Forecast projekt i EL:CON

Af: Jacob Ulbæk Pedersen

Studienummer: 201705416

Cand Merc Business Intelligence

Vejleder: Martin Petri Bagger

Antal anslag inkl. blanktegn del 1: 39.999 + 7 grafer (5.600) + 3 tabeller (2.400) = 47.999

Antal anslag inkl. blanktegn del 2: 6.948

Executive summary

Hos EL:CON A/S har man igennem de seneste år opnået høj vækst. Væksten er opnået både gennem organisk vækst og opkøb, som virksomheden har foretaget. Den store vækst, har sat større krav til virksomhedens mulighed for at følge op på dens økonomiske performance. Derfor begyndte man i EL:CON i starten af 2021 at gentænke måden, man laver opfølgning på økonomisk performance. Som en del af denne proces, blev det besluttet at forsøge at udvikle en datadrevet model, som kan lave forecast på EL:CON's afdelinger på dækningsbidragsniveau, hvilket er formålet med dette projekt.

For at bygge denne model, blev der i projektet først undersøgt, hvordan omsætningen have udviklet sig i tidsperioden fra januar 2017 til november 2021. Her viste det at omsætning havde haft en opadgående trend, samtidig med at omsætningen var præget af månedlig sæsoneffekt. Derudover kunne det også ses, at der var store afvigelser, som ikke kunne forklares ved hjælp af sæsoneffekten eller trenden, hvilket fortalte, at der skulle flere variable i spil, for at kunne forecaste virksomhedens omsætning.

Efterfølgende blev EL:CON's forskellige afdelinger undersøgt. Til at gøre dette blev der udarbejdet en SARIMA model for alle afdelingerne, hvorefter der blev evalueret på dets performance i perioden fra januar 2021 til november 2021. Her blev det tydeligt, at mange af EL:CON's afdelinger bevægede sig i forskellige retninger i forecast perioden fra januar 2021 til december 2022, hvorfor det blev besluttet, at der ikke kunne udarbejdes en forecast model for alle virksomhedens afdelinger. Derfor blev der valgt en afdeling, som vil blive brugt som eksempel. Afdelingen der blev valgt, er Service afdelingen i Aarhus.

For at forecaste på dækningsbidraget i EL:CON blev der udvalgt relevante KPI'er der kunne påvirke de delkomponenter dækningsbidraget består af. Herfra blev der udvalgt otte KPI'er som der vil blive forecastet på. Resultatet af den endelige forecast model var en Root Mean Square Error på 529.570 i valideringsperioden, og 243.256 i testperioden. Det blev konkluderet at den performance ikke var tilstrækkelig god nok, til at forecast modellen var god nok til at blive taget endelig i brug.

På trods af at den udviklede forecast model ikke kunne bruges af EL:CON, har der stadig været resultater fra projektet, som er brugbare for virksomheden. Her kan nævnes den sæsoneffekt der blev vist i virksomhedens omsætning, samt konklusionen om at en af hovedårsagerne til at forecast modellen ikke performede særlig godt, var at forecastet på vareforbrug, ikke var særlig godt. Det viste nemlig, at en af de vigtigste økonomiske performance målingspunkter, er svært af forudsige både gennem datadrevet modellering, men også igennem den forretningsforståelse der eksisterer i virksomheden.

Indholdsfortegnelse

Del 1:

1. Introduktion

EL:CON A/S er en dansk elinstallatør virksomhed, som har hovedkontor i Århus, men har afdelinger placeret rundt i det meste af Danmark. EL:CON har gennem de seneste år oplevet stor vækst. I starten af 2017 var der 669 medarbejdere ansat, mens der nu i slutningen af oktober 2021 er 1.122 medarbejdere ansat. Denne forøgelse er kommet både organisk og gennem opkøb, og forventes at fortsætte i den nærmeste årrække.

Grundet den store vækst EL:CON har oplevet og forventer at opleve i fremtiden, er behovet for at kunne følge op på virksomhedens økonomiske performance også steget. Historisk har man i EL:CON haft et setup, hvor man omkring Q3 året inden har udarbejdet månedsbudgetter for det kommende år, for alle virksomhedens afdelinger. Herefter har man fulgt op på afdelingernes performance efter hver månedsafslutning, ved at sammenholde realiserede tal med budgetterede tal.

I starten af 2021 begyndte man i EL:CON at gentænke denne proces. I stedet for 'kun' at følge op på performance mod et budget, ville man også gerne tilføje en proces, hvor man diskuterer, hvor hver enkelt afdeling er på vej hen. Det blev besluttet at man vil gøre dette, ved at lave rullende 24 måneders forecast for hver afdeling hver måned. Derudover vil man udarbejde targets for hver afdeling, som fastsætter hvordan ledelsen forventer afdelingerne performer frem til og med 2025. Formålet med disse to værktøjer er at skabe en ny type dialog i EL:CON, hvor man ikke kun evaluerer tilbage i tid, men også forsøger at skabe en dialog om, hvordan fremtiden ser ud i gennem rullende forecast, og sammenholde dette med de targets der er fastsat for afdelingerne. Altså vil man kunne følge op på, om en afdeling bevæger sig i den retning som man forventer, og hvis ikke, kan man skabe en dialog om, hvordan man kan få dem tilbage på rette kurs.

I den nuværende proces udarbejdes både budget, rullende forecast og targets ud fra historisk data kombineret med forventninger til den fremtidige udvikling fra EL:CON's chefer og direktører. Dog har man i EL:CON's ledelse haft et ønske om at undersøge, om det er muligt at lave et forecast, som er datadrevet og på den måde kan give et anderledes perspektiv på virksomheden ift. det der er på nuværende tidspunkt. Dette er formålet med denne opgave, og giver i den forbindelse anledning til følgende problemstilling:

Udvikling af datadrevet machine learning model, som kan lave afdelings specifikke forecast på dækningsbidrags niveau i virksomheden EL:CON.

2. Teori

2.1 CRISP-DM

Cross Industry Standard Process for Data Mining (CRIPS-DM) er en model, som beskriver seks forskellige faser, som et data science projekt gennemgår (Datascience-pm). Denne model er blevet anvendt i dette projekt, og der vil nu være en kort redegørelse for, hvad hver fase er, og hvordan man kan se det i dette projekt.

2.1.1 Forretning forståelse

I den første fase af projektet, handler det om at opbygge forståelse for den forretning man laver sit data science projekt for. Derudover er det også vigtigt, at man i dette stadie fastsætter de mål, der er med projektet, og hvilke succeskriterier der er. Det første afsnit i denne opgave omhandler netop denne fase, hvor det blev fastlagt, at målet med denne opgave er at undersøge, hvorvidt det er muligt at bygge en datadrevet model, som kan bruges til at forecaste. Derudover blev der også redegjort for, hvad baggrunden er for at EL:CON har et ønske om at udarbejde projektet.

2.1.2 Data forståelse

Den næste fase handler om at få en forståelse for data i den virksomhed man udarbejder sit projekt for, og især for det data, som kan bruges til at opnå målet med projektet. Denne fase kommer til udtryk i afsnit tre i denne rapport, hvor den generelle datastruktur i EL:CON bliver beskrevet og en evaluering af data kvaliteten bliver udført.

2.1.3 Data forberedelse

Data forberedelsesfasen handler om at vælge hvilken data der skal bruges, samle den data, og til sidst forberede det så det er klar til at blive brugt. Denne fase bliver brugt i det sidste data afsnit, der omhandler dataindsamling, samt de første to afsnit af analyseafsnittet, som viser hvordan data er blevet valgt, i dette projekt.

2.1.4 Modellering

Modellering handler om at udvikle og teste forskellige modeller, der kan bruges til at opnå projektets mål. Herefter vurderes det, hvilken model der har det bet bedste fit. Dette er en iterativ proces hvor man hele tiden ændrer i modellen, indtil man har modelleret en model, som opfylder ens mål bedst muligt. I dette projekt udføres modelleringsfasen i det sidste afsnit af analyseafsnittet, hvor der udvikles en række forskellige modeller.

2.1.5 Evaluering

Den femte fase i et data science projekt er evalueringsfasen, hvor man evaluerer den model, der er blevet udviklet i den forrige fase. Det handler om at evaluere om modellen, opfylder de mål, man havde opstillet for projektet. Denne fase vises i resultatsafsnittet af denne rapport.

2.1.6 Implementering

Hvis det bliver besluttet, at den udarbejdede model skal implementeres, handler denne fase om at planlægge hvordan det skal foregå. Derudover handler det også om at lave en præsentation af de ting, man har fundet ud af i løbet af projektet, som kan kommunikeres til virksomheden. Derudover er det også en analyse af, hvad der er gået godt og dårligt i projektet. I dette projekt bliver dette vist i diskussions- og konklusionsafsnittene (The-modeling-Agency, 2000).

3. Data

Dette afsnit vil give en redegørelse for, hvordan data er indsamlet i denne opgave. Først vil der være en kort forklaring af, hvilket data setup EL:CON har, herefter en redegørelse for den tidsperiode der er valgt at indsamle data for, og til sidst en diskussion af kvaliteten af den indsamlede data.

3.1 Dataindsamling

I EL:CON's data warehouse er der primært kun tabeller som er loadet ind uden transformation, og dermed er en 1:1 replikation af de rå tabeller, som kommer fra virksomhedens ERP system. Det betyder også at man i EL:CON altid udregner KPI'er i dataindtrækket til modeller eller som udregnede kolonner eller measures i Power Pivot i Excel eller Power BI. Derfor er det også besluttet at gøre det på den måde, når relevante KPI'er skal indsamles i denne opgave, hvor det er Power Pivot i Excel, der vil blive anvendt. Det er ikke en optimal løsning, men eftersom målet med denne opgave er at bygge en forecast model, er det uden for opgavens formål at skulle ændre i virksomhedens data setup. Herefter vil de indsamlede KPI'er blive hentet fra Excel og over i Jupyter Notebook, hvor selve forecast modellerne og øvrige analyser vil blive udarbejdet.

Den tidsperiode det er muligt at indsamle data for i EL:CON er januar 2017 til november 2021. Grunden til denne afgrænsning er at man i EL:CON implementerede et nyt ERP system i løbet af 2016 i virksomhedens afdelinger. Derfor er data ikke tilgængeligt fra før det nye ERP system blev implementeret, hvilket gør at der ikke er data for alle afdelinger i alle måneder i 2016. Derfor, for at være sikker på at der er det nødvendige data tilgængeligt for alle afdelinger, blev det besluttet at bruge data fra januar 2017. Grunden til at perioden slutter i november 2021, er at det på det tidspunkt opgaven skrives, er den sidste måned der er blevet afsluttet.

3.2 Datakvalitet

Kvaliteten af data vurderes at være høj. De KPI'er der vil blive brugt i denne opgave er KPI'er som virksomheden selv bruger i deres månedsrapporter, og det er sikret, at de er udregnet på samme måde, som virksomheden selv gør det. Der er dog stadig et par udfordringer, som er værd at nævne.

