imellem, men selv inden for disse specifikke uddannelsesretninger segregeres kønnene – mænd søger også her i højere grad mod uddannelser med højere lønafkast.

Med kønsopdelt data for ansøgerantal samt optagne studerende på alle længerevarende videregående uddannelser i Danmark er det muligt at bestemme omfanget af kønssegregering i uddannelsesvalg. Ydermere, gør løndata for færdige kandidater det muligt at bestemme, hvorvidt der forefindes en sammenhæng mellem kønssegregering blandt optagne studerende og forventet løn efter afsluttet kandidatuddannelse. Sluttelig analyseres det, hvorvidt forventet lønniveau kan forudsige kønsfordelingen blandt ansøgere på længere videregående uddannelser.

Data, metode og etik

Danske universiteters data for antal ansøgere samt optagne på bacheloruddannelser og er offentligt tilgængeligt gennem Uddannelses- og Forskningsministeriet. Data for årene 2013-2016 benyttes for at sikre et stort, repræsentativt sample, samt for at påvise eventuelle nylige forandringer i uddannelsesvalg. Antal af ansøgere og optagne er opdelt efter køn, og desuden indeholder datasættene bachelorstudiernes adgangskvotienter.

Løndata opgjort per uddannelse for alle færdige kandidater er produceret af CEPOS og ligger ligeledes offentligt tilgængeligt. CEPOS har udarbejdet et datasæt for gennemsnitlige bruttolønninger opgjort efter færdiggjort længerevarende videregående uddannelse for alle danskere mellem 25-59 år, men disse data er ikke opdelt efter køn. Da disse løndata er gennemsnitlige og inkluderer både mandlige og kvindelige kandidater fra de respektive uddannelser, mistes dimensionen af kønsforskelle i løn inden for de individuelle uddannelser. På den anden side gør dette løngennemsnit det muligt at isolere den del af lønforskelle, der opstår alene som resultat af kønsforskelle i uddannelsesvalg, hvilket netop er formålet med dette studie.

De to datakilder er forskellige i den forstand, at universiteternes ansøgninger og optag er opgjort for bacheloruddannelser, men løndata for færdige kandidater, i.e. for færdige længerevarende videregående uddannelser. I denne analyse indgår derfor kun bacheloruddannelser, der direkte leder til en kandidatuddannelse, e.g. gennem retskrav på optagelse på en specifik kandidatuddannelse. De to datasæt er derfor forenet ved at identificere par af bachelor- og kandidatuddannelser, e.g. bachelor- og kandidatstudierne i medicin. Dette udelukker for eksempel professionsbacheloruddannelser, der kunne lede til en kandidatuddannelse, og introducerer en bias, da sammenhængen mellem disse bachelor- og kandidatuddannelser ikke er perfekt, e.g. grundet frafald og uddannelsesskift. Ydermere identificeredes 10 uddannelsesretninger, såsom sundhedsvidenskab og samfundsvidenskab. Denne kategorivariabel er essentiel for en vurdering af uddannelsesretningers indflydelse på lønforskelle.

De anvendte data er offentligt tilgængelige, og derfor forefindes ikke rettighedsproblemer. Etisk kan det problematiseres, at nogle uddannelser har meget få optagne, især efter opdeling af køn, således at det er muligt identificere enkelte studerende i datasæt. Dog indgår ingen data på individniveau, hverken i datasættene eller i analysen, og etiske problemer er derfor ikke eksisterende.

Analyse

For perioden 2013-2016 viser data fra de danske universitet klart, at kvinder nu udgør majoriteten af både ansøgere (54,22 %) og optagne (53,10 %) på bacheloruddannelser, der giver direkte adgang til en kandidatuddannelse. På trods af dette, findes der i arbejdsmarkedet fortsat store lønforskelle kønnene imellem, og uddannelsesvalg alene kan forklare en stor del af disse lønforskelle. Det er muligt at forudsige gennemsnitlige lønninger for optagne kvinder og mænd i perioden 2013-2016, ved hjælp af antallet af optagne kvinder og mænd på de individuelle uddannelser samt forventede

