Metodebeskrivelse for model til identifikation af indtjeningspotentiale

Indholdsfortegnelse

1. Baggrund for og formål med projektet

Formålet med dette projekt har været at opbygge en statistisk model, der kan anvendes til at give et databaseret bud på indtjeningspotentialet for den enkelte landbrugsbedrift. Det bliver muligt ved at anvende data fra rigtig mange landbrugsbedrifter til at forudsige indtjeningspotentialet for den enkelte landmand. Herefter køres den enkelte landmands data igennem denne model, der som output kommer med et bud på landmandens indtjeningspotentiale, der kan opnås ved at ændre på en række forskellige parametre.

Motivationen bag er, at det er meget relevant for både landmanden og dennes økonomirådgiver at have et kvantitativt mål for, hvor meget landmanden kan spare ved at foretage en konkret ændring på bedriften. En barriere for dette er, at det kan være meget vanskeligt at overskue mængden af mulige tiltag, og derfor vil der være en tendens til, at de samme råd til at forbedre indtjeningspotentialet går igen fra bedrift til bedrift. Ved at bygge en statistisk model er det muligt at tage højde for den samlede tilstand for en bedrift, når der gives forslag til indtjeningspotentiale.

Det er vores ambition, at en færdig model vil kunne anvendes som et supplement til de nuværende fraktilanalyser og benchmarking, som i modsætningen til denne model er partielle modeller, som udelukkende anvender sammenligning af hver variabel for sig.

Fordelen ved den model, der præsenteres i afsnit 2, er, at den kan kvantificere effekterne af en mulig ændring på forhånd. På den måde kan værktøjet opfattes som en hjælp til både landmanden selv og rådgiveren, da den kan fremhæve de mulige indsatser med størst effekt. Hermed hjælper modellen til at ”finde nålen i høstakken”.

En anden vigtig pointe er, at outputtet fra modellen er fuldstændig individuelt. Det betyder, at de beregnede effekter af en indsats er individuelle for den enkelte bedrift, men med udgangspunkt i mønstre, der

er lært fra hele populationen af landbrugsbedrifter. En udløber af dette er, at én landmand vil have en positiv effekt af at øge udgifterne til dyrlæge og sundhed, mens en anden bedrift – men et andet udgangspunkt – vil have en negativ effekt.

Der er i projektet udelukkende arbejdet med en første udgave af modellen. Der er en række forslag til videreudvikling af modellen, som det er enten tidsmæssige eller teknisk/praktiske årsager ikke har været muligt at implementere i modellen på nuværende tidspunkt. Disse forslag kan ses i afsnit 4.

2. Fremgangsmåde

Dette afsnit indeholder en teoretisk gennemgang af, hvordan vi har opbygget den statistiske model, der skal anvendes til at prædiktere indtjeningspotentiale på en landbrugsbedrift. Herefter gennemgår vi et konkret eksempel på, hvordan modellen er tænkt til at kunne anvendes, samt erfaringer fra en præsentation for landmænd og rådgivere i næste afsnit.

Vi har valgt at operationalisere "indtjeningspotentiale" ved at tage udgangspunkt i en landmands fremstillingspris fra driftsgrensanalysen. Dette valg er foretaget med baggrund i, at

1. nøgletallet fremstillingspris er defineret ens på tværs af driftsgrene, hvilket har betydning for, at modellen kan implementeres på tværs af driftsgrene, og
2. det er relevant, da fremstillingsprisen har udviklet sig til et velkendt nøgletal, som landmænd anvender til at sammenligne sig med f.eks. en ERFA-gruppe.

2.1 Data

Vi har taget udgangspunkt i data fra Økonomidatabasen (forkortes ØDB fremover), da disse data i høj grad er validerede. Da fokus i dette projekt har været på at teste metoden, har vi valgt at anvende data fra en velkendt datakilde, og i stedet fokusere på at implementere en generisk metode.

