Eksamens projekt

My Table of content

Section 1

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 (Kilde). 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. tabellen nedenfor.

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).

TJEK: Kønsopdeling indenfor specifikke fagområder, e.g. tekniske uddannelser: søger mænd også her mod bedre betalte uddannelser?

Med kønsopdelt data for ansøgerantal samt optagne studerende på længere videregående uddannelse er det muligt at bestemme udstrækningen 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 uddannelse. Sluttelig analyseres det, hvorvidt forventet lønniveau kan forudsige kønsfordelingen blandt ansøgere på længere videregående uddannelser.

Data, metode og etik

Data er let offentligt tilgængelig og ligger åbnet på UFM's hjemmeside, derved har der ikke været rettighedsproblemer. Dog kan det problematiseres, at nogle uddannelser har meget få optagne, således at man vil kunne finde frem til de studerendes identitet vha. universiteternes biblioteker med BA opgaver eller muligvist med LinkedIn

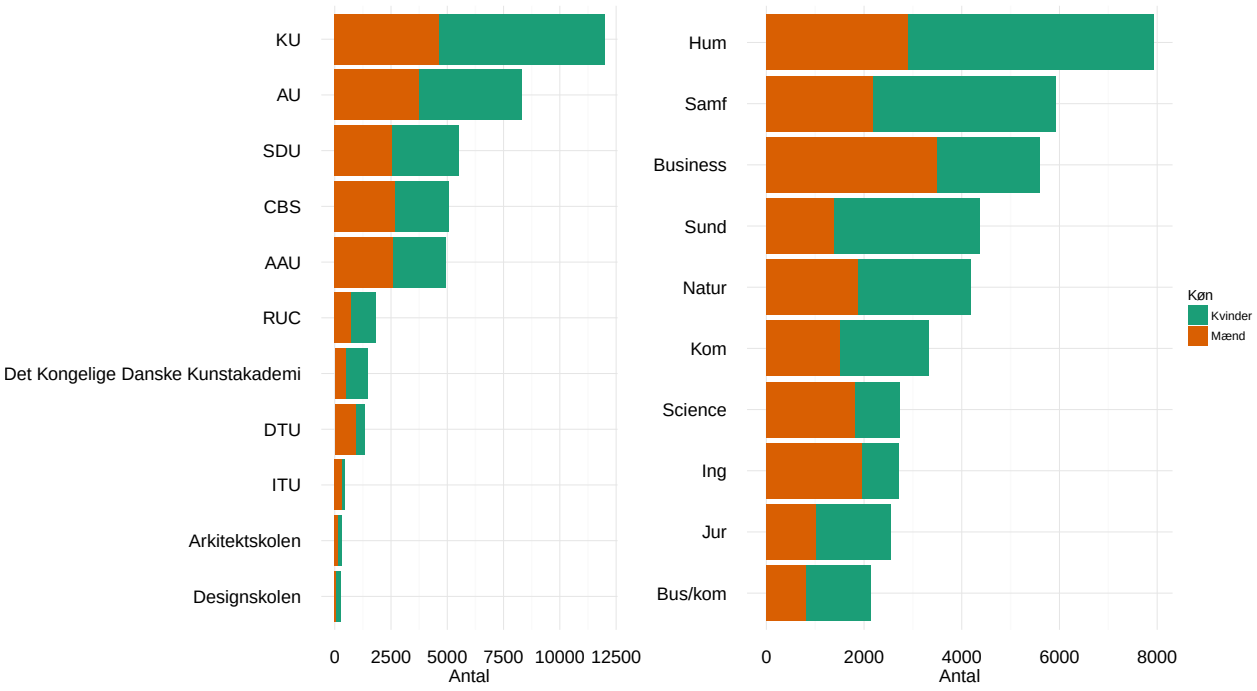
Metode Skrive her om hvilke datasæt vi har hentet og hvilke kilder: For hvert år har vi hentet: 1) ansøgninger fordelt på alle videregående uddannelser 2) optagne fordelt på alle videregående uddannelser 3) Adgangskvotienter for alle videregående uddannelser 4*3

Vi har hentet CEPOS undersøgelsen for bruttoløn opgjort efter uddannelse.

Dernæst gennemgik vi et datasæt kvalitativt henholdvis, at finde matchet mellem BA og færdig uddannelse og give hver uddannelse en retningskategori.