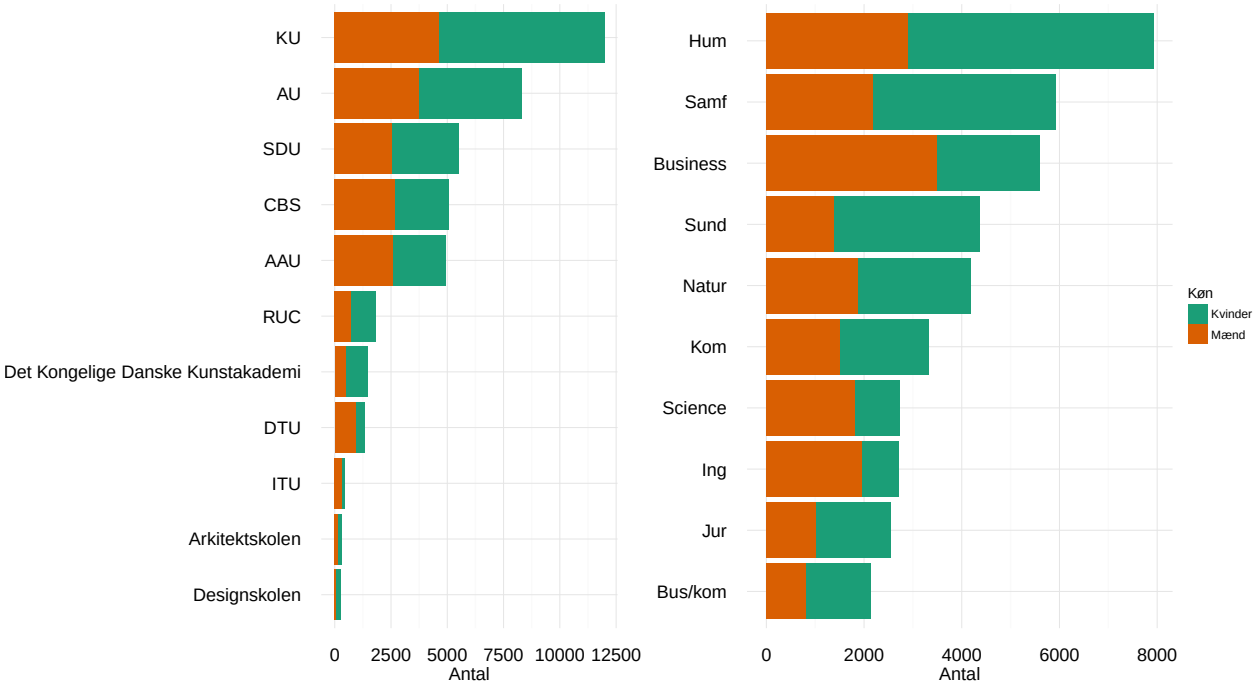
lønninger efter endt uddannelse.

Table 2: Vægtede, forventede lønninger pr. køn og år, 2013-2016

	2013	2014	2015	2016
Mænd	612.272	614.439	617.768	619.243
Kvinder	567.313	568.562	572.960	572.933
Forskel	44.959	45.877	44.809	46.310

De vægtede gennemsnit er vist i Tabel 2. Der er stor forskel i de forudsagte gennemsnitslønninger for kvinder og mænd, og uddannelsesvalg alene har derfor stor betydning for senere lønforskelle i arbejdsmarkedet. Ydermere viser Tabel 2, at de forudsagte lønforskelle har været meget stabile gennem de seneste fire år.