Derudover har vi afgrænset os til i første omgang kun at se på kvægbedrifter. Det har vi gjort ud fra ønsket om at teste den fremgangsmåde, der bliver skitseret i dette afsnit. Valget af driftsgrenen 'kvæg' er arbitrær, og vi har også foretaget analyser ved at anvende modellen på svineproducenter. Erfaringer fra arbejdet med at anvende modellen på data fra en anden driftsgren er, at modellen er generisk, og at en i fremtiden eventuel endelig model forholdsvis nemt vil kunne implementeres på andre driftsgrene.

Udgangspunktet for analysen er det interne regnskab. I kvægmodellen er der yderligere afgrænset til disse tre driftsgrene:

1. Mælk
2. Grovfoder
3. Salgsafgrøder

2.2 Model

Vi har estimeret en statistisk model, der forklarer fremstillingsprisen på produktionsgren 6000 (mælk) ud fra godt 1.000 forklarende variable fra de seneste tre regnskabsår (2016, 2017 og 2018).

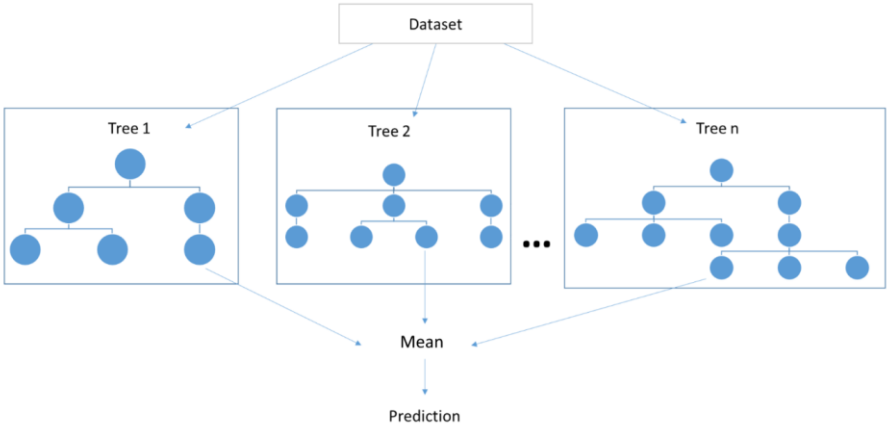
Vi har valgt at anvende Random Forest-algoritmen implementeret i det statistiske programmeringssprog R[1] som "motor". Det har vi gjort ud fra følgende overvejelser:

[1] R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/

1. Det er samme motor, der anvendes i både Risiko- og Rating-modellerne. På den måde har vi kunnet genbruge noget af R-koden fra de to modeller, og vi har også anvendt erfaringer fra arbejdet med disse modeller

2. Det er velkendt i litteraturen, at Random Forest er en algoritme, der ofte opnår meget gode resultater, men som kun har et lille antal parametre, der skal optimeres.

Kort beskrevet er Random Forest en *machine learning*-algoritme, der kan anvendes til både regression og klassifikation. Algoritmen tager udgangspunkt i en basismodel – som oftest, men ikke nødvendigvis, et beslutningstræ, hvilket også gør sig gældende i denne anvendelse.[2] Et beslutningstræ vil typisk have lav bias, men høj varians. Det kan Random Forest-algoritmen udnytte ved at beregne den endelige prædiktion som et gennemsnit[3] over et stort antal beslutningstræer, jf. nedenstående figur:



Et beslutningstræ er en statistisk metode, der forsøger at gruppere data i en række undergrupper på baggrund af variable i datasæt. For hvert "split" undersøges alle variable i datasættet, og den variabel, der giver den bedste opdeling i en undergruppe målt ved størst homogenitet i de to undergrupper, vælges. Herefter kan processen fortsætte, indtil man når et stopkriterium. Alternativt fortsætter algoritmen, indtil alle observationer er adskilt på laveste niveau. Metoden har fået sit navn, fordi den minder lidt om et træ, når den præsenteres grafisk. Det ses i ovenstående figur, hvor hver blå cirkel repræsenterer en variabel fra datasættet, der deler datasættet i to. Den endelige prædiktion findes som et gennemsnit over de observationer, der ender i samme gruppe.