Det første omhandler virksomhedens tidsregistrering. Produktionsmedarbejderne i EL:CON registrer deres tid, og de fleste af disse er timelønnede på månedsbasis. Derfor har de tidsgrænser for, hvornår de senest skal registrere timer, for at få løn for disse. Dog sker det en gang i mellem, at medarbejdere glemmer at få registreret deres timer, og først bliver opmærksom på det, når de får deres løn. Hvis det sker, registrerer de deres timer med dato i den rigtige måned, men timerne påvirker først lønforbruget i den efterfølgende måned. Derfor kan der være tilfælde, hvor det antal timer der betales løn for, ikke er det samme antal

timer, der er registreret i timetabellen, som bruges til at udregne KPI'erne faktureringsgrad, jobtimer og fakturerbar tid.

Den næste udfordring omhandler antallet af produktionsmedarbejdere. EL:CON rapporterer hver måned, hvor mange produktionsmedarbejdere der er i virksomhedens forskellige afdelinger. Dog viser tallet, hvor mange medarbejdere der er i slutningen af hver måned, og ikke hvor mange der er i gennemsnit i hver måned. Derfor blev det besluttet at udregne antal produktionsmedarbejdere på følgende måde:

$$Antal\ produktionsmedarbejdere\ (t) = \frac{Antal\ produktionsmedarbejdere\ ultimo + Antal\ produktionsmedarbejdere\ primo}{2}$$

Derudover er der en anden udfordring vedrørende antal produktionsmedarbejdere. Hvis en medarbejder hjælper til, på sager der hører til i andre afdelinger, vil de timer medarbejderen registrere derfor også ende i en anden afdeling, end den afdeling medarbejderen tælles med i, når der skal tælles antal produktionsmedarbejdere. Det samme gælder lønomkostningen for de timer medarbejderen arbejder for en anden afdeling. Derfor kan der være tilfælde, hvor antallet af timer registreret og lønforbruget ikke hænger sammen med det antal produktionsmedarbejdere der er registreret i afdelingen.

Slutteligt er der en potentiel udfordring ift. vareforbrug. EL:CON har ikke løbende lagerstyring, hvorfor træk fra virksomhedens lager ikke registreres løbende i resultatopgørelsen, men først rammer når der laves lageroptælling ifm. årsregnskabet. På samme måde omkostningsføres køb til varelageret løbende i resultatopgørelsen og reguleres ligeledes først korrekt ifm. den årlige lageroptælling. Dette kan være med til at henholdsvis under- eller overvurdere det reelle vareforbrug pr. måned. Da der som oftest købes varer ind som skal forbruges i samme måned vurderes denne problemstilling dog ikke at have en markant effekt.

4. Metode

Formålet med dette afsnit er at beskrive de modeller og værktøjer, der er blevet brugt i dette projekt. Afsnittet er opdelt i seks underafsnit, som hver giver en kort beskrivelse af et værktøj eller model, der er anvendt i løbet af opgaven, samt en argumentation for, hvorfor det er blevet anvendt.

4.1 Tidsserie dekomposition

Dekomposition af en tidsserie er et værktøj, som opdeler en tidsserie i tre kategorier: trend, sæson og residual. Trend fortæller hvilken retning ens tidsserie overordnet set bevæger sig i. Sæson fortæller om hver måned, hvis det er månedsdata man arbejder med, i gennemsnit er højere eller lavere end den udregnede trend. Den sidste kategori residual, er den del, som man ikke kan forklare gennem en overordnet trend eller sæson effekt. Opdelingen i de tre kategorier sker enten som en additiv dekomposition, hvor de tre kategorier summeret giver det realiserede tal, eller multiplikativv dekomposition, hvor de tre kategorier multipliceret giver det realiserede tal (Otexts).

I dette projekt bruges tidsserie dekomposition, til at opnå en bedre forståelse for EL:CON som helhed men også for bedre at kunne forstå de enkelte afdelinger. Derudover bruges værktøjet også til at undersøge sæsoneffekten, der er i de forskellige KPI'er, når der skal laves transformation af disse, for at sikre at tidsserierne er stationary. Grunden til at værktøjet er valgt, er at det giver et simpelt og overskueligt overblik, samtidig med at det er nemt at kommunikere resultatet videre til andre, som ikke på forhånd har en forståelse for statistiske modeller. En af ulemperne ved tidsserie dekomposition er at det er en naiv metode til at lave et forecast, og derfor er værktøjet heller ikke brugt til det formål (Statsmodels).

4.2 SARIMA

SARIMA er et forecast værktøj, som udarbejder et forecast ved hjælp af mønstre i historisk data på den variabel man ønsker at forecaste. De mønstre værktøjet kigger efter er autokorrelation, partial autokorrelation, stationarity og sæsoneffekt. Autokorrelation er korrelationen mellem en tidsserie og en lagged version af den samme tidsserie, som bruges til at undersøge, om eksempelvis sidste periodes realiserede tal påvirker denne måneds realiserede tal. Partial autokorrelation er korrelationen mellem en tidsserie og en lagged version af den samme tidsserie, men hvor der er fjernet effekten fra de laggede tidsperioder der ligger i mellem. Stationary tidsserie data forekommer hvis data er uafhængig af hvilket tidspunkt i tidsserien man er på, hvilket betyder at stationary tidsserier ikke indeholder sæsoneffekt eller en overordnet trend. Sæsoneffekt undersøger om der er autokorrelation, partial autokorrelation og stationarity i datapunkter fra eksempelvis den samme måned i tidsserien, hvis det er månedsdata man arbejder med (Towardsdatascience, 2021 1).

SARIMA er valgt fremfor ARIMA, fordi at tidsserie dekompositionen udført i opgaven viste, at der er sæsoneffekt i EL:CON's data, og forskellen mellem SARIMA og ARIMA er netop at SARIMA kan tage højde for sæsoneffekt. Grunden til, at der ikke er brugt en SARIMAX model er at formålet med at bruge værktøjet er, at opbygge en bedre forståelse for EL:CON's forskellige afdelinger. Derfor er den mere simple SARIMA model valgt fremfor SARIMAX modellen, som ellers vil kunne inddrage mønstre i historisk data fra andre variable end kun den enkelte variabel, man forsøger at lave forecast på. Når SARIMA modellerne udarbejdes i dette projekt, vælges de specifikke parametre automatisk ved hjælp af pmdarima.auto_arima funktionen i Jupyter Notebook (Pypi).

4.3 Data transformation

Når et forecast udarbejdes, er det bedst at forecaste på en stationary tidsserie, og hvis den er nonstationary transformes ens variabel ved hjælp af forskellige transformationer indtil den er stationary. I denne opgave er det valg at teste stationarity med en Augmented Dickey-Fuller test. Augmented Dickey-Fuller test er en test af stationarity, hvor nulhypotesen er at ens tidsserie er nonstationary, mens den alternative hypotese er at den er stationary. Hertil er der brugt et signifikansniveau på 5%, når testen udføres, hvor nulhypotesen afvises og det dermed påvises at ens tidsserie er stationary, hvis sandsynligheden fra Augmented Dickey-Fuller test er mindre end 5% (Machinelearningmastery, 2016).

I denne opgave er der anvendt forskellige transformationer for at opnå stationarity i de forskellige KPI'er. De fleste KPI'er indeholder en trend, samt sæsoneffekt og derfor er der i flere tilfælde lavet flere transformationer for at opnå stationarity. De transformationer der er anvendt er: fjerne sæsoneffekt udregnet ved hjælp af tidsserie dekomposition, differencing og til sidst ved at tage den naturlige logaritme.

Efter det er sikret at KPI'erne er stationary, er de alle blevet scaled, for at sikre at de alle er på den samme skala. Scalings værktøjet der er anvendt, er min maks scaler, hvilket betyder at alle observationer i tidsserien er mellem nul og en, hvor nul er den mindste observation og en er den højeste observation. Her er det vigtigt at nævne, at scaling er lavet ud fra træningsdatasættet, så der kan være observationer i validering og test datasættet der er mindre end nul eller større end 1, hvis de er større eller mindre end de observationer der er i træningsdatasættet (Towardsdatascience, 2020).

4.4 Valg af KPI'er

Når der skal vælges hvilke KPI'er der skal bruges til forecast modellerne, vil der blive brugt en kombination af korrelationsanalyse mellem den afhængige og de uafhængige variable, og output fra en random forest model der viser hvor vigtige de forskellige KPI'er er for modellen. Derudover, vil der også blive brugt forretningsforståelse før det konkluderes om en KPI skal anvendes. Forretningsforståelse handler om at

sikre at der ikke kun er korrelation mellem variable, men også at der er en logisk forretnings sammenhæng mellem dem. De variable der bliver undersøgt, er både KPI'er fra samme tidsperiode samt laggede versioner af de samme KPI'er, for at se om sidste periodes observationer kan bruges i forecast modellerne.

4.5 Valg af forecast modeller

Til udarbejdelse af forecast modellerne vil regressions funktionen fra PyCaret biblioteket i Python blive brugt. Denne funktion søger automatisk i mellem 25 forskellige modeller og vælger den model der fitter bedst baseret på et performance measure. I denne opgave, når der skal udarbejdes forecast modeller, er det valgt at bruge alle modeller udtagen k-nearest neighbors (KNN). Grunden til dette, er at KNN gav udfordringer når modellerne kørte. Efter alle modellerne er trænet og tunet på trænings datasættet bliver den model der bedst fitter validerings datasættet valgt. En liste, med en kort beskrivelse over de modeller som funktionen bruger kan findes i bilag 1.

Grunden til det er valgt at bruge PyCaret er at den giver mulighed for at teste mange forskellige modeller på en gang, og på den måde kommer bredt omkring alle de forskellige forecast modeller der findes, for at finde den der giver det bedste fit. Derudover kan man automatisk tune modellernes hyper parametre ved hjælp af PyCaret. Eftersom der skal udvikles flere forecast modeller i denne opgave, blev det derfor set som et fordelagtigt valg, så der ikke blev brugt for meget tid på at træne og tune alle forecast modellerne manuelt (PyCaret, 2020).

4.6 Måling af performance

Som nævnt ovenfor, vælges den model fra PyCaret der fitter bedst på validerings datasættet ved hjælp af et performance measure. I denne opgave er det valgt at bruge Root Mean Square Error (RMSE) til at måle performance. RMSE måler, hvor meget ens predictions afviger fra de faktiske værdier. Det gøres ved at udregne gennemsnittet af summen af de kvadrerede afvigelser mellem de predictions en model laver, og de faktiske værdier. Herefter tages kvadratroden af det tal, for at finde RMSE. RMSE er valgt, fordi den straffer store afvigelser, da det ikke ønskes at have en forecast model, der har store afvigelser i nogle måneder (Thedatascientist). RMSE vil også blive brugt som performance measure, når performance på validering og test datasættet skal evalueres. Alternativer som kunne være blevet anvendt, er Mean Absolute Error (MAE) og Mean Square Error (MSE).