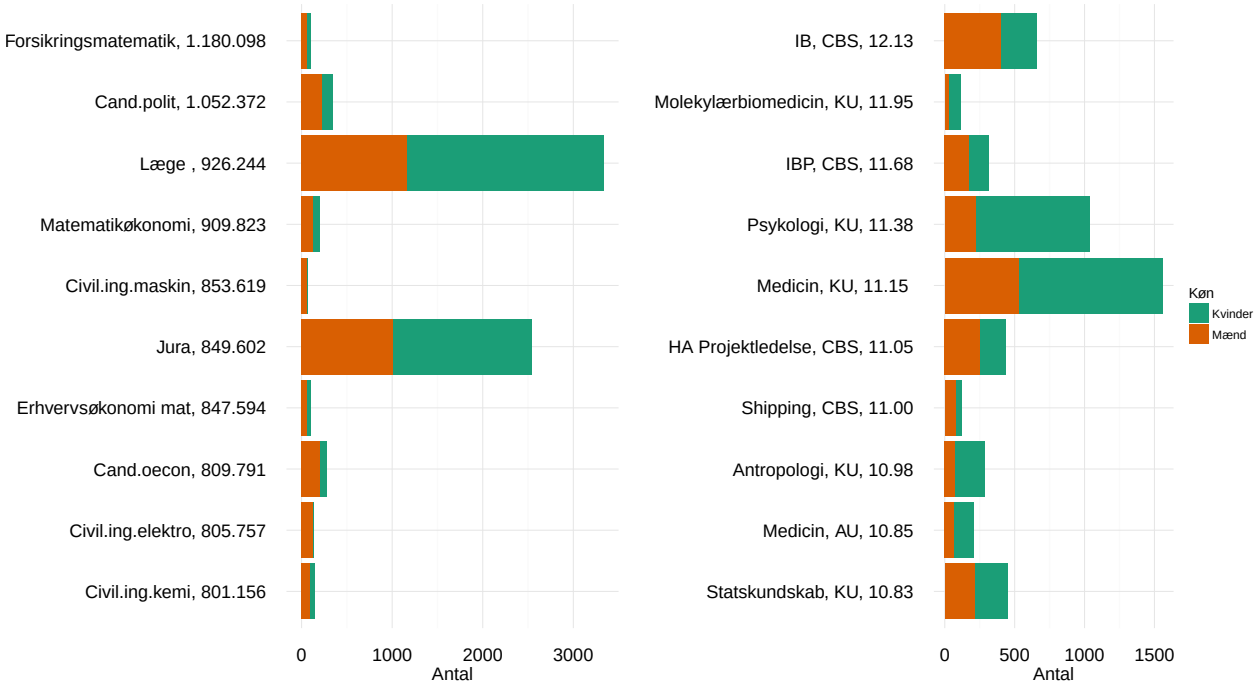
Figur 1: Gennemsnitlig antal ansøgere fordelt på køn ved danske uddannelse institutioner og retninger, 2013-16



Figur 1 viser de absolutte antal af kvindelige og mandlige ansøgere fordelt på de relevante uddan-

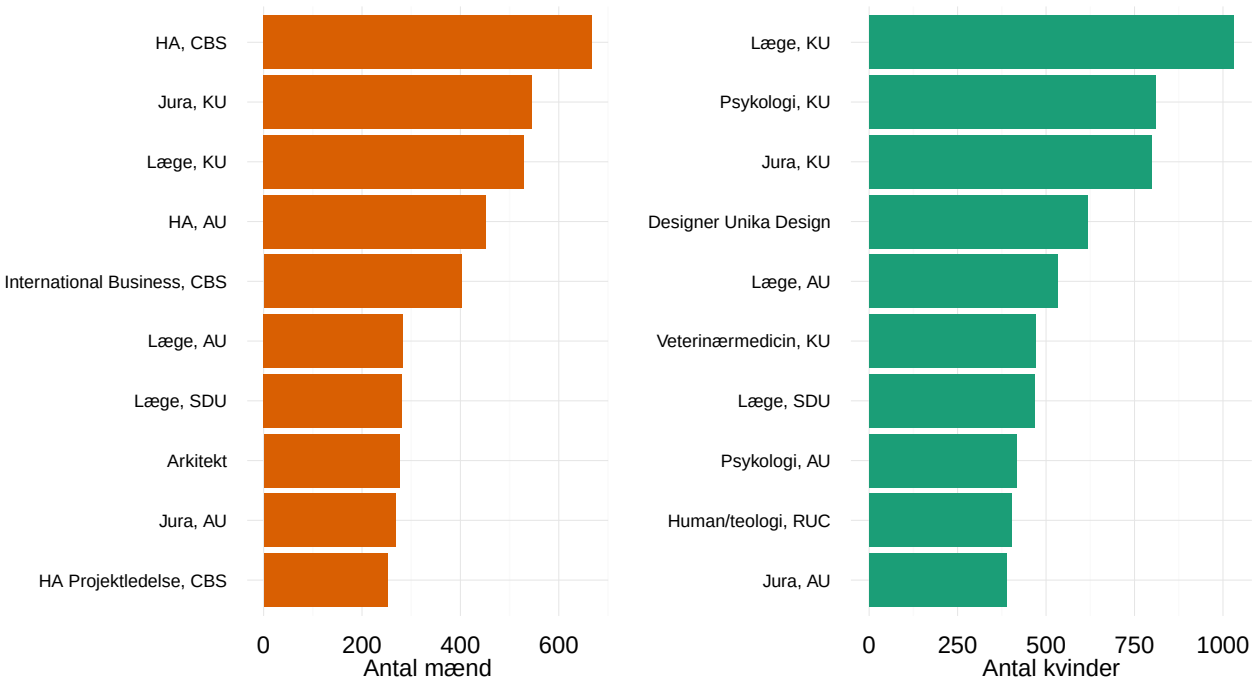
nelsesinstitutioner. Kvinder udgør majoriteten på de fleste traditionelle universiteter, men billedet er anderledes på CBS, DTU og ITU. Desuden kan det også ses, at studieretninger som ingeniør, science og business er domineret af mandlige ansøgere. Dette giver sig også til udtryk i de gennemsnitlige antal af kvindelige og mandlige ansøger på det højest lønnede uddannelser, Figur 2.