For at mindske variansen ved at tage et gennemsnit over mange beslutningstræer er det nødvendigt, at de enkelte beslutningstræer har lav korrelation. Hvis ikke, så mindskes effekten. I ekstremtilfældet med perfekt korrelation mellem de enkelte beslutningstræer er det effektive antal træer kun ét, og der opnås ingen fald i modellens varians ved at anvende mange træer. Random Forest-algoritmen de-korrelerer de enkelte træer ved

1. at bygge hvert træ på en stikprøve fra det fulde data[4], og
2. kun at anvende en delmængde af de forklarende variable, når et enkelt træ bygges.

[2] Beslutningstræer anvendes også som basismodel i andre *machine learning*-algoritmer som f.eks. *bagging* og *boosting*, der dog kombinerer de enkelte beslutningstræer på forskellige vis. *Bagging* minder meget om Random Forest.
[3] Ved regression kan man også anvende medianen – ved klassifikation anvendes majoritetsreglen.
[4] Denne stikprøve findes typisk via *bootstraping*, dvs. man udvælger *N* observationer fra det oprindelige datasæt (hvor *N* er det samlede antal observationer i datasættet), hvor alle rækker har samme sandsynlighed for at blive valgt i hver iteration. Det gør, at den samme observation kan være medtaget flere gange, når et beslutningstræ bygges, men ikke er medtaget ved opbygning af andre træer.

2.3 Simulering af effekten

Med udgangspunkt i den estimerede Random Forest-model jf. ovenstående afsnit 2.2 har vi simuleret effekten af at ændre på 25 forskellige, men nøje udvalgte variable.[5] For flere af variablene gælder det, at det på forhånd er svært at vide, om der er en positiv effekt ved at sænke eller øge værdien af variablen. Derfor har vi også undersøgt begge effekter. Det er vores erfaringer, at effekterne i høj grad er afhængige af den enkelte bedrift. Dette uddybes i nedenstående afsnit 3, hvor forløbelige resultater fra modellen præsenteres.

Det er vigtigt, at de variable, der ændres eller "skrues" på for efterfølgende at kunne prædiktere effekten af denne ændring, er nøje udvalgt. Vi har taget udgangspunkt i et scenarie, hvor en variabel enten sænkes til 5% fraktilen for hele datasættet, eller hæves til 95% fraktilen. Det medførte imidlertid, at nogle af variablene blev ændret til værdier, der ud fra en faglig betragtning var meget urealistiske.

I stedet har vi fundet de nye værdier på denne måde:

1. Først opdeles bedrifterne i to hovedgrupper afhængigt af, om de har konventionel eller økologisk produktion

2. Herefter opdeles bedrifterne i fem grupper på baggrund af nøgletallet "Antal årskøer". Den variabel, som skal tildeles en værdi, bliver sat til at være lig med enten 5% eller 95% fraktilen i denne gruppe

3. Til sidst simuleres modellen fra afsnit 2.2 med denne nye værdi. Forskellen på den faktiske fremstillingspris for året (2018) og den prædikterede fremstillingspris, når der tages højde for denne ændring, gemmes