5. Analyse

Dette afsnit er en redegørelse for den analyse der er foretaget i projektet. Det er opdelt i tre underafsnit, hvor det første omhandler en general analyse af EL:CON som virksomhed. De efterfølgende to afsnit omhandler modelopsætning, træning og valg af modeller og til sidst udarbejdelsen af et forecast, for en udvalgt afdeling.

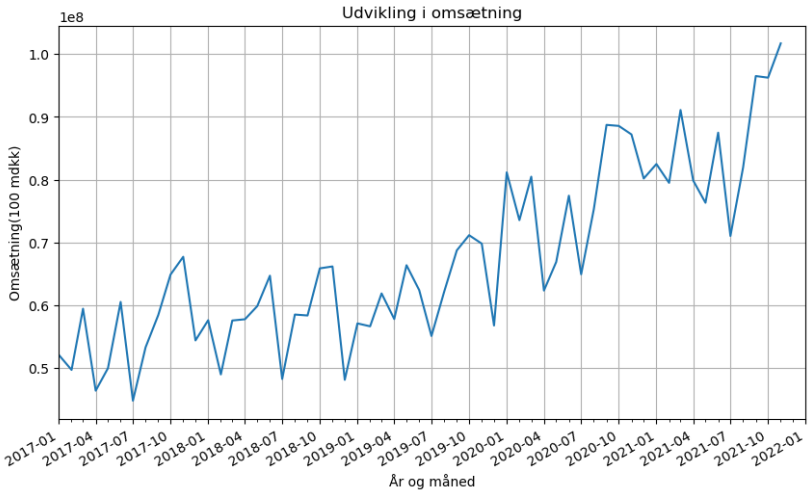
5.1 Opbygning af forståelse for EL:CON

For at kunne lave et forecast i en virksomhed, er det vigtigt at have en forståelse for virksomheden, og især det man gerne vil forecaste. Målet med dette projekt er at forecaste dækningsbidraget på afdelingsniveau i EL:CON. For at kunne gøre dette, er det blevet valgt at fokusere på de elementer, som påvirker dækningsbidraget, og forecaste disse. Ligningen for hvordan dækningsbidrag udregnes er følgende:

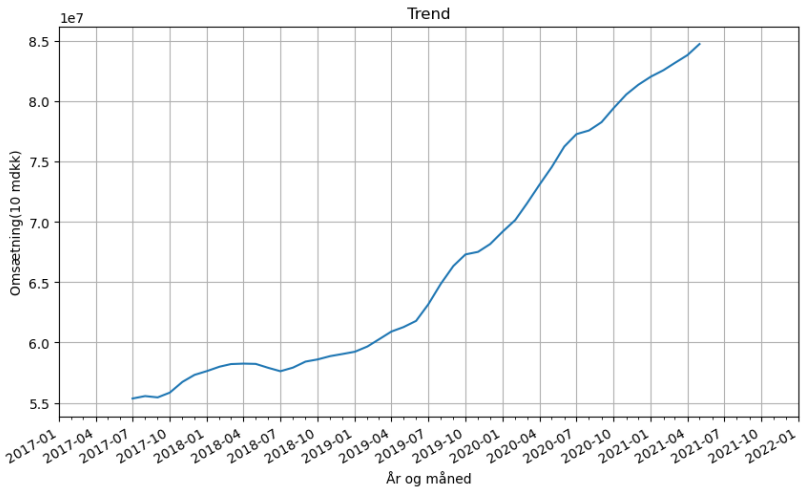
$$Dækningsbidrag = Omsætning - Vareforbrug - Lønforbrug$$

Derfor er det omsætning, vareforbrug og lønforbrug, der hver især vil blive forecastet, og herefter vil det forecastede dækningsbidrag blive udregnet ud fra ovenstående ligning. Grunden til dette er valgt, er at man i EL:CON også ønskede at have et forecast på især omsætning, da det er noget man i virksomheden gerne vil have fokus på.

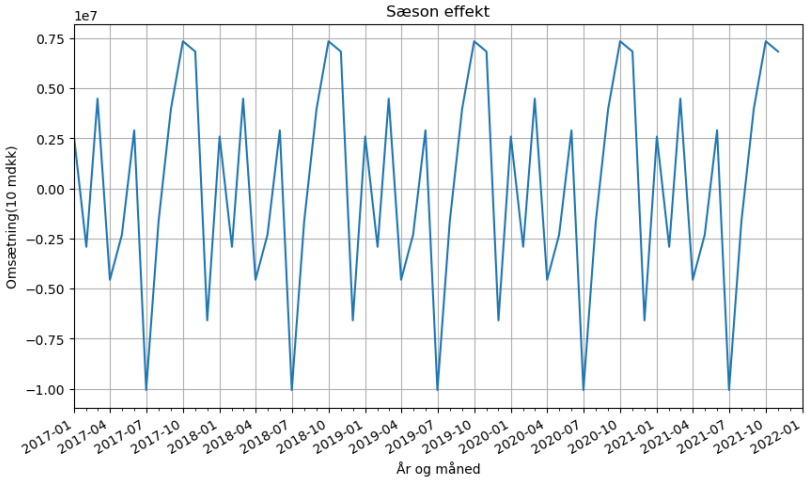
For at opbygge en forståelse af, hvad EL:CON er for en virksomhed vil der først blive lavet en dekomposition af omsætningen total for hele virksomheden. Det er valgt, fordi det giver et godt indblik i hvor høj grad sæsoneffekt påvirker virksomhedens aktivitet, samt giver en forståelse for den overordnede trend, virksomheden har bevæget sig i historisk set. En additiv dekomposition giver følgende output, og koden kan ses i bilag 2:



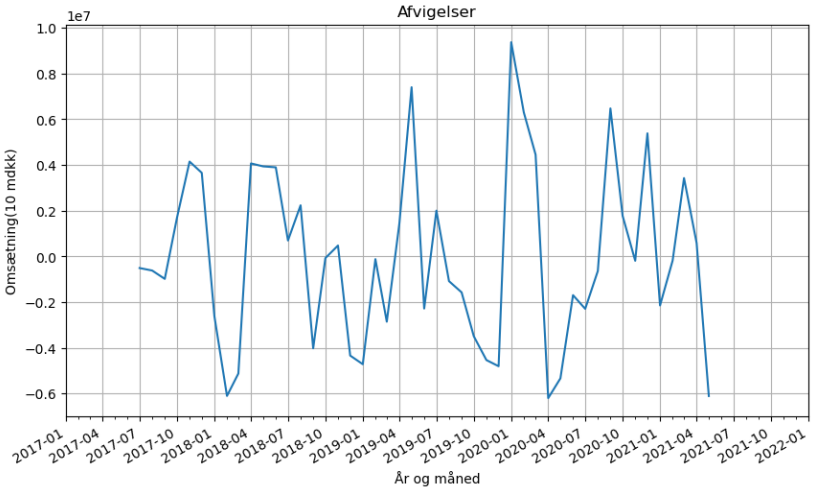
Graf 5.1: Udvikling i omsætning i 100 mdkk i perioden januar 2017 til november 2022.



Graf 5.2: Trend effekt output fra additiv dekomposition af omsætning i periode fra januar 2017 til november 2021.



Graf 5.3: Sæson effekt output fra additiv dekomposition af omsætning i periode fra januar 2017 til november 2021.



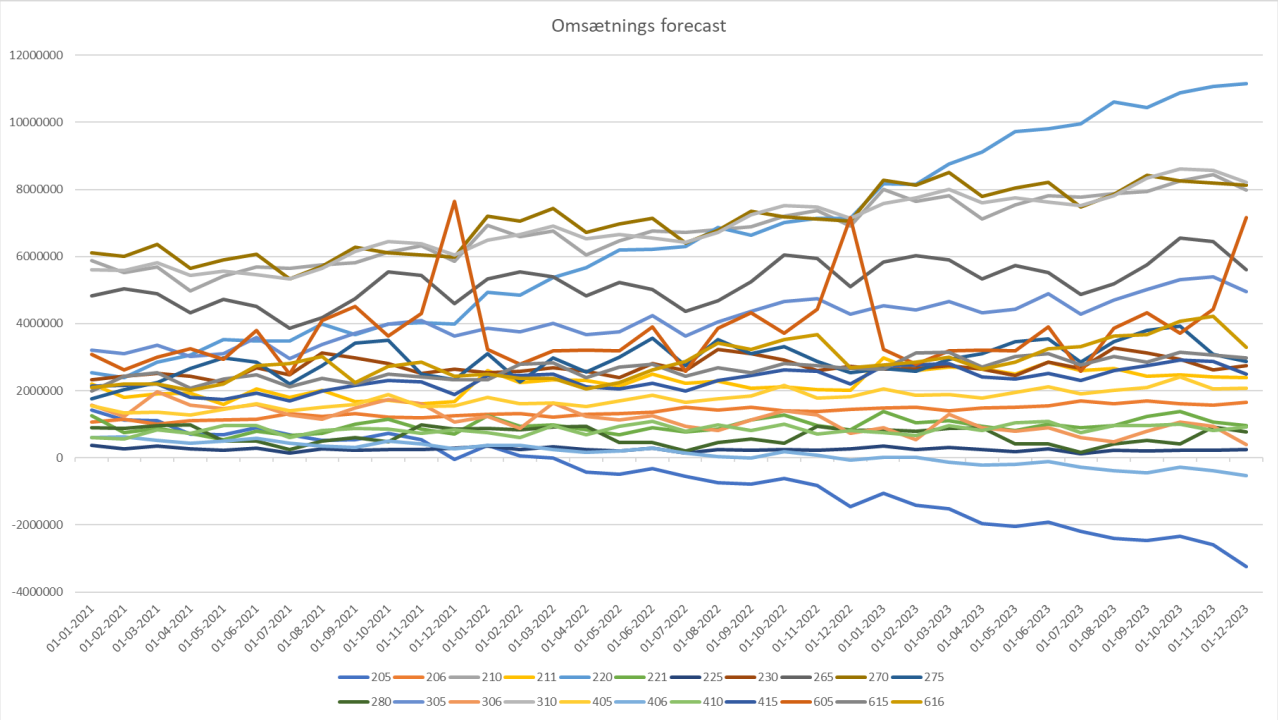
Graf 5.4: Residual output fra additiv dekomposition af omsætning i periode fra januar 2017 til november 2021.

