Projekt hos Thise Mejeri

Kenneth Gottfredsen, Eva Rauff og Sanne Sørensen

1/1/23

1 Problemfelt

Det er vanskeligt at forudsige hvor meget koldskål Thise skal producere til deres kunder (Kjer 2022). Sommervejret påvirker kundernes efterspørgsel på koldskål (Holland 2022). Hos Thise Mejeri bruger medarbejderne vejrudsigten, og mange års erfaring når de skal vurdere, hvor mange ltr. koldskål, der skal produceres (Jensen 2022). I en artikel fra TV Midtvest fortæller adm. direktør Poul Pedersen fra Thise Mejeri at: *"Vi kigger på tal og vejrudsigter som aldrig før. Så beslutter vi os for et tal, og vi plejer at være gode til at gætte." - (ibid).* Det er et erhvervsøkonomisk problem, hvis man udelukkende bruger vejrudsigten og gætteri til at forudsige, hvor meget koldskål produktionsafdelingen skal producere. Gætteriet medfører en betydelig risiko, da man ikke kan garantere, at alt koldskålen bliver solgt - selv når vejret er godt! (ibid). Konsekvensen kan føre til et stort økonomisk tab, fordi den mængde koldskål som ikke bliver solgt i butikkerne, leveres tilbage til Thises eget lager igen. Afhænger COOPs efterspørgsel på koldskål kun af vejret? Eller er der andre mekanismer end vejret, som forårsager en stigende eller faldende efterspørgsel på koldskål? Formålet med undersøgelsen er, at undersøge hvordan Thises datamodenhedsproces hænger sammen med COOPs efterspørgsel på koldskål. Undersøgelsen skal bidrage med en multibel lineær regressionsmodel, som kan give en tilnærmelsesvis præcis forudsigelse af, hvor mange ltr. koldskål COOP vil efterspørge fremadrettet.

1.1 Problemformulering

På baggrund af ovenstående problemfelt bliver der i næste afsnit formuleret et hovedspørgsmål og nogle underspørgsmål, de skal besvare den erhvervsøkonomiske problemstillingen.

1.1.1 Hovedspørgsmål

"Hvordan kan Thises produktionsafdeling forbedre deres datamodenhed og dermed forbedre udnyttelsen af deres egne data til produktionsplanlægning?"

Hovedspørgsmålet er *løsningsorienteret*. Fordi spørgsmålet skal bidrage med datainitiativer som Thise selv kan benytte i sig af under produktionsplanlægningen. Fordi datainitiativerne vil forbedre deres evne til at bruge data, den selv producerer, til og skabe større økonomisk værdi.

1.1.2 Underspørgsmål 1

"Hvor befinder Thises produktionsafdeling sig på nuværende tidspunkt datamodenhedsmæssigt?"

Formålet med spørgsmålet er, at *beskrive* hvor langt i datamodenhedsprocessen Thises produktionsafdeling befinder sig på nuværende tidspunkt.

1.1.3 Underspørgsmål 2

"Hvilke faktorer påvirker COOPs efterspørgsel af koldskål?"

Formålet med spørgsmålet er, at *beskrive* hvilke variable der hænger sammen med COOPs efterspørgslen på koldskål.

1.1.4 Underspørgsmål 3

"Hvilken model kan Thises produktionsafdeling bruge til at forudsige COOPs efterspørgsel af koldskål fremadrettet?"

Formålet med spørgsmålet er, at *identificere* den bedste regressionsmodel og de variable som produktionsafdeling skal bruge, når de skal forsøge at forudsige Coops efterspørgsel af koldskål.

1.1.5 Underspørgsmål 4

"Kan undersøgelsens resultater anvendes til at styrke produktionsplanlægningen?"

Formålet med spørgsmålet er, at *perspektivere* undersøgelsens statistiske analyse til en anden produktionskontekst.

2 Videnskabsteori

Et paradigme er et tankesystem (Bergfors 2021). Sammenhængen mellem Thises datamodenhedsproces, vejrforholdene og efterspørgslen af koldskål er kompleks. Der er derfor behov for både kvalitativ og kvantitativ viden til, at besvare problemstillingen.