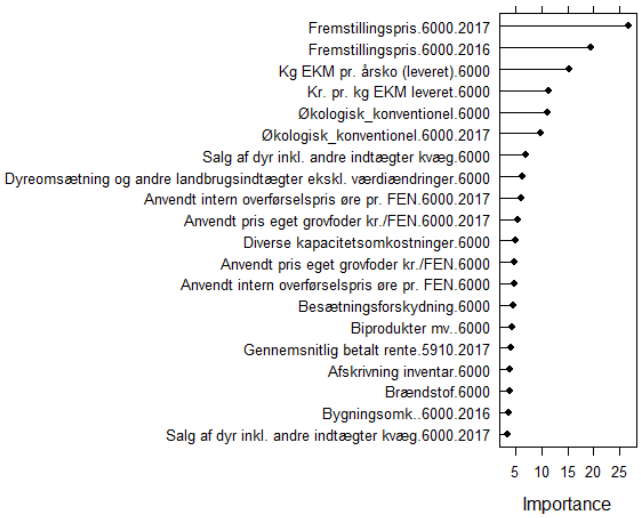
4. Denne proces gentages for alle variable.

Det afsluttende afsnit 4 indeholder bl.a. flere forslag til alternative fremgangsmåder.

3. Resultater

Figuren nedenfor viser de vigtigste variable i Random Forest-modellen.[6] Der er en tendens til, at fremstillingsprisen er persistent, da det er de to foregående års fremstillingspris, der er mest forklarende i forhold til dette års fremstillingspris. Det betyder, at der er en tendens til, at bedrifter med enten et højt eller lavt niveau for fremstillingspris vil fortsætte med samme niveau fremover.

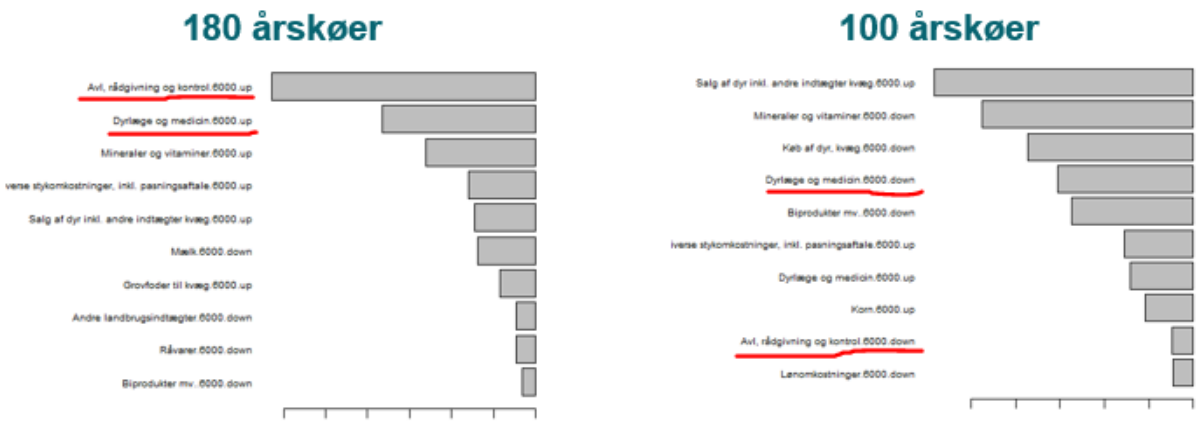
[5] En liste med alle variable kan ses i appendikset i afsnit 5.1)
[6] Den enkelte variabels betydning måles som et gennemsnit over det antal gange, den pågældende variabel er blevet anvendt til et "split" i et beslutningstræ, vægtet med, hvor meget dette split betyder for beslutningstræets evne til at danne gode prædiktioner. Hvis en variabel slet ikke anvendes til et split vil variablen have en *importance* på 0.



Derudover er indikatoren for økologi/konventionel også en meget relevant parameter, hvilket giver god mening ud fra en faglig betragtning, da afregningsprisen i hele den undersøgte periode er signifikant højere for økologisk mælk.