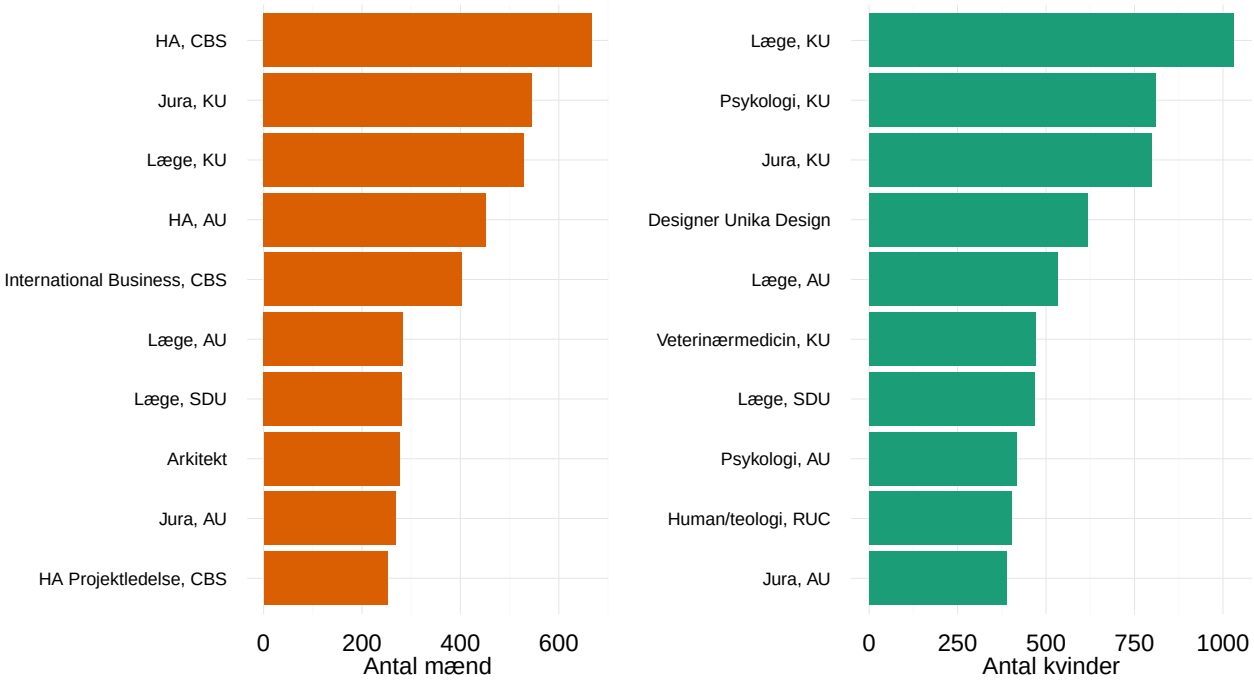
Vi joinerne alle uddannelserne sammen på deres unikke udannelsesnr. Dernæst joinet lønnen sammen fra CEPOS undersøgelsen ved benytte placeringsnr som nøgle.

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



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

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



Tabel 2: 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.

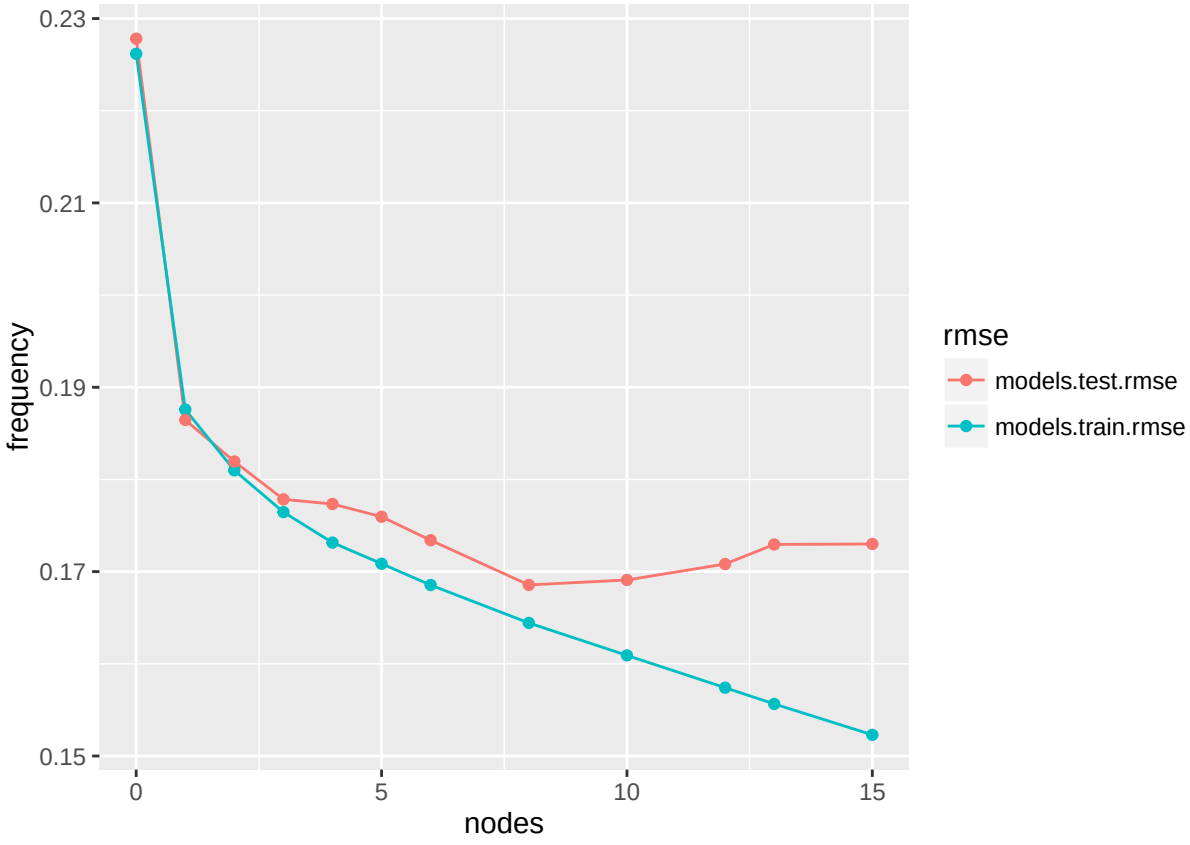
Ved brug af en *machine learning* algoritme udarbejdes et regressionstræ. Denne type regressionstræ benytter en delmængde af de tilgængelige data, kaldet træningssættet, til at klassificere studierne ved at "lære" forskelle i strukturer mellem uddannelser, og dermed forudsige udfaldet af kvindeandel. 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.