Undersøgelsen styrer hen mod det *realistiske* paradigme, fordi tilgangen både fokuserer på tal og sproget. Formålet er at forklare sammenhængene med tal, og dermed beskrive, forklare og forstå hvorfor tallene ser ud, som de gør. Antagelsen er at Thises datamodenshedsproces hænger sammen med Coops efterspørgsmål på koldskål samt andre vejr-variables effekter herpå. *Ontologi* er en antagelse om, hvordan man anskuer den verden, problemstillingen indgår i (Bergfors 2021). Den ontologiske opfattelse er, at virkeligheden hos Thise Mejeri eksisterer uafhængigt af medarbejdernes egne opfattelser, da sandheden om Thises datamodenhed og efterspørgslen på koldskål forstås objektivt (Bergfors 2021). *Epistemologi* handler om, hvordan man anskaffer viden om en problemstilling. Undersøgelsen bruger både kvalitative og kvantitative data som bruges i kombination med hinanden, da forklaringer skal kunne generaliseres - dette kaldes for metodekombination (ibid)

2.1 Undersøgelsesdesign

Et undersøgelsesdesign er en strategisk plan som skal besvare en problemstilling ud fra empiriske data (Bergfors 2021). Vi anvender et *statisk* undersøgelsesdesign til, at besvare vores problemstilling. Fordi designet i sig selv giver et her - og nu billede. Variablerne måles og observeres fx. fra perioden 1/4/22-30/8/22, hvilket er en specifik periode. Dertil undersøges relationen mellem vejrvariable og efterspørgslen på koldskål i deres nuværende tilstand (ibid).

Designtypen er et *casestudium*, fordi vores problemstilling tager afsæt i en nutidig kontekst, og fordi et casestudie undersøger komplekse sammenhænge i dybden (ibid). Casen omhandler Thise Mejeri som virksomhed. Formålet er at *beskrive* og *forstå* det komplekse samspil der hænger sammen med Thises datamodenhedsproces, vejret og COOPs efterspørgsel på koldskål.

2.2 Metodologi

Metodologi henviser til de forskellige metoder, man kan bruge til at indsamle data med (ibid). Der anvendes eksisterende kvantitative data til en regressionsanalyse. Denne data skal skabe større forståelse for, hvorfor forskellige vejr-variable påvirker Coops efterspørgsel af koldskål. Der bruges eksisterende sekundære vejrdata hentet med en API fra Danmarks meteorologiske instituttet (DMI), og produktionsdata om koldskål hentet fra Thises VMI-system. Som supplement til de kvantitative data kombineres der med nogle kvalitative data i form af to semistrukturerede interviews med to informanter. Den ene informant arbejder i produktionsafdelingen, og den anden arbejder i salgsafdelingen. Formålet med den kvalitative data er, at disse dybdegående oplysninger kan bruges til og forstå datamodenhed ud fra et subjektivt medarbejderperspektiv.

Den statistiske metode som anvendes i denne undersøgelse kaldes for superviseret metode, fordi den tager udgangspunkt i en afhængig variabel. Den specifikke metode er en *multibel lineær regression*. Den bruges til at forudsige efterspørgslen på koldskål i liter ud fra effekten af flere uafhængige vejr-variabler. Dette betegnes som prædiktion (Hastie et.al 2021). Den generelle formel er:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 ... + \hat{\beta}_p x_p + \epsilon$$

$\hat{y}$ er den forudsagte værdi af Y og $\hat{f}$ er et estimat for f. $\hat{Y}$ er desuden den afhængige variabel efterspørgsel i liter. x_i er udvalgte uafhængige variabler. $\hat{y} = f(X)$ indeholder variation som vi kan reducere ved, at bruge den korrekte SL-metode til, at beregne f med. Men det vil aldrig være en fejlfri model vi ender ud med. Fordi estimatet ϵ er tilfældige fejl eller støj, man ikke kan gøre noget ved og den type fejl (Hastie et.al 2021).

3 Thises forretningsmodel

Et business model canvas er et illustrativt og strategisk værktøj som bruges til, at forstå hvordan en forretningsmodel skaber økonomisk værdi (Ostervalder & Pigneur 2010). Det bruges i denne sammenhæng til, at forstå Thises forretningsmodel.

Thise Mejeri bruger et VMI system som salgskanal. Et VMI system er et leverandørstyret system. COOP bestiller deres koldskål igennem dette system, og Thise producere dernæst koldskålen og opbevarer det på deres eget lager. Dernæst transportere de selv koldskålen ud til COOPs butikker. Thise har en intern aftale med COOP om, at de skal have mindst 80% af af produkterne på lageret. De resterende 20% supplerer de op med vha. Pareto forecasting, som er et endnu et system i deres salgskanal. 80/20 reglen fra Pareto fungerer som et management værktøj, da Thise hele tiden skal tilpasse deres salgsstrategier for at opretholde deres leveringsservice på ca. 98%.