3.1 Et konkret eksempel

Nedenstående figur viser resultatet af testkørsler for to forskellige bedrifter med henholdsvis 180 årskøer (figuren til venstre) og 100 årskøer (til højre). Længden på hver søjle angiver indtjeningspotentialet ved at ændre den pågældende variabel enten op eller ned. Retningen er angivet som enten "up" eller "down" efter det sidste punktum på hver label. Alle effekter er negative, fordi indtjeningspotentialet er modelleret som et *fald* i fremstillingspris, hvilket alt andet lige vil give en højere indtjening.



Det fremgår tydeligt, at estimatet for den enkelte bedrifts indtjeningspotentialer er individuelt og tager højde for bedriftens nuværende udgifter. På figuren til venstre er de to effekter med højest indtjeningspotentiale at øge udgifter til "Avl, rådgivning og kontrol" henholdsvis "Dyrlæge og medicin". I figuren til højre er der positive indtjeningspotentialer ved at mindske udgifterne på de samme to områder (der er understreget med rødt i figuren).[7]

3.2 Gennemgang for rådgivere

Med henblik på at teste den praktiske anvendelighed præsenterede vi den 27. september 2019 de forløbelige resultater fra en gruppe bestående af omkring 10 landmænd og rådgivere fra næsten lige så

[7] Kategorien "Dyrlæge og medicin" fremgår to gange i figuren til højre, da det er besparelser at hente ved både at sænke og øge udgifterne. Det skal fortolkes på denne måde, at den pågældende bedrift ligger placeret højere end 95% fraktilen for udgifter til "Dyrlæge og medicin", så at sætte udgifterne lig 95% fraktilen er reelt en besparelse.

mange forskellige rådgivningsvirksomheder. Dette afsnit indeholder vores refleksioner over tilbagemeldingerne på baggrund af denne præsentation.

Den primære tilbagemelding i forhold til den model, vi præsenterede på mødet – og som også er beskrevet i dette notat – var, at det er meget positivt, at modellen er bedriftsspecifik, forstået på den måde, at den anvender data fra cirka 1.000 bedrifter til at estimere sammenhænge, men at outputtet tager højde for en specifik bedrifts data og dermed udgangspunktet for forandringer. Der var også flere, der påpegede, at denne egenskab gør modellen velegnet som et supplement til de eksisterende benchmarkværktøjer.

Der var flere rådgivere, der meldte tilbage, at de forestiller sig, at modellen vil blive forbedret, hvis der i højere grad inkluderes produktionsdata. Som beskrevet i afsnit 2.1 er modellen bygget på data fra SEGES' Økonomidatabase, og der er kun medtaget udvalgte og overordnede produktionsdata som f.eks. antal køer, produceret mælk (kg) samt de elementer, der indgår i fremstillingsprisen. Ved at inkludere produktionsgrensspecifikke data fra f.eks. produktionsstyringsprogrammet DMS vil det blive muligt at vurdere, om det er en yderligere indsats i forhold til f.eks. reproduktion eller forbedring af celletal, der vil give den højeste effekt på bundlinjen.

Den sidste tilbagemelding, der bør fremhæves her, handler om, at det ikke er så vigtigt, at outputtet fra modellen bygger på en statistisk model, der kan virke som lidt af en "black box" for landmænd og rådgivere.[8] Det er derimod vigtigt, at modeloutputtet har et format, som er formuleret i et kendt sprog, og som det er muligt at relatere til landmandens hverdag. Denne tilbagemelding hænger også sammen med, at det bør øge værdien af modellen, hvis der bliver tilføjet produktionsdata.

4. Forslag og anbefalinger

På baggrund af erfaringer og tilbagemelding fra landmænd og rådgivere vurderer vi, at modellen i sin nuværende form er færdig i første udgave. Det skyldes blandt andet, at modellen fungerer på konkrete bedrifter, som det fremgik af afsnit 3.1. Derudover vægter den positive tilbagemelding fra landmænd og rådgivere naturligvis højt.

Da der alene er tale om en første udgave, er der en række forslag til videreudvikling af modellen, som er opstillet og beskrevet nedenfor.