Figur 2: Gennemsnitlige antal ansøgere fordelt på køn for top 10 uddannelser målt på bruttoløn, kr. og top 10 gennemsntlig højst kvotient, 2013-16



De højest lønnede uddannelser er alle dominerede af mandlige ansøgere, med undtagelse af to: professionerne lægevidenskab og jura. Dette replicerer det velkendte – og internationale – billede af kvinders koncentration på længerevarende videregående uddannelser, der giver adgang til professioner (Connolly and Gregory 2007). Inden for professionerne, såsom lægevidenskab, jura og psykologi, er det klart defineret, hvilke niveauer af human kapital, der er nødvendige for at praktisere. I professionerne er kvinder og mænd med de samme kvalifikationer derfor i høj grad ansat i lignende stillinger – der er mindre rum for diskrimination (Connolly and Gregory 2007).

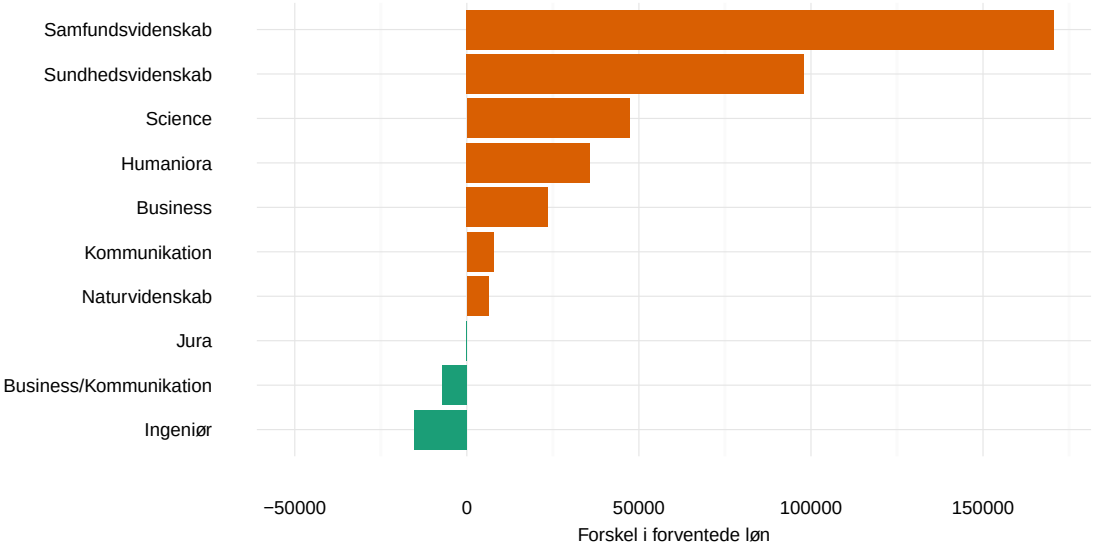
Figur 3: Top 10 gennemsnitlig mest søgte uddannelse for kønnene, 2013-2016



Denne søgen mod professionerne afspejler sig tydeligt i kvinders uddannelsesvalg, både i forhold til de top 10 mest søgte uddannelser og i forhold til de 10 uddannelser med de højeste adgangskvotienter, Figur 3. Her ses en tydelig tendens: mænd søger mod højtlønnede businessuddannelser, hvor kvinder i langt højere grad søger mod professionerne lægevidenskab, psykologi, jura samt veterinærvidenskab.