Pareto er et godt værktøj til, at sikre at Thise opnår forskellige KPI-målsætninger. Problemet med Pareto er, at det er abonnementbaseret og dyrt, fik vi fortalt af Bjarne som er senior demand planner (Justesen 2022: 59). Der er stor risiko forbundet ved, at have et VMI system. Bliver et produkt ikke solgt, leveres det tilbage til Thises hovedlager, hvor det registreres på en spild konto (ibid). Derfor er det vigtigt, at have nogle pålidelige forecasting modeller der kan forudsige efterspørgslen af produkter med stor præcision. Thise får nemlig først betaling når varen er solgt i COOPs butikker.

Det kræver et tæt samarbejde, at have et VMI system kørende. Derfor er det vigtigt at Thise overholder deres produktvilkår til COOP, da de er bundet af klausulaftaler som kan medføre forskellige sanktioner af økonomisk karakter, så frem de ikke overholdes. Stærke kunderelationer er derfor vigtige i den her sammenhæng.

4 Thises datamodenhed

Alexandramodellen er en datamodenhedsmodel. Den bruges til at finde af hvor langt en virksomhed er i datamodenhedsprocessen, og hvordan den kan bruge forskellige strategier til og blive mere datadrevne (Bækby et. al 2017). Modellen bruges til at undersøge hvor datamodne Thise er på nuværende tidspunkt.

Thise Mejeri er på nuværende tidspunkt i fase 1, da de registrerer data for at skabe en forståelse og værdi for virksomhedens drift (ibid). Produktionsplanlæggeren har via. Navision adgang til alt virksomhedsdata som fx. lagerstyring og finansiel styring gennem deres Windows styresystem. Dette er en on-premises-løsning. Navision er et ERP softwaresystem. Det bruges til at styre virksomhedens forretningsprocesser og dermed effektivisere driften. og kan dermed planlægge fremtidig produktionen. At de anvender programmet direkte gennem deres Windows styresystem gør virksomheden sårbar overfor systemnedbrud, hvilket de oplever ca 5 gange om året. Kompetencemæssigt, ifølge Bjarne, har Thise Mejeri en brist da de ikke har ansat nogle medarbejdere med relevant analyse

erfaring (Justesen 2022: 45). De er derfor afhængige af den eksterne udbydere til at forecaste med de data, der bliver indsamlet.

Produktionsplanlæggeren har via. Navision adgang til alt virksomhedsdata gennem deres Windows styresystem, hvilket er en on-premises-løsning. Navision er ERP software til at håndtere økonomi data til og træffe beslutninger ud fra. og kan dermed planlægge fremtidig produktionen. At de anvender programmet direkte gennem deres Windows styresystem gør virksomheden sårbar overfor systemnedbrud, hvilket de oplever ca 5 gange om året. Kompetencemæssigt, ifølge Bjarne, har Thise Mejeri en brist da de ikke har ansat nogle medarbejder med relevant analyse erfaring. De er derfor afhængige af den eksterne udbydere til at forecaste med de data, der bliver indsamlet (ibid).

5 Introduktion til dataanalysen

Thise Mejeri har i dag svært ved, at forecaste hvor stor efterspørgslen bliver på Koldskål i løbet af deres koldskålssæson, som ca. løber i ugerne 13-36. Thise Mejeri ønsker, at opretholde deres leverings service på 98% på alle de produkter de har lovet levering på fortæller Søren Jensen.

5.1 Baggrund

Den kvantitative del af analysen er afgrænset til COOP butikker i nærheden af Landbohøjskolen, hvis beliggenhed er i Københavnområdet. Butikkerne afgiver ordre til Thises fjernlager, hvorefter koldskålen bliver leveret ud til butikkerne.

5.2 Formål

Formålet med analysen er derfor, at udregne en multibel lineær regressionsmodel som bedst kan forudsige butikkernes efterspørgsel på koldskål i området omkring Landbohøjskolen. Derudover vil vi også finde ud af, hvordan vejret og andre vejr-relateret faktorer påvirker butikkernes efterspørgsel på koldskål. Tilgangen kan løse Thises forretningsproblem. Undersøgelsens dataminingproblem går ud på, at identificere de forskellige vejr-variablers effekt på efterspørgspørgslen af koldskål.