1. Inkludere produktionsdata. Som nævnt i forlængelse af vores oplæg for gruppen af landmænd og rådgivere, vil det være interessant at opdatere modellen med data fra produktion, f.eks. DMS (kvæg) eller Cloudfarms og Agrosoft (svin). Den metode, der er beskrevet i dette notat, vil nemt kunne udvides til at inkludere disse data og samtidig simulere effekten af ændringer.

2. Højere frekvens på data. Modellen inkl. simulering af effekter kan nemt opdateres, når der kommer nye data. Det betyder, at der er et potentiale for at anvende modellen flere gange i løbet af året, hvis der anvendes data, der bliver opdateret med en højere frekvens end årlig.

3. Da det er vigtigt, at hver variabel ændres til en værdi, der både giver faglig mening og er databaseret,[9] kan det evt. være givtigt i stedet for at opdele bedrifterne på baggrund af en cluster-analyse (f.eks. *k-means*) og anvende 5% henholdsvis 95% fraktilerne fra disse grupper som de nye værdier.

[8] Denne bemærkning kan muligvis hænge sammen med, at mange landmænd allerede anvender værktøjer til f.eks. planlægning af markdriften, der bygger på statistiske modeller.
[9] Det skyldes, at den endelige model gerne skal kunne køre på en ny bedrift uafhængig af menneskelig involvering. Et alternativ hertil vil være, hvis en landmand eller rådgiver selv har mulighed for at indtaste en ny værdi.

4. Et alternativ til at anvende ovenstående fraktil-tilgang vil være i stedet for at simulere en ændring i hver variabel på eksempelvis $\{-50\%, -40\%, \ldots, -10\%, 10\%, \ldots, 40\%, 50\%\}$ for at opnå et estimat for det marginale indtjeningspotentiale

5. Det er også muligt at anvende prædefinerede scenarier, hvor flere korrelerede variable ændres samtidigt. Ulempen ved dette er, at scenarierne skal defineres på forhånd, men det kan evt. kombineres med forslaget i punkt 4)

6. Et muligt færdigt produkt vil kunne implementeres på den måde, at en landmand logger ind på landmand.dk med sit DLI-login og bestiller en 2-4 siders pdf-rapport med de grafer, der er vist i eksemplerne i afsnit 3.

5. Appendiks

5.1 De 25 udvalgte variable til simulering

Landbrugsarealer dyrkbart, ha
Landbrugsareal dyrkbart, ha (forpagtet) (inkl. andel af 310880,882,885)
Årskøer, Stk.
Kg EKM leveret
Årsopdræt, Stk.
Udskiftningsprocent
Bruttoudbytte fordelt pr. produktionsgren i driftsgrensanalyse
Bruttoudbytte som ikke stammer fra grovfoder
Sum af dyrlæge, medicin, avl rådgivning og kontrol, diverse vedr. husdyr og andet
summen af foderomkostninger i driftsgrensanalysen
Sum af energi, maskinstation, vedligehold, forsikring og diverse kapacitetsomkostninger fra driftsgrens-analysen på kvæg og svin. Samlet omtales det som "øvrige kapacitetsomkostninger"
Sum af energi, maskinstation, vedligehold, forsikring og diverse kapacitetsomkostninger fra driftsgrens-analyse grovfoder
Stykomkostninger markbrug fra Driftsgren grovfoder
Fremstillingspris, grovfoder, øre pr. FEN
Kapitalomkostninger konteret - til driftsgren grovfoder
Kapitalomkostninger konteret - til driftsgren
kapitalomkostninger fordelt på grovfoder i driftsgrensanalyse
Kapitalomkostninger fordelt pr. produktionsgren i driftsgrensanalysen
Mælk
Bruttoudbytte
Lønomkostninger
Mælk
Lønomkostninger
Ejeraflønning
Udbytte, grovfoderafgrøder, FEN