Figur 2 viser ligeledes, at de 10 uddannelser med højest lønnede kandidater alle er tekniske og matematiske, med undtagelse af de to professioner, lægevidenskab og jura, hvor kvinder udgør majoriteten af ansøgerne. Mænds dominans i de højest lønnede tekniske og matematiske fag, beviser at kønsstereotype antagelser bliver cementeret i både i uddannelsesvalg og i arbejdsmarkedet: traditionelle, mandsdominerede faggrupper oplever højere lønninger, og disse faggrupper tiltrækker en stor overvægt af mandelige ansøgere – lønulighed er dermed med en cirkulær proces. Grundet store lønforskelle på tværs uddannelsesretninger, e.g. ingeniørvidenskab og jura, kan lønforskelle kønnene imellem derfor potentielt forklares gennem divergerende kønspræferencer for uddannelsesretninger, e.g. flere mænd søger mod de højtlønnede uddannelsesretninger business og ingeniørvidenskab. På den anden side er det muligt, at mænd også koncentrerer sig i de højest lønnede uddannelser inden for de forskellige uddannelsesretninger, og uddannelsesretning dermed kun er en lille del af forklaringen.

Figur 4: Forskel i forventede, vægtede lønninger pr. studieretning



Figur 4 viser, at dette faktisk er tilfældet: med undtagelse af business/kommunikation samt ingeniørvidenskab koncentrerer mænd sig i de bedst lønnede uddannelser, også inden for specifikke uddannelsesretninger. Dermed er kønsspecialisering og -koncentration i specifikke uddannelsesretninger kun en delvis forklaring af lønforskelle mellem kvinder og mænd med videregående uddannelser. For at finde et mere robust billede af kønsforskelle i uddannelsesvalg, introduceres udvidede modeller i det følgende afsnit.

Modeller

I et forsøg på at forklare kønsfordelingen på længere videregående uddannelser udregnes andelen af kvinder optaget på alle de forskellige længere videregående uddannelser i Danmark. Disse kvindeandele benyttes herefter som afhængig variabel i en række økonometriske modeller. Først benyttes en logit-regressionsmodel, da den afhængige variabel, kvindeandele er proportionel. Dernæst anvendes en statistisk læringsmodel til at opstille et decision tree for processen bag de kønsspecifikke uddannelsesvalg.

Tabel 3: Logit-modeller,
marginal effekter

Model	(1)	(2)	(3)	(4)
Konstant	0.238***	0.350***	0.271***	0.274***
	(0.005)	(0.010)	(0.010)	(0.010)
Indkomst i 100.000 kr.	-0.035***	-0.042***	-0.040***	-0.039***
	(0.001)	(0.001)	(0.001)	(0.001)
Business		-0.174***	-0.172***	-0.157***
		(0.008)	(0.008)	(0.008)
Humaniora		-0.042***	-0.018*	-0.016*
		(0.008)	(0.008)	(0.008)
Ingeniør		-0.291***	-0.264***	-0.270***
		(0.009)	(0.009)	(0.009)
Jura		0.109***	0.077***	0.111***
		(0.010)	(0.010)	(0.011)
Kommunikation		-0.085***	-0.065***	-0.067***
		(0.008)	(0.009)	(0.009)
Naturvidenskab		-0.057***	-0.034***	-0.036***
		(0.008)	(0.008)	(0.008)
Samfundsvidenskab		-0.016*	-0.019*	-0.011
		(0.008)	(0.008)	(0.008)
Science		-0.295***	-0.264***	-0.264***
		(0.009)	(0.009)	(0.009)
Sundhedsvidenskab		0.184***	0.142***	0.158***
		(0.010)	(0.010)	(0.010)
Adgangskvotient			0.011***	0.011***
			(0.000)	(0.000)
Totalt optag i hundrede				-0.009***
				(0.001)
McFadden R-sq.	0.1	0.4	0.4	0.4
Deviance	19854.7	12942.2	12271.8	12206.5