5.3 Importer data til R

Først importeres datasættet ind i R-Studio:

5.4 Tidying og transformering af datasæt

Nu har vi fået indlæst datasættet. Det næste skridt er gøre strukturen i vores dataframe nemmere at arbejde med og mere læsevenlig. Processen kaldes for tidy data, det betyder at hver variabel har en kolonne, hver observation en række, samt hver observationsenhed er i en tabel (Wickham 2022). Det gør analysearbejdet nemmere. Først rekodes nogle af variablerne, så de stemmer overens med hvad der står i opgavebesvarelsen.

For at indhente data fra DMI laves en HTTP GET-anmodning til en API fra DMI. Vi skal bruge adgangen til at få de relevante vejr-variable som vi senere skal bruge i vores analyse. API'en leverer til slut et objekt i JSON format som bliver transformeret om til en dataframe i stedet for en liste. Først bruges base_url og info_url til at anmode om vejrdata fra DMI's API. req_url bruges til at udvælge specifikke parametre fra API´en.

I næste kodechunk vil vi transformere den data vi har hentet fra vores API-kald til nogle mere brugbare data. Først bruges base_url og dernæst info_url til at anmode om vejrdata fra DMI's API. req_url bruges til at udvælge specifikke parametre fra API´en. Derefter bruges `pivot_wider()` til at fordele variablerne ud i deres egne separate kolonner. mutate-funktionen bruges til at konvertere kolonnen målingstidspunkt til datoformat. `Separate()` bruges til at opdele kolonnen målingstidspunkt i to separate kolonner som navngives 'date' og 'time'. `Filter(str_sub)`funktionen udvælger rækker, der indeholder de første fire karakterer: "12:0".

I nedestående kode-chunk bruges `left_join()` til at returnere alle rækkerne fra x, samt alle kolonner langs x og y (Wickham 2022). Udvalgte dataframes sammenkobles fra data1 og data2 til data3. Da alle kolonnerne er lige lange, er der ingen missing værdier. Dernæst anvendes mutate til, at oprette fire nye variabler i data3 kaldet temp_gt25_3_dage. `Lag()` er brugt til at lave variablerne, som har opfanget forsinkede værdier fra temp1, temp2 og temp3. Afslutningsvis dannes variablen 'temp_gt25_3_dage', som måler de dage hvor der har været mere end 3 dage i træk med >= 25 grader. Det er en dummyvariabel fordi der bruges `if_else`. Relevante variabler bliver beskrevet når der tolkes på modelparametre i regressionsanalysen.

5.5 Datavisualisering og eksplorativ analyse

Nu er data blevet gjort tidy. Næste skridt er at undersøge hvilke faktorer der kan hænge sammen med butikkernes efterspørgslen af koldskål. I afsnittet begynder den eksplorative analyse. `geom_density()` funktionen bruges til, at forstå fordelingen og til at forudsige den forventede fordeling af efterspørgslen på koldskål. Man kan se at at spredningen af observationerne er størst omkring 520 liter, dette er gennemsnittet. Observationerne er normalfordelte. At de er normalfordelt er en fordel, fordi den lineære regressionsmodel er en parametrisk test, hvor ét af kravene er data skal være normaltfordelt (Hastie et.al 2021). At kravet er opfyldt gør model-parametrene er mere pålidelige.

I følgende kode-chunk er der vist et boxplot. Det skal vise den statistiske variationen ift. butikkens forventede lagerbeholdning og efterspørgslen på koldskål. Man kan se at median-efterspørgslen

stiger når fra høj til lav forventet lagerbeholdning af koldskål. Det tyder på at der er en signifikant sammenhæng mellem de 2 variabler.

I næste kode-chunk er der lavet et boxplot som viser fordelingen af efterspørgslen i forhold til om 25% af butikkerne er løbet tør for kammerjunkere eller ej. På baggrund af plottet kan man se at hvis butikkerne ikke har kammerjunkere på lageret så falder efterspørgslen. Det betyder efterspørgslen på koldskål stiger hvis butikkerne er løbet tør for kammerjunkere.

Forneden er et boxplot som viser sammenhængen mellem måned og efterspørgslen af koldskål. Det er tydeligt at se at median-efterspørgslen stiger fra april-juli hvorefter den efterspørgslen igen falder i august.

I nedestående kodechunk er der udvalgt 6 kontinuerte vejr-variabler, fordi ønskes er, at undersøge om disse er korreleret med hinanden, og om deres indbyrdes korrelation er statistisk signifikant. Der anvendes en `chart.Correlation()` til at foretage en korrelationsanalyse.