Graf 5.1 viser udviklingen i den realiserede omsætning for EL:CON fra januar 2017 til november 2021. Her ses det at omsætningen har været stigende i perioden, hvor den i januar 2017 var cirka 50 millioner kroner til at den i november 2021 er cirka 100 millioner kroner. Derudover viser den også at omsætningen svinger op og ned i høj grad hver måned. Begge disse konklusioner understøttes af graf 5.2, der viser, at den overordnede trend er opadgående, og graf 5.3 der viser sæsoneffekten, hvor især juli og december kan fremhæves som måneder med lavere aktivitet, mens oktober og november er måneder med højere aktivitet.

Graf 5.4 viser den omsætning der ikke kan forklares gennem trend og sæsoneffekt. Her ses det, at der i en måned er helt op til næsten en million i afvigelse, hvilket er meget taget i betragtning at omsætningen i den måned var cirka 80 millioner. Det fortæller os, at der er flere effekter der skal tages højde for, udover sæsoneffekt og den overordnede trend, hvis der skal laves et forecast på omsætningen i EL:CON.

EL:CON har på nuværende tidspunkt 35 afdelinger fordelt på lokationer i hele landet. Disse afdelinger er inddelt i fire forskellige divisioner, som er: Service vest, Service øst, Entreprise og Specialkompetencer. I opstartet af projektet blev det besluttet, at der ikke skulle forecastes på afdelingerne i Entreprise divisionen, da historiske tal ikke på samme måde kan bruges til at forecaste deres aktivitet, da det i højere grad handler om deres ordrebeholdning på nuværende tidspunkt. Derudover er de afdelinger som EL:CON har opkøbt i perioden fra 2017 og frem, heller ikke blevet inddraget, da de ikke har den nødvendige datamængde.

Efter der er blevet frasorteret afdelinger ud fra ovenstående kriterier, er der 22 tilbage. For at få en bedre forståelse, for disse afdelinger, er der udarbejdet en SARIMA model for alle afdelinger. Modellernes komponenter er fundet ved hjælp af automatisk søgning af hvad, der bedst passer på afdelingens tal i en træningsperiode fra januar 2017 til december 2020. Modellen forecaster herefter 24 måneder ud i fremtiden fra januar 2021 til december 2022, hvor den bliver målt på, hvor godt den forecaster perioden fra januar 2021 til november 2021, hvor der er realiserede tal. Resultat af dette, kan ses i følgende graf og tabel:



Graf 5.5: 24 måneders omsætnings forecast output fra SARIMA modeller for 22 afdelinger i EL:CON.

Afdeling	205	206	210	211	220	221	225	230	265	270	275
ModelDescription	ARIMA(1,2,3)(0,1,1)[12]	ARIMA(1,2,1)(0,1,1)[12]	ARIMA(0,1,1)(0,1,1)[12]	ARIMA(0,1,1)(0,1,2)[12]	ARIMA(0,2,2)(0,1,1)[12]	ARIMA(2,1,3)(0,1,1)[12]	ARIMA(0,1,1)(0,1,1)[12]	ARIMA(1,2,1)(0,1,1)[12]	ARIMA(0,1,1)(0,1,1)[12]	ARIMA(1,1,0)(0,1,1)[12]	ARIMA(0,1,2)(1,1,1)[12]
RMSE	1.090.249	289.525	914.526	420.685	1.757.566	223.724	69.434	508.181	712.416	2.261.640	1.037.315

Afdeling	280	305	306	310	405	406	410	415	605	615	616
ModelDescription	ARIMA(0,1,1)(0,1,1)[12]	ARIMA(0,1,1)(0,1,1)[12]	ARIMA(0,0,0)(0,1,0)[12] intercept	ARIMA(0,1,2)(0,1,1)[12]	ARIMA(0,1,1)(0,1,1)[12]	ARIMA(0,1,1)(0,1,1)[12] intercept	ARIMA(0,1,0)(1,1,0)[12]	ARIMA(0,1,0)(0,1,1)[12]	ARIMA(0,0,0)(0,1,2)[12]	ARIMA(3,1,0)(0,1,1)[12]	ARIMA(0,1,0)(2,1,0)[12]
RMSE	246.234	356.159	306.297	450.208	407.354	286.698	275.199	489.335	763.105	488.988	1.735.023

Tabel 5.1: 24 måneders omsætnings forecast performance og type af SARIMA model valgt for 22 afdelinger i EL:CON.

Outputtet af SARIMA modellerne viser tydeligt, at der er forskel på afdelingerne i EL:CON. Kigger man på graf 5.5 ses det at nogle få afdelinger ender med negativ omsætning i forecast perioden, mens en anden afdeling firedobler sin omsætning indenfor forecast perioden. Derudover ses det også, at nogle afdelinger nærmest er en flad linje, hvor det ikke virker som om, der er nogen sæsoneffekt eller trend, hvorimod andre afdelinger svinger meget fra måned til måned. Kigger man på tabel 5.1, ses det også, at det er meget forskelligt, hvilke SARIMA modeller, der bedst fitter afdelingernes omsætningstal i træningsperioden.

Derudover ses det også at der er stor forskel på hvor god en RMSE der er på de forecastede tal, hvor det bedst fit har en RMSE på 69.434 og det dårligste på 2.261.640. Baseret på ovenstående analyse, konkluderes det, at det ikke vil være muligt at lave en forecast model, der passer på alle EL:CON's

afdelinger på en gang. Derfor vil der nu blive udvalgt en afdeling, som der vil blive udarbejdet en forecast model for. Koden for hvordan modellerne er lavet, kan ses i bilag 2.

I samarbejde med interessenterne i EL:CON blev det besluttet at den afdeling der først skulle udarbejdes et forecast for er Service afdelingen i Aarhus. Det blev valgt, fordi det er en af de større afdelinger i EL:CON, og derfor er interessant at undersøge. Samtidig blev det vurderet, at det vil give mening at vælge en afdeling, hvor afdelingens chef var lokaliseret samme sted som der hvor praktikprojektet bliver udarbejdet. På den måde vil spørgsmål hurtigere kunne blive besvaret, når de skulle komme undervejs.

5.2 Modelopsætning

Som beskrevet tidligere, skal der laves et forecast både på omsætning, vareforbrug og lønforbrug som så efterfølgende lægges sammen som et forecast for dækningsbidraget. For at opbygge modeller der kan forecaste disse, skal der udvælges hvilke parametre, der skal inddrages i modellen. For at gøre det, er der i samarbejde med EL:CON udvalgt KPI'er som potentielt kunne have en effekt på en eller flere af de tre regnskabstal der skal forecastes. En oversigt med forklaring af disse kan ses i bilag 3. Før analysen bliver igangsat, vil der blive udført en test for stationarity af alle de KPI'er der er udvalgt, hvor data vil blive transformeret for at sikre at der arbejdes med stationary data. Herefter vil alle variable blive scaled. En oversigt over de transformationer der er udført på KPI'erne, kan ses i bilag 4.

Derudover vil datasættet også blive opdelt i tre dele: træning, validering og test. Træningsdatasættet består af alle observationer fra januar 2017 til december 2020, valideringsdatasættet består af observationerne fra januar 2021 til august 2021 og test datasættet består af observationerne fra september 2021 til november 2021.

For at vælge hvilke KPI'er der skal bruges i forecast modellerne, er der både blev kigget på korrelationsanalyse, samtidig med at der er udarbejdet en random forest model, hvor der undersøges, hvor vigtig hver parameter er for modellen. Både korrelationsanalysen og random forest model analyserne er lavet ud fra træningsdatasættet. Det er blevet gjort for de KPI'er der skal bruges for at kunne forecaste dækningsbidraget, og resultatet af denne analyse kan ses i bilag 5.

Som resultat af ovenstående analyse, er det besluttet at der skal udarbejdes en forecast model for følgende parametre: Arbejdsdage – ferie, Faktureringsgrad, Antal produktionsmedarbejdere, Fakturerbar tid, Jobtimer, Omsætning, Vareforbrug, Lønforbrug og Dækningsbidrag. Her vil arbejdsdage – ferie, jobtimer og dækningsbidrag dog blive udregnet ud fra de andre KPI'er eller historisk data, og der vil derfor ikke blive lavet modeller for disse. Det betyder altså at der skal udarbejdes seks modeller, hvilket til sidst vil resultere

i et forecast for dækningsbidraget i afdelingen. En oversigt over korrelationsanalyse og random forest modellerne for de forskellige KPI'er kan ses i bilag 2.

5.3 Valg og træning af forecast modeller

Der er nu udvalgt, hvilke KPI'er der skal indgå i de forskellige forecast modeller og derfor er det nu tid til at træne og vælge hvilket forecast model værktøj, der skal bruges. Det gøres ved hjælp af PyCaret, som bruger træningsdatasættet til at træne modellerne, som er indeholdt i biblioteket, hvorefter valideringsdatasættet bruges til at vælge den model der passer bedst. Her er det vigtigt at nævne, at alle modellerne er blevet tuned før modellen bliver valgt, for at sikre at det er den rigtige model der bliver valgt. En samlet oversigt over de seks modeller der er blevet valgt kan ses i bilag 6, mens koden kan ses i bilag 2.

Det sidse der vil blive udført i analyse afsnittet, er at der udarbejdes et forecast. Som nævnt tidligere er målet med denne opgave at udarbejde at 24 måneders forecast, og derfor er det også det der vil blive gjort her. Tidsperioden er fra januar 2021 til december 2022. Eftersom KPI'erne der bruges i de forskellige forecast modeller er afhængig af hinanden, vil forecastet blive udarbejdet i et loop, hvor der tages højde for de indbyrdes afhængigheder. Et eksempel på dette er at omsætning bruger vareforbrug som variabel i dens forecast model. Derfor er det nødvendigt at lave et forecast på vareforbrug, før der laves et forecast på omsætning. Eftersom der er nogle modeller, der også bruger laggede variable, er det også kun muligt at lave en periodes forecast ad gangen. I den første periode, vil den laggede observation, være det realiserede tal, for december 2020, mens det efterfølgende vil være det forecastede tal fra perioden inden. Rækkefølgen for, hvilke modeller der forecastes på i alle perioder er: Antal produktionsmedarbejdere, fakturerbar tid, faktureringsgrad, lønforbrug, vareforbrug og til sidst omsætning.

Som nævnt tidligere er der tre KPI'er der ikke er udarbejdet forecast modeller for. De er dækningsbidrag, Arbejdsdage – ferie og jobtimer. Dækningsbidraget vil blive udregnet ud fra ligningen vist tidligere. Arbejdsdage – ferie, bliver forecastet, ved at undersøge hvor mange arbejdsdage der er i hver måned i perioden. Herefter vil det blive udregnet, hvor stor en andel af et års ferie, der i gennemsnit er blevet afholdt i hver måned. Dette tal vil så blive ganget med 30, eftersom der er 30 feriedage på et år. Herefter fratrækkes det fra det antal arbejdsdage der er i måneden, hvilket giver, hvor mange Arbejdsdage – ferie der er i hver måned. Jobtimer bliver udregnet ud fra følgende formel:

$$Jobtimer = \frac{Fakturerbar\ tid}{Faktureringsgrad}$$

Grunden til det er valgt at bruge ovenstående formel, til at forecaste jobtimer, er at sikre at de matematiske sammenhænge der er mellem KPI'erne fastholdes i forecastet. Den fulde kode fra forecastet kan ses i bilag 2.