Machine learning

I forrige afsnit var formålet at opstille en model, som kan give indsigt i kausale sammenhænge mellem kvindeandelen af de optagne studerende og de valgte forklarende variable. I kontrast til det, er formålet med dette afsnit at opstille en model, som kan forudsige kvindeandelen af de optagne studerende på baggrund af udvalgte inputvariable.

Ved brug af en *machine learning* algoritme udarbejdes et regressionstræ. Denne type regressionstræ benytter en delmængde af de tilgængelige data, kaldet træningssættet, til at klassificere studierne ved at "lære" forskelle i strukturer mellem uddannelser, og dermed forudsige udfaldet af kvindeandel. Metoden kaldes *supervised learning*, da algoritmen udregner og dernæst viser, hvilke karakteristika, der bestemmer divergerende niveauer af kvindeandele på uddannelser i træningssættet. I dette tilfælde benyttes inputvariablene forventet gennemsnitsindkomst, studieretning, adgangskvotient samt uddannelsesstørrelse i et forsøg på at forudsige kvindeandelen af de optagene studerende på en given uddannelse.

Intuitionen bag regressionstræet er, at træningsdata gentagne gange opdeles i mindre dele. De enkelte knuder i træet illustrerer denne opdelingsproces. Til de enkelte knuder hører en mindre model, som igen splitter data op i mindre dele ved at fremsætte nye "spørgsmål", der besvares binært. Processen kaldes segmentering og er relativ simpel ift. andre algoritmer, dermed kan segmenteringsmodellen fortolkes intuitivt og nemt illustreres grafisk. Samtidig er det muligt at udarbejde et beslutningstræ med lige så mange knuder, som der findes observationer i træningsdatasættet og derved lade RMSE konvergere mod 0. Mange knuder gør modellens beskrivelse af træningsdata meget nøjagtig, men de gør samtidig modellen ude af stand til at give præcise forudsigelser på nye data på grund af overfit. Derfor er valget af et optimalt antal knuder centralt.