Efterspørgel og humidity er ikke korreleret med hinanden, og dermed ikke statistisk signifikant. Beslutningen er derfor, at humidity ikke vil blive inkluderet i analysen. Efterspørgsel og den gennemsnitlige temperatur per time har en moderat korrelations koefficient på 0.38. P-værdien er meget lav med tre stjerner, det betyder at sammenhængen er signifikant, og det er derfor usandsynligt at opnå et mere ekstremt resultat, hvis man foretog en ny undersøgelse. Mindre end 5% af korrelationen skyldes derfor tilfældighed. Alle temperatur-variablerne er tæt på 1, hvilket betyder at de har stærk samvariation. Dette kaldes for multikolinearitet. Det vil sige, hvis de blev brugt i den endelige model ville det være vanskeligt, at fortolke på koefficienterne. En Model med høj multikolinearitet bliver mindre præcis og mindre pålidelig. For at reducere multikolinearitetet fjernes de øvrige temperatur-variable.

Som førnævnt var der en moderat signifikant sammenhæng mellem efterspørgslen og gennemsnits temperaturen. Derfor bruges temp_mean_past1h som den uafhængige effekt i næste kodechunk. Der anvendes `predict()` til at konstruere et 95 prædiktionsinterval, efterfulgt af `geom_smooth()` til og visualisere sammenhængen med et scatterplot.

Ud fra scatterplottet kan man se, at forholdet mellem den gennemsnitlige temperatur og butikkernes efterspørgsel på koldskål er moderat lineært. Fordi hældningen på tendenslinjen er positiv. Det antages at når gennemsnits temperaturen stiger én enhed, vil efterspørgslen stige tilsvarende, det 'forholdsvis' mange af datapunkterne er placeret omkring tendenslinjen. Der anvendes lineær regression, fordi det er en simpel metode, Og det er nemt at tolke på model-parametrene.

Mange af datapunkterne ligger også langt væk fra tendenslinjen, der udtrykker en stigende gennemsnitlig efterspørgsel på koldskål i liter. Det indikerer at der er stor varians og potentiel bias tilstede. En mere kompleks model kan derfor anvendes til, at forklare sammenhængen. Flere af observationerne er placeret udenfor dette bånd, hvorfor det er besluttet at anvende et prædiktionsinterval i stedet - der er den røde stiplede linje. Formålet er med andre ord, at medregne usikkerheden omkring de individuelle værdier og ikke usikkerheden omkring gennemsnittet.

Når den gennemsnitlige temperatur hver time er 30 °C, er efterspørgslen på koldskål for én ny observation 629.87 liter. Ved samme temperatur vil butikkernes efterspørgslen af koldskål med

95% sikkerhed være [369.30:890.43]. Man kan på baggrund af nedestående output tydeligt se, at hvis den gennemsnitlige temperatur i °C stiger, stiger butikkernes efterspørgsel på koldskål tilsvarende.

5.6 Træning på træningsdata

først trænes modellen på træningsdata, fordi vi gerne vil tilpasse modelparametrene. Der anvendes træningsdata til, at fintune vores regressionsmodel. Efter modellen er blevet trænet godt igennem, bliver den afprøvet på testdata, da man gerne vil undersøge hvor god modellen er til, at forudsige en så præcis efterspørgsel på koldskål som mulig. Vurderingen af modelpræcisionen bestemmes ud fra den laveste MSE værdi. MSE måler hvor langt den forudsagte værdi for en observation er fra den faktiske værdi for en observation. Er MSE lille er der den forudsagte værdi tæt på den faktiske værdi, er MSE stor er den forudsagte værdi langt fra den faktiske værdi. MSE er således et udtryk for, hvor præcis den udvalgte model er til, at forudsige efterspørgslen af koldskål (Hastie et.al 2021). MSE skal være så tæt på 0 som muligt.

Antallet af variabler i det samlede datasæt er mindre end antallet af observationer. Derfor bruges backward-selection til, at udvælge de uafhængige variabler som fremadrettet skal indgå i modellerne. Det vil sige, at vi tilføjer alle variable ind på højre side af ligningen, og fjerner dem med den højeste p-værdi indtil der kun er signifikante uafhængige variable tilbage (Hastie et.al 2021). Denne teknik kan hjælpe med at reducere unødvendig varians i den udvalgte model. men på samme tid er den effektiv til, at identificere vigtige relationer i datasættet (ibid).

5.7 Test på testdata

I det foregående afsnit blev den gennemsnitlige MSE beregnet for hver af de fire modeller på træningsdata. Vi er egentlig ligeglade med disse MSE værdier. Det er mere interessant at se hvor præcise forudsigelserne er på testdata. Træningsdata anvendes som førnævnt til, at udvælge signifikante uafhængige variable og tilpasse modelparametrene.