6. Resultat

Der vil i dette afsnit blive analyseret på det forecast der er blevet lavet, for de forskellige KPI'er. Der vil primært være fokus på dækningsbidrag, eftersom det er et forecast på dækningsbidrag, der er hovedformålet med denne opgave. Der vil både blive evalueret, hvad performance er på validering og på træningsdatasættet.

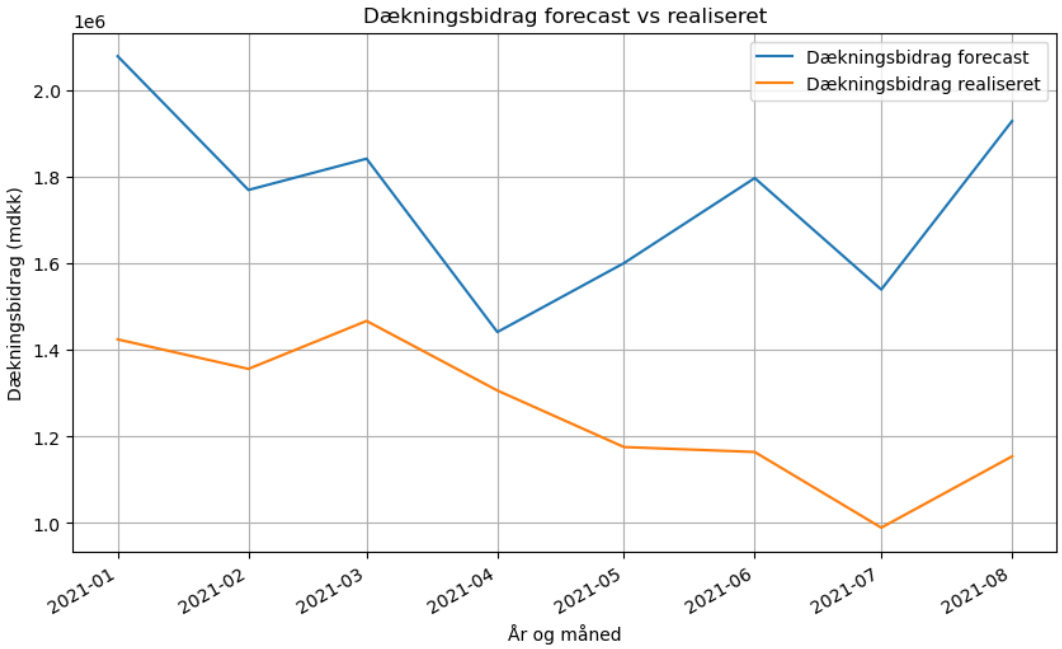
6.1 Valideringsperformance

Nedenstående tabel viser RMSE for de otte forskellige KPI'er der er blevet forecastet på. Derudover bliver det vist, hvad gennemsnittet er af de realiserede tal, samt de forecastede tal, for at kunne evaluere størrelsen på RMSE. Derudover giver det også en indikation på, om modellerne over- eller undervurderer værdierne i dets forecast.

Validation performance			
KPI	RMSE	Gennemsnit realiseret	Gennemsnit forecast
Dækningsbidrag	529.570	1.254.598	1.749.835
Omsætning	581.748	4.739.817	5.069.145
Vareforbrug	344.589	1.372.559	1.161.601
Lønforbrug	146.627	2.112.660	2.157.708
Faktureringsgrad	0,027	0,919	0,901
Jobtimer	1.271	7.558	8.710
Fakturerbar tid	992	6.948	7.848
Antal produktionsmedarbejdere	1,29	65,63	65,73

Tabel 6.1: Validation performance af forecast modeller, samt gennemsnitlig realiseret og forecastede værdier.

Tabel 6.1 ovenfor viser, at forecastet på dækningsbidrag har en RMSE på 529.570, hvilket vurderes som en dårlig performance, baseret på at gennemsnittet for dækningsbidraget i perioden er 1.254.598 kroner. Det ses også i tabellen, at gennemsnittet for forecastet er cirka 500.000 kroner større end det realiserede dækningsbidrag. Som tidligere nævnt, bliver dækningsbidraget udregnet ud fra omsætning, vareforbrug og lønforbrug. Hvis vi kigger på performance på disse KPI'er kan det ses, at grunden til at dækningsbidraget overvurderes i så høj grad, skyldes at omsætningen også overvurderes, samtidig med at vareforbruget undervurderes. Især vareforbruget vurderes at have en dårlig performance med en RMSE på 344.589 sammenlignet med et gennemsnitlig realiseret tal på 1.372.559 kroner. Nedenstående graf 6.1 viser den forecastede og realiserede værdi af dækningsbidraget i valideringsperioden:



Graf 6.1: Dækningsbidrags forecast og realiserede værdier i mdkk i valideringsperioden.

Grafen viser det samme, som det der blev beskrevet før, nemlig at dækningsbidraget overvurderes i høj grad. Faktisk er det ikke nogen måneder i valideringsperioden, hvor det realiserede dækningsbidrag er større end det forecastede dækningsbidrag. Grafen understøtter konklusion om at forecastet for dækningsbidraget ikke har særlig god performance. Dog kan det ses, at forecastet ser ud til at ramme sæsoneffekten på en tilfredsstillende måde, sammenlignet med den realiserede sæsoneffekt, hvilket er positivt.

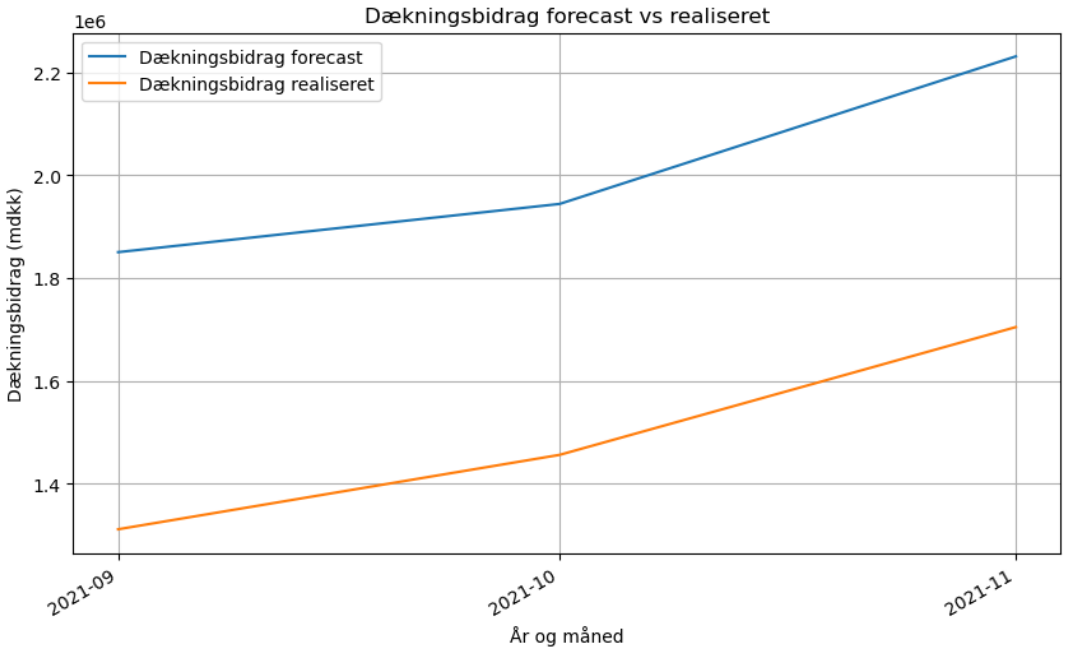
6.2 Test performance

Tabel 6.2 nedenfor viser performance af forecastet på de otte KPI'er, der er forecastet. Tabellen er sat op på samme måde som tabellen med validerings performance.

Test performance			
KPI	RMSE	Gennemsnit realiseret	Gennemsnit forecast
Dækningsbidrag	243.256	1.490.875	1.733.309
Omsætning	315.538	5.606.485	5.596.970
Vareforbrug	429.772	1.627.220	1.257.375
Lønforbrug	211.670	2.488.390	2.606.286
Faktureringsgrad	0,009	0,896	0,901
Jobtimer	1.491	8.998	10.131
Fakturerbar tid	1.371	8.062	9.128
Antal produktionsmedarbejdere	0,65	72,33	72,93

Tabel 6.2: Test performance af forecast modeller, samt gennemsnitlig realiseret og forecastede værdier.

Tabellen viser at forecastet i testperioden viser de samme tendenser som det gjorde i valideringsperioden, hvor det dog ser bedre ud når man kigger på RMSE for dækningsbidraget, som cirka er blevet halveret. Dækningsbidraget er overvurderet med cirka 250.000 kroner hver måned, hvilket skyldes at vareforbruget er undervurderet med cirka samme mængde. Dog ses det, at den gennemsnitlige forecastede og realiserede omsætning er tættere på hinanden i testperioden ift. valideringsperioden. Derudover er RMSE blevet bedre for omsætning, end det var i valideringsperioden, hvilket indikerer at modellen performer bedre på omsætning i testperioden end den gjorde i valideringsperioden. Det er værd at nævne at performance på lønforbrug er god både i valideringsperioden og testperioden. Derudover kan det ses at performance bliver dårligere i testperioden for både jobtimer og fakturerbar tid, hvilket er bekymrende, fordi det indikerer, at forecastet i den efterfølgende forecast periode, vil bevæge sig længere væk fra det, der vil blive realiseret i den periode.



Graf 6.2: Dækningsbidrags forecast og realiserede værdier i mdkk i testperioden.

Graf 6.2 ovenfor viser den forecastede og realiserede værdi af dækningsbidraget i testperioden. Her ses det at tendensen er den samme som i valideringsperioden, hvor dækningsbidraget overvurderes i alle måneder, men at forecastet er god til at fange sæsoneffekten, der er i dækningsbidraget.

Det er blevet vurderet at forecast modellen ikke kan bruges som et forecast værktøj for EL:CONs afdelinger, baseret på ovenstående performance og derfor blev det også besluttet ikke at udarbejde andre forecast modeller på andre afdelinger. Dog er der visse ting fra projektet, som er brugbar information for virksomheden, og som vil blive brugt i fremtiden. Blandt andet viste projektet, at performance på

forecastet på vareforbruget var dårligt, hvilket har skabt en dialog omkring, hvor svært det i virkeligheden er at forudse virksomhedens vareforbrug. Grunden til dette, er at der er forskel på hvilke type opgaver man arbejder på. Nogle opgaver kræver et højt vareforbrug for at kunne laves, mens andre opgaver kræver et lavere vareforbrug for at kunne laves. Vekslingen mellem de to opgavetyper, er noget, der er udfordrende at forudsige, især når man forsøger at forecaste 24 måneder ud i fremtiden.