Den alternative valideringsmetohode, krydsvalidering, minder meget om ovenstående metode - dog betyder krydsvalidering at træningdata opdeles til 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. For hver af de udvalgte staf-keofficienter

vælges det undertræ, som minimerer cost-funktionen

En af fordelene ved krydsvalidering er formindsket risiko for stikprøve-bias, da data tilfældigt opdeles til træning og test. Udover krydsvalidering kunne bootstraaping også benyttes til at lave syntetetiske stikprøver.

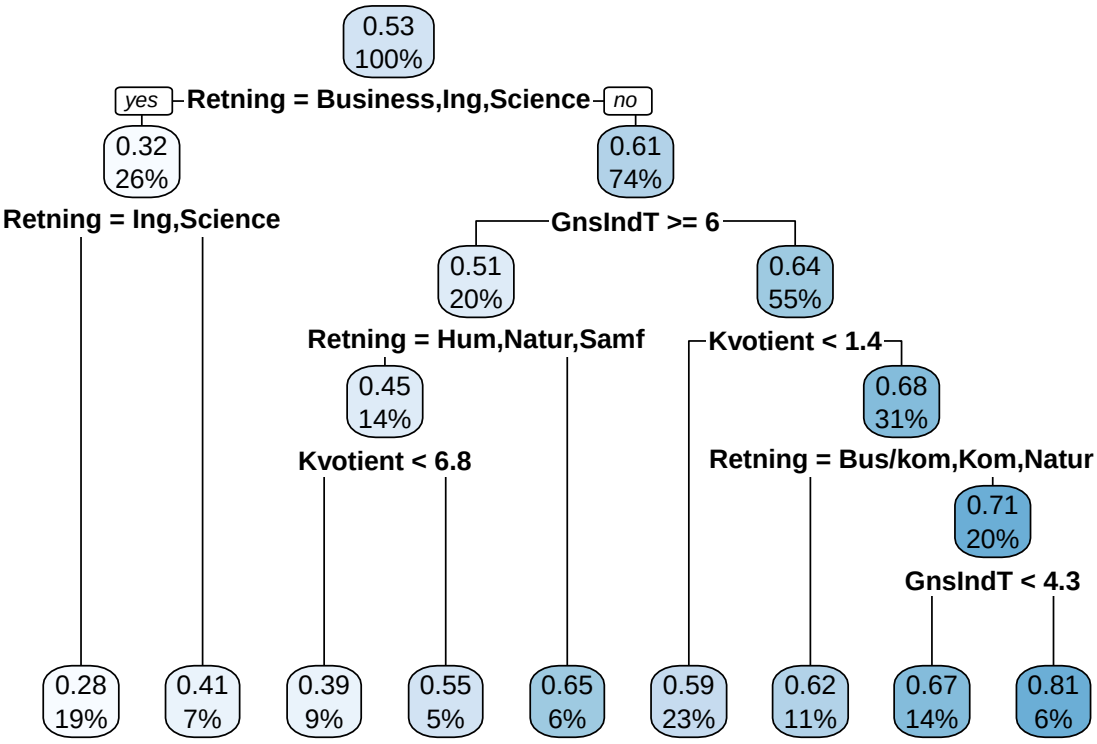
Regressionstræet

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 samfundsviden-skab, 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 busi-ness og 0,29 for ingeniør eller science. Dermed ser vi, at retningerne ingeniør og science umiddelbart er nog-le 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.


0.53
100%
yes
Retning = Business,Ing,Science
no
0.32
26%
0.61
74%
Retning = Ing,Science
GnsIndT >= 6
0.51
20%
0.64
55%
Retning = Hum,Natur,Samf
Kvotient < 1.4
0.45
14%
0.68
31%
Kvotient < 6.8
Retning = Bus/kom,Kom,Natur
0.71
20%
GnsIndT < 4.3
0.28
19%
0.41
7%
0.39
9%
0.55
5%
0.65
6%
0.59
23%
0.62
11%
0.67
14%
0.81
6%