Dog er det værd at nævne, at den data undersøgelsen er baseret på simulerede data. Dvs. at f er kendt allerede. Den virkelige sandhed om efterspørgslen af koldskål vides dog ikke. Men hvis der bliver udtrukket nogle testdata ud fra data3, kan man validere hvor godt en model performer på disse testdata, når modelkompleksiteten øges. Kompleksiteten kan øges ved, at de kontinuerte variable opløftes i flere potenser, eller ved og inkludere flere uafhængige variabler.

5.8 Valg af metode til at teste model performance

Man kan producere testdata på flere måder. Der anvendes en $LOOCV$ metode, fordi data3 kun indeholder 151 observationer i alt. Fordelen ved fremgangsmåden er, at den træner på alle observa-

tionerne, undtagen ét datapunkt. Processen gentages i dette tilfælde 150 gange. Derefter beregnes en gennemsnitlig MSE score, som udtrykker hvor god modelpræcisionen er (Hastie et.al 2021).

Problemet med metoden er, at det kræver stor computerkraft, det er fordi modellen trænes k gange (ibid).

6 Resultater

Nu er de fire modeller blevet trænet på træningsdata og testet godt igennem på testdata. Resultaterne tager kun udgangspunkt i koefficienterne fra de fire testmodeller. Modellen med den laveste kvadrerede RMSE, og den højeste R^2 er den model som har størst prædiktiv præcision.

Baselinemodellen har kun den afhængige variabel i ligningen. $\hat{\beta}_0$ er den gennemsnitlige efterspørgsel på koldskål ved 520.23 ltr. $RMSEtest = 139.20$. Bliver Temp_mean_past1h inkluderet i den simple model, falder $RMSEtest = 130.36$, det betyder at modellen 'fitter' bedre på data og at modellen batter lidt mere. Inkluderes de kategoriske faktorer i modellen med moderat kompleksitet, falder $RMSEtest = 88.55$. Dette er den model hvor den forudsagte værdi er tættest på den faktiske værdi. Dertil er $adjR^2 = 0.63\%$, det referer til den proportion af butikkernes efterspurgte koldskål som bliver forklaret af de uafhængige variabler. De forklarer altså 63% af den samlede varians i datasættet.

Bliver modellen mere kompleks, bliver præcisionen ikke bedre - ofte gælder det modsatte! I den komplekse model stiger $RMSEtest = 89.37$ og $adjR^2 = 0.62\%$, dvs. den begynder at falde. Desuden bliver Temp_mean_past1h^22 insignifikant, da hældningen er 0. Dette skyldes overfitting. Dette kommes der nærmere ind på i diskussionsafsnittet.

Den moderate model er blevet udvalg til den mest præcise model. Fortolkningen af koefficienterne er, når de øvrige variabler holdes konstant:

- Er den forventede lagerbeholdning på mellemste niveau, falder butikkernes gennemsnitlige efterspørgsel på koldskål med -76.57 ltr, i forhold til butikker med en lav forventet lagerbeholdningen.
- Er den forventede lagerbeholdning på højeste niveau, falder butikkernes gennemsnitlige efterspørgsel på koldskål med -82.67 ltr, i forhold til butikker med en lav forventet lagerbeholdning. Ændres referencegruppen, stiger efterspørgslen tilsvarende.

Det giver umiddelbart god mening, da butikkerne ikke vil risikere at bestille for meget koldskål. Da der er risiko for, at den ikke bliver solgt, og dermed øges riskoen for, at koldskålen overskrider den sidste salgsdato.

- Er 25% af butikkerne i det pågældende område ikke løbet tør for kammerjunkere, falder butikkernes gennemsnitlige efterspørgsel med -71.89 ltr, sammenlignet med de 25% af butikkerne i det pågældende som er løbet tør for kammerjunkere. Ændres referencegruppen, stiger efterspørgslen tilsvarende.

Butikkerne vil gerne lave mersalg og dermed sælge kammerjunkere sammen koldskål, det kan også indikere at forbrugerne synes koldskål og kammerjunkere skal spises sammen.

- I maj måned stiger butikkernes efterspørgsel på koldskål gennemsnitligt med 84.37 ltr, i forhold til april måned. I juni måned stiger efterspørgslen gennemsnitligt med 129.51 ltr, i forhold til april måned. I juli måned stiger efterspørgslen gennemsnitligt med 155.95 ltr, i forhold til april måned. I august falder efterspørgslen ned til 85.10 ltr, sammenlignet med april måned.