Derudover har det været interessant for virksomheden at få tydeliggjort den sæsoneffekt der altid har italesat af ledelsen. Der er som tidligere nævnt meget udtalt og er for virksomheden meget relevant fortsat at kunne forecaste på når der laves 24 måneders rullende forecast.

7. Diskussion

I dette afsnit vil der blive diskuteret, hvilke muligheder der er, for at forbedre performance på den udviklede forecast model.

Som nævnt tidligere, kommer den dårligste forecast performance fra vareforbrugs forecastet. Derfor er det også det KPI der er mest interessant at undersøge, hvis man vil forsøge at forbedre performance. Her er det især interessant at kigge på korrelationsanalysen der blev lavet, da der skulle vælges KPI'er til vareforbrugs forecast modellen. Sammenligner man denne korrelationsanalyse med de øvrige korrelationsanalyser fra de andre KPI'er er det tydeligt, at vareforbruget ikke er så stærkt korreleret til særlig mange af de øvrige KPI'er. Derfor vil det være relevant at undersøge, om det er muligt at udarbejde ekstra KPI'er som potentielt kunne være bedre til at forudsige vareforbruget. Grunden til det ikke er gjort i dette projekt, er at det som nævnt tidligere er svært at forudsige de KPI'er, som påvirker vareforbruget. Her er der eksempelvis tale om hvilket miks af opgavetyper, man arbejder på.

En anden mulighed man kunne tage i brug, for at forsøge at forbedre forecastet på vareforbruget, er at forsøge at skabe en korrelation mellem vareforbrug og nogle af de øvrige variable i træningsdatasættet. Det kunne eksempelvis være jobtimer, hvor man kunne transformere vareforbruget ud fra den fordeling som jobtimerne følger. Altså kan man udregne hvor stor en anden af et års jobtimer der er realiseret i en måned, og så bruge den andel til at vurdere, hvor stor en andel af vareforbruget der også skal være realiseret i den måde. På den måde, vil man skabe en korrelation mellem jobtimer og vareforbrug, som vil gøre at forecast modellen for vareforbrug vil blive mere afhængig af jobtimerne, end det ellers var før. Det vil mindske problemet med, at vareforbruget nogle gange har måneder hvor det er ekstraordinært højt, uden man egentlig kan forklare hvorfor.

Et sidste alternativ man kunne bruge til at forbedre performance, vil være at bruge residualen fra resultatet fra tidsserie dekomposition. Hvis man fratrækker residualen fra de realiserede tal, vil man også kunne fjerne de ekstraordinære udsving, som der ses i vareforbruget.

8. Konklusion

Målet med denne opgave var at udvikle en datadreven machine learning model, som kunne forecaste dækningsbidraget for EL:CON's afdelinger. For at gøre det, blev der først udført en analyse af omsætningen i EL:CON på totalt niveau. Målet med denne var at opbygge en forståelse af virksomheden som helhed. Denne analyse var en additiv dekomposition, og viste at virksomhedens omsætning, har haft en positiv trend gennem de seneste år. Derudover viste den også, at virksomhedens omsætning er præget af sæsoneffekt. Til sidst viste analysen også, at der hver måned er store afvigelser, som ikke kan forklares gennem sæsoneffekt eller den trend omsætningen bevæger sig.

Herefter blev der udført en analyse af virksomhedens afdelinger. Det blev udført ved at udarbejde en SARIMA model på omsætningen fra 22 af virksomhedens afdelinger, og lave et forecast ud fra disse modeller. Resultatet af dette viste, at afdelingerne bevægede sig i meget forskellige retninger i forecast perioden på 24 måneder. Derfor blev det tydeligt, at der ikke kunne udarbejdes en forecast model, der ville virke for alle afdelinger, hvorfor det blev valgt at starte med en afdeling som eksempel. Her faldt valget på Service afdelingen i Aarhus.

For at udarbejde en forecast model der kan forecaste på dækningsbidraget i den valgte afdeling, blev dækningsbidraget delt op i tre dele: omsætning, vareforbrug og lønforbrug. Til hver af de tre KPI'er blev det besluttet at udvikle en forecast model. For at finde ud af hvilke KPI'er der skulle bruges i de forskellige forecast modeller, blev der i samarbejde med EL:CON udvalgt en række mulige KPI'er. Herefter blev de mest relevante valgt til hver model, ud fra korrelationsanalyse og variable importance output fra random forest modeller. Det resulterede i at der skulle udarbejdes forecast for otte KPI'er, før der til sidst vil blive udregnet et forecast på dækningsbidraget. For at udarbejde de forecast modeller der skulle bruges, blev det valgt at bruge PyCaret biblioteket som automatisk træner en masse forskellige forecast modeller, hvor den model der havde den bedste performance på validerings datasættet blev valgt som model.

Til sidst blev de udviklet et forecast for dækningsbidraget for Service afdelingen i Aarhus. Resultatet af denne var en RMSE på 529.570 i valideringsperioden, og 243.256 i testperioden. Det blev vurderet til ikke at være en tilstrækkelig god performance, og derfor blev det besluttet ikke at arbejde videre med forecast projektet. En af hovedårsagerne til at forecastet havde så dårlig en performance, var at forecastet for vareforbruget var undervurderet ift. de realiserede tal i hele perioden. På trods af at forecast modellen der blev udviklet i projektet ikke have tilstrækkelig god performance til at være brugbart for ELC:ON, var der alligevel nogle resultater fra opgaven, som har været brugbare. Her i blandt er eksempelvis resultaterne fra dekompositionen af virksomhedens omsætning og et fokus på hvorfor det er så svært at forudsige vareforbrug, som ellers er et vigtigt KPI for virksomheden.

9. Kilder

Analyticsvidhya, 2020. AdaBoost and Gradient Boost – Comparitive Study Between 2 Popular Ensemble Model Techniques. Lokaliseret den 3. januar 2022 på https://www.analyticsvidhya.com/blog/2020/10/adaboost-and-gradient-boost-comparitive-study-between-2-popular-ensemble-model-techniques/

Corporatefinanceinstitute. Elastic Net. Lokaliseret den 3. januar 2022 på https://corporatefinanceinstitute.com/resources/knowledge/other/elastic-net/

Datascience-pm. What is CRISP DM? Lokaliseret den 3. januar 2022 på https://www.datascience-pm.com/crisp-dm-2/

Machinelearningmastery, 2016. How to Check if Time Series Data is Stationary with Python. Lokaliseret den 3. januar 2022 på https://machinelearningmastery.com/time-series-data-stationary-python/?nowprocket=1

Otexts. 6.3 Classical decomposition. Lokaliseret den 3. januar 2022 på https://otexts.com/fpp2/classical-decomposition.html

PyCaret, 2020. Regression. Lokaliseret den 3. januar 2022 på https://pycaret.readthedocs.io/en/latest/api/regression.html

Pypi. Pndarima. Lokaliseret den 3. januar 2022 på https://pypi.org/project/pmdarima/

Sckit-learn 1.1. 1.1. Linear Models. Lokaliseret den 3. januar 2022 på https://scikit-learn.org/stable/modules/linear_model.html

Sckit-learn 1.10. 1.10. Decision Trees. Lokaliseret den 3. januar 2022 på https://scikit-learn.org/stable/modules/tree.html#tree

Sckit-learn 1.11. 1.11. Ensemble methods. Lokaliseret den 3. januar 2022 på https://scikit-learn.org/stable/modules/ensemble.html

Sckit-learn 1.17. 1.17. Neural network models (supervised). Lokaliseret den 3. januar 2022 på https://scikit-learn.org/stable/modules/neural_networks_supervised.html

Sckit-learn 1.3. 1.3. Kernel ridge regression. Lokaliseret den 3. januar 2022 på https://scikit-learn.org/stable/modules/kernel_ridge.html

Stat.yale. Linear Regression. Lokaliseret den 3. januar 2022 på http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm

StatisticsHowTo 1. Lasso Regression: Simple Definition. Lokaliseret den 3. januar 2022 på https://www.statisticshowto.com/lasso-regression/

StatisticsHowTo 2. Ridge Regression: Simple Definition. Lokaliseret den 3. januar 2022 på https://www.statisticshowto.com/ridge-regression/

Statsmodels. Statsmodels.tsa.seasonal.seasonal_decompose. Lokaliseret den 3. januar 2022 på https://www.statsmodels.org/dev/generated/statsmodels.tsa.seasonal.seasonal_decompose.html

Thedatascientist. Performance measures: RMSE and MAE. Lokaliseret den 3. januar 2022 på https://thedatascientist.com/performance-measures-rmse-mae/

The-modeling-Agency, 2000. CRISP-DM 1.0. Lokaliseret den 3. januar 2022 på https://www.the-modeling-agency.com/crisp-dm.pdf

Towardsdatascience, 2020. Everything you need to know about Min-Max normalization: A Python tutorial. Lokaliseret den 3. januar 2022 på https://towardsdatascience.com/everything-you-need-to-know-about-min-max-normalization-in-python-b79592732b79

Towardsdatascience, 2020. Machine Learning Basics with the K-Nearest Neighbors Algorithm. Lokaliseret den 3. januar 2022 på https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

Towardsdatascience, 2021 1. Understanding Autocorrelation in Time Series Analysis. Lokaliseret den 3. januar 2022 på https://towardsdatascience.com/understanding-autocorrelation-in-time-series-analysis-322ad52f2199

Towardsdatascience, 2021 2. CatBoost regression in 6 minutes. Lokaliseret den 3. januar 2022 på https://towardsdatascience.com/catboost-regression-in-6-minutes-3487f3e5b329

Del 2:

1. Redegørelse for valgte praktik sted og praktik opgave

Jeg har været i praktik i virksomheden EL:CON i perioden fra august til december 2021. Jeg vil ikke bruge mere tid på at introducere EL:CON, da der allerede har været en beskrivelse af virksomheden i del 1 af denne opgave.

Jeg har været ansat som studentermedhjælper hos EL:CON i en stilling som Business Controller siden februar 2019. Derfor var det også oplagt for mig, efter jeg havde valgt at jeg gerne vil i praktik at gøre det der. Det var det fordi, jeg på forhånd havde at dybdegående kendskab til virksomheden både forretningsmæssigt men også datamæssigt, da det er noget jeg har arbejdet meget med, i den tid jeg har været ansat der. En af fordelene ved at jeg har været i praktik ved en virksomhed jeg i forvejen kendte godt, er netop at jeg fra starten af kunne være med til at vælge, hvilket projekt jeg gerne ville arbejde med, hvilket jeg satte meget pris på. Derudover var det naturligvis også en fordel at jeg fra starten af kendte til virksomhedens struktur, så jeg ikke skulle bruge lang tid på at sætte mig ind i dette.

