Regression Tree

I et ønske om at kunne forudsige kvindeandelen af de optagede studerende på baggrund af udvalgte inputvariable udarbejdes et *regression tree.* Denne type beslutningstræ vil på baggrund af flere oplysninger forsøge at give et unikt resultat i form af en talværdi. I dette tilfælde benyttes variablene gennemsnitsindkomst, retning og uddannelsesinstitution som baggrundsoplysninger i forsøget på at forudsige kvindeandelen af de optagede studerende.

Idéen bag beslutningstræet, er at datasættet gentagne gange opdeles i mindre dele. De enkelte knuder i træet illustrerer denne opdelingsproces idet der til de enkelte knuder hører en mindre model, som netop splitter data op i mindre dele ved at fremsætte et spørgsmål, der skal besvares med ja eller nej. Vi ser altså, at der ved den første knude i træet tages udgangspunkt i hele datasættet. Vi kan herfra se, at modellen på baggrund af alle data for de tre 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 retninger der fokuseres på. Det er allerede ved dette skridt i træet interessant at se modellens forudsigelser. Vi ser, 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 indskrænke 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 science, 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…

Det er naturligvis muligt at udarbejde et beslutningstræ med utrolig mange knuder i den tro, at dette vil gøre modellens forudsigelser mere præcise. Det kan dog ske at træet passer for godt til det datasæt der ligger til grund for dette, hvormed det risikeres at modellens forudsigelser ikke vil være gode. (<http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Fourth%20Printing.pdf>, s. 307-311)

I udarbejdelsen af beslutningstræet udtrak vi tilfældigt halvdelen af observationerne i det oprindelige datasæt, hvilke blev brugt som *training set* og den resterende del af data fungerede således som *validation set.* Dette betyder at beslutningstræet udarbejdes på baggrund af træningsdatasættet og derefter testes på testdata, således at vi kan vurdere modellen. *cross validation* og benyttes

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