Det betyder, at butikkernes gennemsnitlige efterspørgsel på koldskål stiger hen over sommeren til og med august, hvor sæsonen nærmer sig slutningen (Holland 2022)

- Har der været 25°C eller varmere i mere end tre dage, så falder butikkernes gennemsnitlige efterspørgsel på koldskål med −66.74 ltr.

Det kan være en indikation på, at forbrugerne bliver trætte af at spise koldskål, når det bliver for varmt over en længere periode.

- Stiger den gennemsnitlige temperatur målt pr. time med én °C, stiger butikkernes efterspørgsel på koldskål i gennemsnit med 4.25 ltr.

Øget sommervarme hænger moderat sammen med butikkernes efterspørgsel på koldskål. Det stemmer overens med eksisterende viden på området (Kjer 2022).

	Baseline	Simpel	Moderat	Kompleks
Forvent_lager (mellem)			−76.57∗∗∗ {19.82}	−74.55∗∗∗ {19.72}
Forvent_lager (høj)			−82.67∗∗∗ {22.58}	−87.00∗∗∗ {22.81}
Weekend_helligdag (ja)			113.01∗∗∗ {15.00}	107.33∗∗∗ {15.16}
Kamjunk (nej)			−71.89∗∗∗ {17.50}	−76.52∗∗∗ {19.82}
Temp_gt25_3_dage			−82.00∗ {34.23}	−66.74∗ {34.90}
Måned (maj)			84.37∗∗∗ {24.54}	101.69∗∗∗ {23.36}
Måned (juni)			129.51∗∗∗ {32.79}	162.71∗∗∗ {27.87}
Måned (juli)			155.95∗∗∗ {32.79}	155.947∗∗∗ {29.12}
Måned (august)			85.10∗ {34.76}	197.44∗ {30.96}

	Baseline	Simpel	Moderat	Kompleks
Temp_mean_past1h		9.46∗∗∗	4.249∗	0
		{1.89}	{2.07}	
Uafhængige variable	0	1	6	6
Skæring $\widehat{\beta}_0$	520.23∗∗∗	346.04∗∗∗	379.93∗∗∗	434.78∗∗∗
Model P-værdi	∗∗∗	∗∗∗	∗∗∗	∗∗∗
RMSE_træning	139.20	128.74	82.10	83.21
RMSE_test	139.20	130.36	88.55	89.37
R^2		0.15%	0.65%	0.64%
Justeret R^2		0.14%	0.63%	0.62%
Observationer	151	151	151	151

Tabel 1. Resultater fra fire testmodeller. Referencegrupper () for faktorerne er: kamjunkja, forvent_lagerlav, månedapril. {} referer til standardfejlen. Note: $* = P < 0.1; ** = P < 0.05; *** = P < 0.01$

7 Diskussion

I det forrige afsnit blev der fortolket på resultaterne fra modellen med en moderat kompleksitet. Der begrundes for, hvorfor Temp_mean_past1h bliver mindre signifikant i takt med kompleksiteten øges.

Først laves der et histogram med `ggplot()` der viser forskellen i RMSE for alle modellerne på hhv. test og træningsdata ud fra modelkomplesiteten.

Forskellen i trænings- og test RMSE er jvf. histogrammet ikke ens når kompleksiteten øges. I den simple model er test-MSE'en større end trænings-MSE'en. Det samme mønster i den moderate og den komplekse gælder. Det skyldes at vores SL-metode 'tuner' modellen for meget, så den køre modellen for hårdt på træningsdata. På den måde opfanger modellen tilfældigheder fra træningsdata, i stedet for egenskaber ved den f, vi ikke kender til. Mønstret fra træningsdata stemmer dermed ikke overens med mønstret i testdatasættet (Hastie 2022). I histogrammet kan man også se, at forskellen mellem RMSE i den moderate og den komplekse model er meget lille. Alt andet lige, vælges den moderate ud fra et princip om sparsommelighed frem for den komplekse model (ibid).

For det første hænger det optimale modeldevalg også sammen med præcision i forudsigelse og fortolkningen af parametrene. Stiger fleksibiliteten bliver det svære at tolke på parametrene. Falder falder fleksibiliteten bliver det nemmere at tolke på parametrene. Da Temp_mean_past1h blev opløftet til Temp_mean_past1h^22 i den komplekse model, blev p værdien for $\widehat{\beta}_1$ for Temp_mean_past1h^22 insignifikant ved $Pr = 0.58\%$. $\widehat{\beta}_1$ er derfor ikke tilstrækkelig langt fra 0.