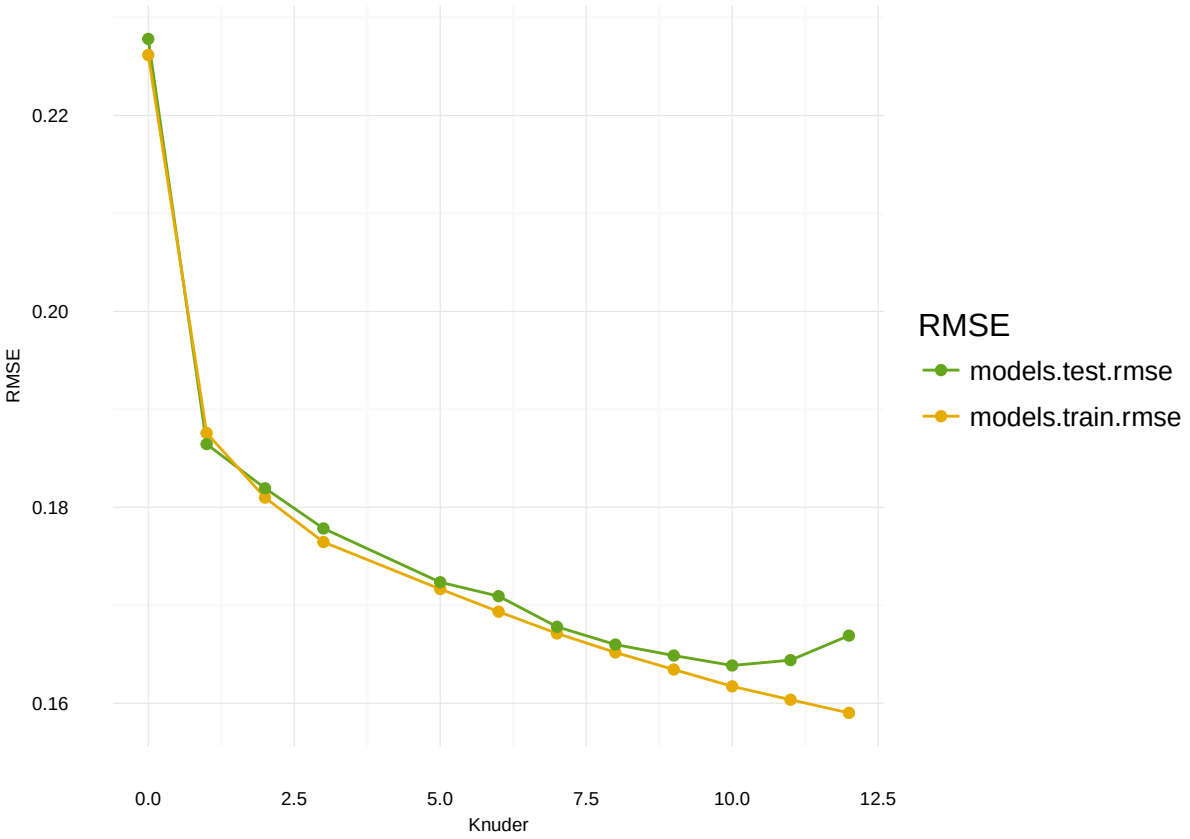
Validering

Der findes forskellige metoder til udvælgelsen af antallet knuder, heriblandt kryds-valideringsmetoden, som ofte giver mere robuste resultater end simpel validering. På trods af dette, og at kryds-validering er standardmetoden i algoritmepakken, benyttes her simpel validering. Årsagen til dette er, at simpel validering giver mulighed for grafisk, at forstå udvælgelsen af knuder.

Indledningsvist udtrækkes halvdelen af observationerne, på tilfældigvis, i det oprindelige datasæt. Udtrækket bliver brugt som træningsdata, mens den resterende del af data benyttes til validitets, også

kaldes testdata. Dernæst udarbejdes et stort regressionstræ, som beskæres indtil det optimale under-træ opnås. Det optimale under-træ er det under-træ, som mindsker RMSE mellem forudsigelserne og observationerne i testdata. I dette tilfæde, frembringes, ved denne process, et regressionstræ som består af 8 knuder. Dette fremgår af grafen nedenfor, hvor RMSE er en funktion af knuder. For testdata har RMSE har globalt minimum i 8, mens RMSE for træningsdata konvergere mod 0, præcis som ventet. Det udtrykker altså en balancen mellem bias ved lille et træ og overfitting ved et stort træ.

Figur 5:



Den alternative valideringsmetode, krydsvalidering, minder meget om ovenstående simple valideringsmetode - dog betyder krydsvalidering at træningsdata opdeles i flere undersæt, og derved skal flere modeller estimeres og beskæres. Denne process er beregningstung, og derfor opstilles en cost-funktion for RMSE, hvori et kompleksitetsled formuleres. Kompleksitetsleddet er produktet af en positiv strafkoefficient og antal af knuder. Givet en række udvalgte strafkoefficienter estimeres blot en sekvens af de ellers mange kombinationer af modeller. Gennem en rekursiv krydsningsprocess vælges efterfølgende den strafkoefficient, der minimerer cost-funktionen for modellen i træningssættet.

En fordel ved krydsvalidering er en lavere risiko for stikprøve-bias, da træningsdata opsplittes i flere tilfældige undersæt. Udover krydsvalidering kunne bootstraaping også benyttes til at lave syntetetiske stikprøver. Den grafiske fortolkning bag model- eller knudevalg - som vist i Figur X7 - er dog kun mulig ved simpel validering.