Derudover vidste jeg, at der var flere større dataprojekter som man i EL:CON godt kunne tænke sig at få udført. Derfor var der også flere muligheder oppe at vende, da jeg sammen med min chef skulle blive enig om, hvad jeg skulle lave i min praktikperiode. Her diskuterede vi en analyse af virksomhedens entreprise projekter, som skulle bruges til risikostyring, justering af virksomhedens data warehouse setup, og til sidst udvikling af en forecast model som kunne forecaste dækningsbidrag for afdelingerne i virksomheden.

Som beskrevet i del 1 af opgaven, endte vi med at beslutte at jeg skulle bruge min tid, på at arbejde med forecast af dækningsbidrag i EL:CON. Grunden til vi valgte netop det projekt, er at det er noget der er stor fokus på i virksomheden. Der er der i form af en ændring i måden der bliver lavet budgetter og targets. Derfor var det oplagt at vælge det projekt, der er mest aktuel på nuværende tidspunkt.

2. Beskrivelse af praktikperioden

Da det blev aftalt med EL:CON at jeg skulle være i praktik hos dem, blev det også aftalt, at jeg skulle bruge tre dage i ugen på at arbejde på mit praktikprojekt, og to dage i ugen på at løse de opgaver, jeg normalt lavede i virksomheden som studentermedhjælper. På den måde, vil jeg have lige så god tid til at lave mine normale opgaver, som jeg tidligere havde haft. Dog vidste det sig hurtigt, at det var svært for mig at opretholde den aftale, da mængden af opgaver jeg fik der ikke relaterede sig til praktikprojektet blev større i løbet af den tid jeg var der. Derfor er det endt med, at jeg ikke har brugt lige så lang tid på at arbejde på mit praktik projekt, som vi oprindeligt havde aftalt, men til gengæld har jeg fået løst en masse andre

opgaver, som jeg ellers normalt ikke vil have haft tid til. Disse opgaver har også været relevante i forhold til mit studie, og derfor har jeg ikke anset det som værende et stort problem.

Det lagde dog et stort pres på mig, da jeg hele tiden i min praktikperiode, skulle prioritere mit praktik projekt mod andre opgaver jeg modtog. Derfor vil jeg også nævne det, som en af ulemperne ved at være i praktik, samme sted som jeg har været ansat som studentermedhjælper. Der var dog aldrig noget pres fra virksomheden om, at jeg skulle nedprioritere mit praktikprojekt fremfor de øvrige opgaver. Faktisk var min chef god til at forsøge på at sørge for, jeg altid havde så meget tid som muligt til at arbejde på projektet. Grunden til, at jeg endte med at bruge mindre tid på praktikprojektet end vi først havde aftalt, skyldtes altså at jeg valgte at prioritere de øvrige opgaver højere i nogle tilfælde i løbet af projektet.

En af de svagheder jeg har lært ved om mig selv i løbet af min praktikperiode, er at jeg havde svært ved at sige nej til at hjælpe med øvrige opgaver, på tidspunkter hvor jeg havde afsat tid til at arbejde på mit praktikprojekt. Den sidste måned jeg var i praktik, blev dette problem dog bedre, da jeg sammen med min chef besluttede, at jeg skulle arbejde mere hjemmefra, når jeg arbejdede på mit praktikprojekt. På den måde, blev det nemmere for mig, ikke at blive involveret i øvrige opgaver, når jeg skulle arbejde på mit praktikprojekt.

En anden udfordring, som jeg stødte på i løbet af min praktikperiode, var mangel på sparring i forbindelse med udarbejdelse af forecast modellerne. I EL:CON er der ikke nogen ansat, som har erfaring eller viden i forhold til machine learning, og derfor manglede jeg altså til tider nogle jeg kunne spare med, når jeg løb ind i udfordringer, eller havde spørgsmål jeg gerne vil diskutere med nogen. I løbet af min praktikperiode lærte jeg, at jeg som person sætter pris på at arbejde i en afdeling, hvor jeg har nogen jeg kan spare med, i forhold til de opgaver jeg sidder og arbejder med. Dog skal det siges, at der intet problem var i forhold til sparring, når det handlede om forretningsforståelse, da jeg hurtigt kunne diskutere det med min chef, og de øvrige kollegaer i vores afdeling.

I min praktikperiode er jeg blevet bekræftet i, hvor meget jeg sætter pris på at arbejde i en stilling, hvor man har frihed. Med det mener jeg, både prioritering af opgaver, som nævnt tidligere, men også styring af arbejdstider. Hos EL:CON har jeg i min praktikperiode, selv haft mulighed for at styre, hvornår jeg er mødt på arbejde og hvornår jeg er taget hjem igen, så længe jeg har nået det antal timer vi har aftalt. Det er en måde at strukturere ens arbejdstider på, som jeg har sat meget pris på, og som helt sikkert er noget, jeg vil prioritere at have som mulighed, i mine fremtidige ansættelsesaftaler også.

Derudover er der i den afdeling i EL:CON jeg er ansat i, og har været i praktik i, stor tillid til medarbejderne i forhold til at de løser deres opgaver. Man skal altså ikke hele tiden rapportere til ens chef eller leder i

forhold til hvad man har brugt sin tid på og hvordan man løser sine opgaver. I stedet er der mulighed for selv at tage fat i ens chef eller øvrige i teamet, hvis man har brug for hjælp eller sparring til de opgaver man sidder med, eller til prioritering af arbejdsopgaver. Dette er en måde at arbejde på, som jeg i min praktikperiode er blevet bekræftet i, at jeg trives i.

Til sidst vil jeg nævne, at jeg også er blevet bekræftet i, hvor vigtigt det er for mig, at arbejde i en afdeling, hvor man samarbejder godt. På trods af at jeg som tidligere har nævnt, har manglet lidt sparring i forhold til machine learning i løbet af min praktikperiode, har jeg sat meget pris på den afdeling jeg har været en del af. Det jeg især har værdsat, er at man kan have en til tider uformel dialog omkring de opgaver man sidder og løser, og hele tiden har mulighed for hurtigt at sparre med hinanden, hvis man har brug for hjælp eller har spørgsmål til noget af det man sidder med. Det er helt sikkert også noget, jeg vil vægte højt i min fremtidige karriere.

Bilag

Bilag 1 Overblik og beskrivelse af modeller i PyCaret

1. 'lr' - Linear Regression

En linear regression model, er en model som forsøger at modellere sammenhængen mellem de uafhængige og den afhængige variabel, ved at forklare sammenhængen mellem de to som en lineær ligning. Denne type regulering kaldes for L1-regulering. (Stat.yale)

2. 'lasso' - Lasso Regression

Lasso regression fungerer som linear regression, men hvor koefficienterne i modellen bliver straffet. Målet med lasso regression er at lave mere simple modeller, og kan også bruges til at vælge variable, da nogle koefficienter i modellen kan blive sat til nul, så variablen ikke bliver brugt i modellen. Denne type regulering kaldes for L2-regulering (StatisticsHowto 1).

3. 'ridge' - Ridge Regression

Ridge regression er også en model der fungerer ligesom linear regression, men forskellen er igen at koefficienterne i modellen bliver straffet. Forskellen mellem lasso og ridge regression er at i ridge regression bliver alle koefficienter straffet med samme mængde, men i lasso regression kan det være forskelligt hvor meget de forskellige koefficienter bliver straffet (StatisticsHowto 2).

4. 'en' - Elastic Net

Elastic net er en kombination af Lasso og Ridge regression. Det betyder altså at den bruger både L1 og L2 regulering til at straffe koefficienterne i modellen (Corporatefinanceinstitute).

5. 'lar' - Least Angle Regression

Least Angle Regression er en model der foretager foreward stepwise udvælgelse, hvilket betyder at den starter med nul variabel, og herefter tilføjer de variable som er mest korreleret med den afhængige variabel (Sckit-learn 1.1).

6. 'llar' - Lasso Least Angle Regression

Lasso Least Angle Regression fungerer ligesom Least Angle Regression, men i stedet for at tilføje variable efter hver step, forøger man koefficienterne i modellen i forhold til hvordan de er korreleret med afvigelserne i modellen (Sckit-learn 1.1).

7. 'omp' - Orthogonal Matching Pursuit

Orthogonal Matching Pursuit er ligesom Least Angle Regression en foreward stepwise udvælgelse, men hvor den i hvert step tilføjer den variable med en koefficient forskellig fra nul, der er mest korreleret med afvigelserne i modellen, hvorefter den opdaterer afvigelserne i modellen (Sckit-learn 1.1).

8. 'br' - Bayesian Ridge

Bayesian Ridge er en model hvor, hvor den afhængige variable estimeres ud fra en sandsynlighedsmodel, hvor estimatet antages at følge en gaussian distribution (Sckit-learn 1.1).

9. 'ard' - Automatic Relevance Determination

Automatic Relevance Determination følger en gaussian distribution ligesom bayesian ridge, men forskellen er at hver vægt har sin egen standard afvigelse, som bruges når den afhængige variabel skal estimeres (Sckit-learn 1.1).

10. 'par' - Passive Aggressive Regressor

Passive Aggresive Regressor er en model som bliver trænet på en måde, hvor den bliver givet små mængder af data af gangen og optimerer parametre på baggrund af den nye data der bliver givet (Sckit-learn 1.1).

11. 'ransac' - Random Sample Consensus

Random Sample Consensus modellen splitter datapunkter mellem inliers og outliers, og forsøger at optimere antallet af inliers. En inlier er et datapunkt som ligger indenfor en bestemt grænse ift. hvor meget den må afvige fra det estimerede, hvor en outlier er den der ligger udenfor. Den endelige model er kun trænet på de datapunkter der er kategoriseret som inlier (Sckit-learn 1.1).

12. 'tr' - TheilSen Regressor

TheilSen Regressor er en model der generaliserer ved hjælp af median værdier, hvilket gør at den er bedre til at håndtere outliers end eksempelvis linear regression (Sckit-learn 1.1).

13. 'huber' - Huber Regressor

Huber Regressor er en model der på samme måde som Random Sample Consensus splitter data mellem inliers og outlier, men i stedet for at fjerne outliers, giver den mindre vægt til dem (Sckit-learn 1.1).

14. 'kr' - Kernel Ridge

Kernel Ridge er en model der kombinerer ridge regression med en kernel. En kernel transformerer data så ikke lineær ligning kan bruges i en lineær ligning (Sckit-learn 1.3).

15. ‘svm’ - Support Vector Regression

Support Vector Regression fungerer ligesom Kernel Ridge, men forskellen er den loss funktion de to modeller bruger. Begge bruger L2-regulering, men Support Vector Regression bruger også insensitive loss, hvor en Kernel Ridge i stedet bruger squared error loss (Sckit-learn 1.3).