Dette gjorde det vanskeligere at tolke på denne paramter. Det var fordi modellen opfangede for mange fejl og for meget støj i form af bias.

For det andet hænger modelvalget sammen det bias/variance tradeoff som forekommer, hvis man øger modelkompleksiteten i jagten på identificere den model som har lavest varians og bias. Varians er ændringen i $\hat{f}$, når den beregnes på nye data, fx. vha valideringsmetoder - dog er værdien aldrig helt den samme! Stiger fleksibiliteten øges variansen. Bias opstår når man forsøger, at forudsige et komplekst fænomen med en for simpel metode. Stiger fleksibiliteten reduceres bias. Efterspørgslen på koldskål er som førnævnt et multidimensionelt fænomen. Vi forsøgte at reducere denne bias ved og udvide den simple lineære model med Temp_mean_past1h uden alle de kategoriske faktorvariabler, da de ikke kan indgå i en polynomisk regression. Det resulterede ikke i en forbedring i RMSE, eller en stigning i R^2. Hverken da den blev opløftet i 3 og 22 potens. Valget blev derfor, at beholde faktorerne i den moderate model, fordi de til sammen forklarer 63% af den samlede varians. Andre variable kunne havde indflydelse på butikkernes efterspørgsel på koldskål. Gennemsnitlige antal solskinstimer, maksimale vindhastighed, samt de private forbrugers socioøkonomiske forhold kunne bidrage med flere dimensioner til den moderate model. Beslutningen blev derfor, at vælge den multible lineære model med moderat kompleksitet. I næste kapitel udrulles anbefalingerne.

8 Anbefalinger

For at kunne designe og udvikle det bedste dataprodukt bruges et Data Product Canvas til, at kortlægge nøgleområder i form af anbefalinger. Thise skal implementere disse for, at øge deres datamodenhed og dermed øge deres økonomiske indtjeningspotentiale. Der præsenteres fem overordnede anbefalinger:

1. Opgrader deres data science stack til en mere moderne version. Start med at opgradere Navision til også at være cloud-baseret, i stedet for udelukkende at bruge det via. Windows styresystemet.
2. Opstil nye regressionsmodeller ud fra alt deres historiske produktions data. Brug den multible lineære regression fra analysen som modelskabelon, og reproducer den i en anden produktionskontekst.
3. Få medarbejderne til at opfatte sig selv som værende en del af en moderne datamodenhedskultur. Første skridt er at ansatte en dataanalytiker som skal accelerer datamodenhedsprocessen fra fase til fase to. I fase to begynder man at strukturere dataopsamlingen for, at lukke risikohuller der er når forskellige systemer skal samarbejde.
4. Få ledelsen til at være spydspidsen når datamodenhedsprocessen skal acceleres fremadrettet.
5. Brug R-Studio som programmeringssprog fordi det er gratis og det arbejder godt sammen med andre programmer som fx. SQL, hvilket Thise allerede bruger som en del af deres data science stack.

9 Konklusion

I dette afsnit besvares problemformuleringen.

10 Litteratur

Bækby, R. & Kølsen, C. (marts. 2017). *„Find din vej i dataindsatsen"*. I: Alexandrainstituttet.

Hastie, T. & James, G. (august. 2021). *„An introduction to statistical learning"*. I: Springer. 2 udgave.

Holland, S. (jul. 2022). *„Vejret afgør sommerens mængde af koldskål"*. I: Fødevareforbundet.

Link: https://www.nnf.dk/nyheder/2021/juli/vejret-afgor-sommerens-maengde-af-koldskal/

Jensen, M. L (jul. 2022). *„Thise skruer gevaldigt op for koldskålsproduktionen"*. I: Tv Midtvest.

Link: https://www.tvmidtvest.dk/skive/thise-mejeri-skruer-gevaldigt-op-for-koldskaalsproduktionen

Kjer, U. (sep. 2022). *„Sommervejret var 3 pct. bedre end sidste år"*. I: Mejeriforeningen.

Link: https://mejeri.dk/nyheder/sommervejret-var-3-pct-bedre-end-sidste-ar/

Osterwalder, A. & Pigneur, Y. (2010). *„Business Model Generation "*. 1. udgave. John Wiley & Sons.

Picopublish (feb. 2022). *„Datamodenhed handler om at blive bedre til at anvende egne data i værdiskabende sammenhænge"*. I: Picopublish.