Regressionstræet

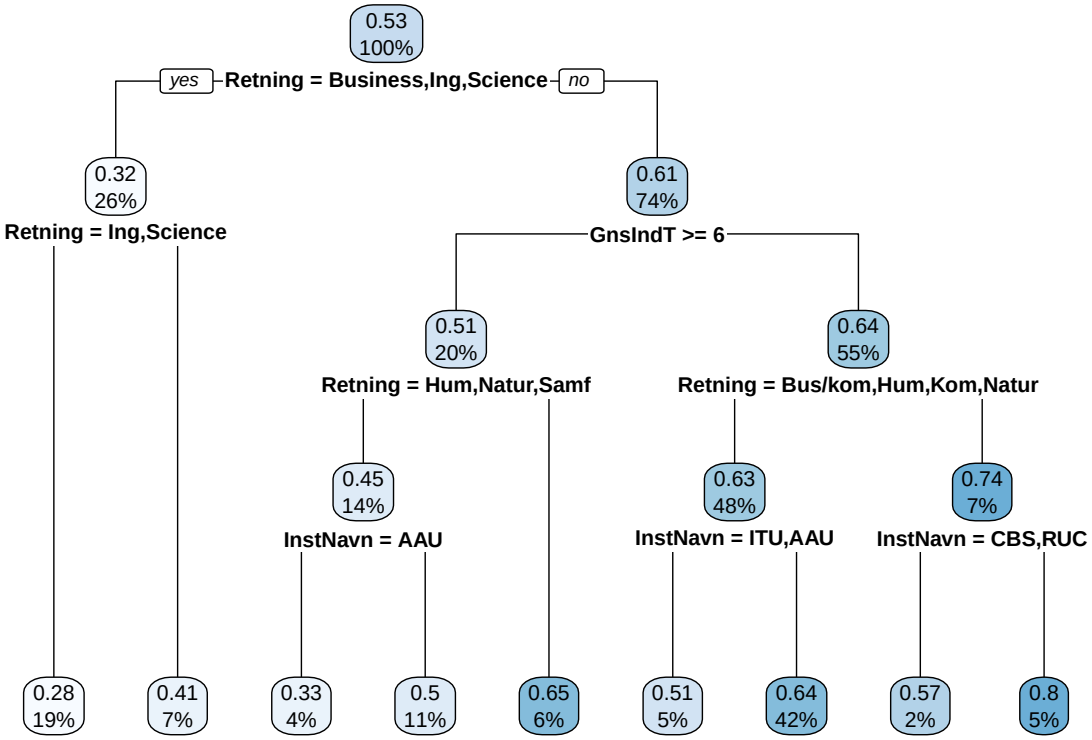
Ved den første knude i træet tages udgangspunkt i hele datasættet. Det ses her, at modellen på baggrund af alle data for de fire inputvariable beregner et gennemsnit for kvindeandelen af optagede til 0,52. Modellen fremsætter i første knude spørgsmålet "Retning = business, ingeniør eller science", og data opdeles dermed efter svaret på dette. Ydermere beregnes en ny værdi for den forventede kvindeandel af de optagede studerende, givet de nye oplysninger om, hvilke uddannelsesretninger under de to undertræer fokuserer på. Allerede ved dette skridt i træet er modellens forudsigelser interessante. Det fremgår, at i tilfældet hvor retningen for uddannelserne er business, ingeniør eller science vil den forventede værdi af kvindeandelen af de optagede studerende falde fra 0,52 til 0,32. Er der derimod tale om de resterende retninger for uddannelserne som eksempelvis samfundsvidenskab, sundhed eller humaniora vil den forventede kvindeandel stige til 0,6. Dette viser, at vi ved at indskræn-ke fokus til bestemte retninger inden for uddannelsesudvalget kan se en forskel i fordelingen af mænd og kvinder i uddannelsessystemet. Ser vi nu på de 28 % af data, hvor retningen er business, ingeniør eller sci-ence, vil modellen forudsige kvindeandelen af de optagede studerende til at være 0,4 hvis retningen er business og 0,29 for ingeniør eller science. Dermed ser vi, at retningerne ingeniør og science umiddelbart er nogle af de uddannelseskategorier, der ikke tiltrækker mange kvinder.

Det er interessant at se på knude...

Træet viser at det i beslutningsprocessen er inputvariablen retning der er den vigtigste, efterfulgt af gennem-snitsindkomst og afsluttende uddannelsesinstitution.

Kan bruges deskriptivt og
Hver knude viser - Den forudsagte andel af kvinder - Hvor procent af observationerne det gælder.

Figur 6: Regressionstræ



Konklusion