16. ‘knn’ - K Neighbors Regressor

K Neighbors Regressor er en model, hvor den estimerer den afhængige variabel, ved at finde de k punkter i trænings datasættet hvor de uafhængige variable minder mest om hinanden, og tage gennemsnittet af den afhængige variabel for de k punkter den finder i trænings datasættet (Towardsdatascience, 2020). Denne er ikke brugt i denne opgave.

17. ‘dt’ - Decision Tree Regressor

Decision Tree Regressor er en model der ved hjælp af simple enten eller regler estimerer værdien af den afhængige variabel. Kompleksiteten af modellen afhænger af dybden på træet, hvor dybden fortæller hvor mange enten eller regler man skal i gennem inden man estimerer (Sckit-learn 1.10).

18. ‘rf’ - Random Forest Regressor

Random Forest Regressor er en bagging-tree estimator, hvilket betyder at den opdeler datasættet i tilfældige subsets og ud fra disse bygger et decision tree for alle subsets. Når der skal laves et estimat på den afhængige variabel, tages der et gennemsnit af de forskellige estimater, der kommer fra de forskellige træer (Sckit-learn 1.11).

19. ‘et’ - Extra Trees Regressor

Extra Trees Regressor fungerer på samme måde som Random Forest Regressor, men forskellen er at når data skal splittes i enten eller regler, gør den det tilfældigt i Extra Trees Regressor, hvorimod Random Forest finder det mest optimale split (Sckit-learn 1.11).

20. ‘ada’ - AdaBoost Regressor

AdaBoost Regressor er en boosting model, som betyder at modellerne bliver lavet i rækkefølge, hvor eksempelvis den anden model forsøger at rette på det den første model estimerede forkert. I AdaBoost Regressor bruger man simple modeller når man laver et boosted tree. Derudover laves der flere boosted trees i AdaBoost Regressor, hvor hver model har forskellige vægte som bruges når der skal laves en endelig forudsigelse (Sckit-learn 1.11).

21. ‘gbr’ - Gradient Boosting Regressor

Gradient Boosting Regressor minder om AdaBoost Regressor men forskellen er at det er afvigelserne fra den første model, der bliver givet videre til den næste model. Efter hver step forsøger man at minimere en loss function, eksempelvis RMSE, ved at ændre en smule på ens parametre. Retningen på ændringen afhænger af afvigelsen, og den learning rate man bruger i modellen (Analyticsvidhya, 2020).

22. ‘mlp’ - MLP Regressor

MLP Regressor er et neural netværk, hvilket betyder at den tager input fra de uafhængige variable, som så bliver transformeret i hidden layers og til sidst kommer ud i output layer. Målet er at minimere en loss function, som er squared errors (Sckit-learn 1.17).

23. ‘xgboost’ - Extreme Gradient Boosting

Extreme Gradient Boosting er en model som minder om Gradient Boosting Regressor, men den bruger avancerede L1 og L2-reguleringer, hvilket gør at dens performance ofte er bedre (Sckit-learn 1.11).

24. ‘lightgbm’ - Light Gradient Boosting Machine

Light Gradient Boosting Machine er en mere simpel version af en Gradient Boosting Decision Tree, som har fokus på at formindske tiden det tager at træne en model, og dermed gør det muligt at bruge større datasæt en man tidligere har kunne (Sckit-learn 1.11).

25. ‘catboost’ - CatBoost Regressor

CatBoost Regressor er ligesom Light Gradient Boosting Machine et forsøg på at lave en mere simpel og hurtigere version af en Gradient Boosting Decision Tree (Towardsdatascience, 2021 2).

Bilag 2 Kode fra Jupyter Notebook

Bilag 2 kan ses i vedhæftede HTML fil.

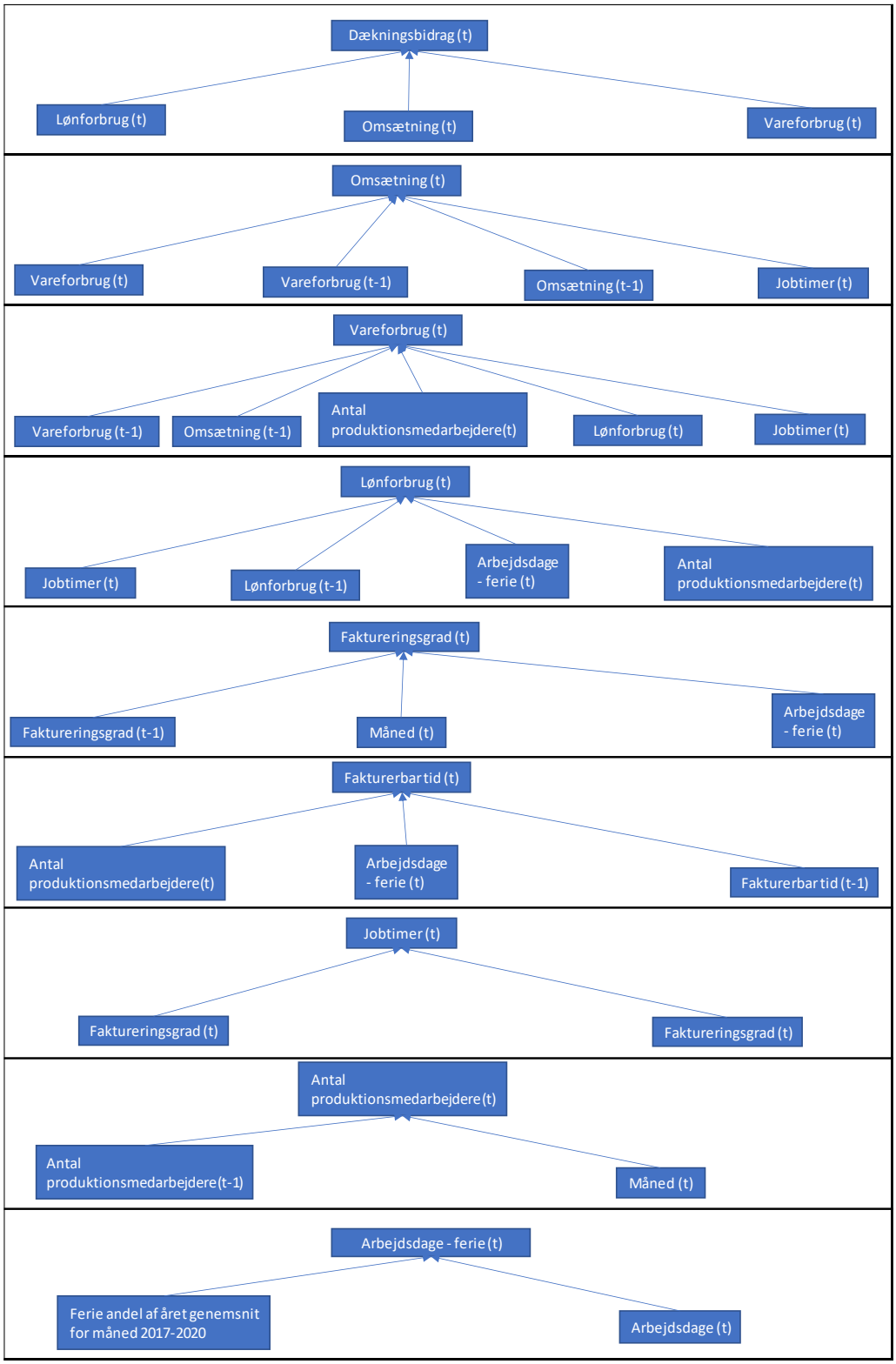
Bilag 3 Overblik over KPI'er med forklaring

KPI	Forklaring
Omsætning	Omsætning måles i kroner, og måler hvor meget omsætningen der har været i en måned. I EL:CON tages der højde for igangværende arbejder når der udregnes omsætning.
Vareforbrug	Vareforbrug måles i kroner, og måler hvor meget vareforbrug der har været i en måned.
Lønforbrug	Lønforbrug måles i kroner, og måler hvor meget vareforbrug der har været i en måned.
Faktureringsgrad	Faktureringsgrad måles i % og udregnes i EL:CON på følgende måde: Fakturerbar tid/jobtimer
Antal produktionsmedarbejdere	Antal produktionsmedarbejdere måles ud fra hvor mange medarbejdere der er kategoriseret som produktionsmedarbejder i virksomhedens lønsystem. Antal produktionsmedarbejdere udregnes på følgende måde: Antal produktionsmedarbejdere = (antal produktionsmedarbejdere ultimo + antal produktionsmedarbejdere primo)/2
Arbejdsdage - ferie	Arbejdsdage - ferie måles i antal dage og udregnes ved at tælle hvor mange arbejdsdage der her været i en måned. Herefter udregnes hvor mange feriedage der har været i den samme måned, ved at tage antallet af ferietimer der er registreret i den måned og dividere det med det samlede antal ferietimer der er registreret i det år. Denne andel ganges herefter med 30, som er det antal feriedage der er på et år, og det tal fratrækkes så antal arbejdsdage.
Fakturerbar tid	Fakturerbar tid måles i timer og er en tidstype der bruges i EL:CON, som indikerer at den time der bliver registreret kan faktureres til en kunde. Fakturerbar tid er således summen af de timer der er registreret som fakturerbar tid i en måned.
Jobtimer	Jobtimer måles i timer og består i EL:CON af tidstyperne, fakturerbar tid, sygdom, kursus og intern tid. Jobtimer er således summen af de timer der er registreret med en af de fire tidstyper i en måned.
År	År variablen fortæller hvilket år, den nuværende observation tilhører.
Måned	Måned variablen fortæller hvilken måned, den nuværende observation tilhører. Måned variablen er lavet så den består af 11 variable, hvor der altså er en variabel for hver måned undtagen januar, og alle variable er dummy variable, hvilket betyder at deres værdi er 0, hvis det ikke er den aktuelle måned, og 1 hvis det er den aktuelle måned

Bilag 4 Overblik over transformationer udført på KPI'er

KPI	Naturlig logaritme	Fjernet sæsoneffekt	Differenteret	Min maks skaleret
Omsætning		X	X	X
Vareforbrug		X	X	X
Lønforbrug		X	X	X
Faktureringsgrad				X
Antal produktionsmedarbejdere	X	X	X	X
Arbejdsdage - ferie				X
Fakturerbar tid	X	X	X	X
Jobtimer		X	X	X

Bilag 5 Struktur af forecast modeller



Bilag 6 Valgte forecast modeller

KPI	Model
Omsætning	MLP Regressor
Vareforbrug	Random Sample Consensus
Lønforbrug	CatBoost Regressor
Faktureringsgrad	Decision Tree Regressor
Fakturerbar tid	Random Sample Consensus
Antal produktionsmedarbejdere	Random Forest Regressor