Data scientist case

Jeppe Karnøe Knudsen

Oktober 2022

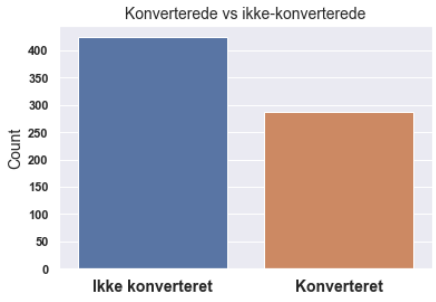
Indledning

Jeg er blevet givet et datasæt over brugere, som er kategoriseret som henholdsvis konverterede og ikke-konverterede. Derudover rummer datasættet de kategoriske parametre køn, kundesegment, eksisterende konti, og branch samt de numeriske parametre alder, initial fee, familiestørrelse og relaterede kunder. I denne rapport vil jeg forsøge at bestemme hvilke af disse parametre der er mest afgørende for kategoriseringen af brugere som værende konverterede.

I de første afsnit renser jeg datasættet og præsenterer et overblik over de forskellige parametre. Herefter implementerer jeg en Decision Tree-model, som giver estimater for hvilke parametre er de vigtigste i modellen. Til sidst diskuterer jeg disse konklusioner og peger på yderligere mulige undersøgelser.

Dataoverblik

Som indledende trin i min dataanalyse kigger jeg på målet for mine undersøgelser, konvertering, hvor jeg har undersøgt antallet af konverterede i forhold til det samlede antal af observerede personer. Jeg har fundet at 342 ud af det samlede antal af personer på 891 er konverteret. Dette giver en andel af konverterede for hele datasættet på 38.3% som illustreret i figur 1.



Figur 1: Overblik over konverterede og ikke-konverterede.

En god analyse forudsætter et godt datasæt. For at rense datasættet har jeg undersøgt antallet af manglende datapunkter indenfor hver af de givne variable. Dette har vist at følgende variable har det angivne antal tomme felter:

- Age: 177
- Branch: 2
- Credit Account ID: 687

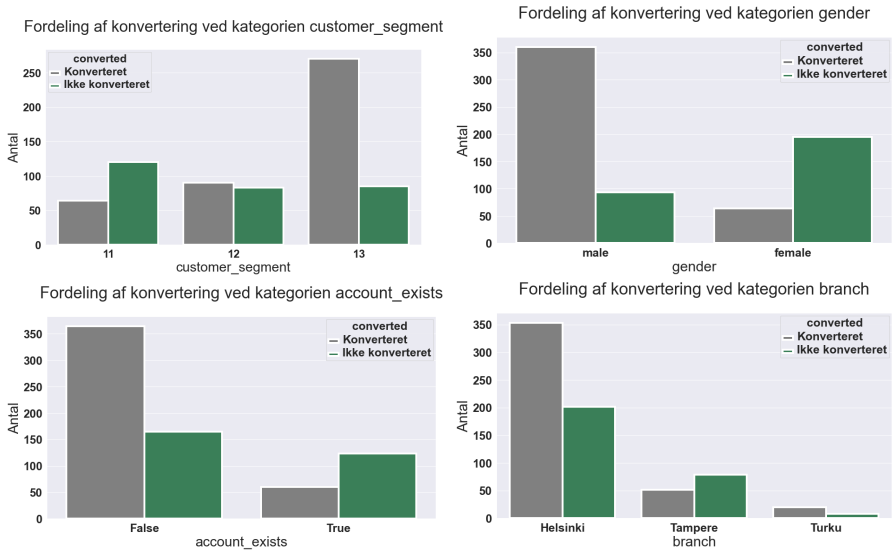
Altså er der en størstedel uden registrerede konti og en betragtelig mængde af manglende aldersbestemmelser og kun 2 personer med manglende branchbestemmelse. Disse håndterer jeg som følgende:

- **Konto ID:** De uidentificerede konto-ID'er er fundet ud fra den givne Hash for None-værdier. Her har jeg valgt at lave en ny kategorisk variabel for datasættet kaldet account_exists. Denne variabel er sand eller falsk, og bestemmer om den observerede har en registreret konto eller ej. Jeg ville muligvis kunne bruge regelmæssigheder fra de få med en identificeret konto til at finde estimater for disse, men jeg vurderede at min tid ville være givet bedre ud ved andre problemstillinger.
- **Alder:** En inspektion af de manglende datapunkter for alder viser, at 52 ud af de 177 – eller 29 % – er konverterede, hvilket er markant lavere end de cirka 60 % konverterede for alle observationer, og en eksklusion af disse datapunkter ville resultere i en uproportionelt stor mængde af ikke-konverterede. Man kunne vælge at forsøge at udlede estimater for alder, eventuelt ud fra median- eller gennemsnitsværdier for alder i datasættet, eller eksempelvis ved en Machine Learning algoritme. Dog har jeg valgt at fokusere mine kræfter andetsteds, og udelader observationer med manglende alder fra datasættet, selvom det kan have konsekvenser for mine konklusioner.
- **Branch:** Her mangler kun to værdier og man kunne undersøge om der er overlap ift. andre datapunkter, men grundet det lave antal og under hensyntagen til disse brugeres andre dataværdier såsom initial fee, vurderer jeg ikke at de påvirker det overordnede datasæt afgørende.

Efter denne håndtering af manglende datapunkter vil jeg i det følgende præsentere et overblik over fordeling af konverterede og ikke-konverterede i hver variabel i datasættet. Her har jeg delt data op i kategoriske og numeriske variable, og jeg vil gennemgå disse overkategorier individuelt.

Kategorisk data

I figur 2 er fordelingen af konverterede og ikke-konverterede plottet for alle de givne kategorier. Disse giver en række indledende observationer for hver kategoriske variabel:



Figur 2: Overblik over kategorisk data.

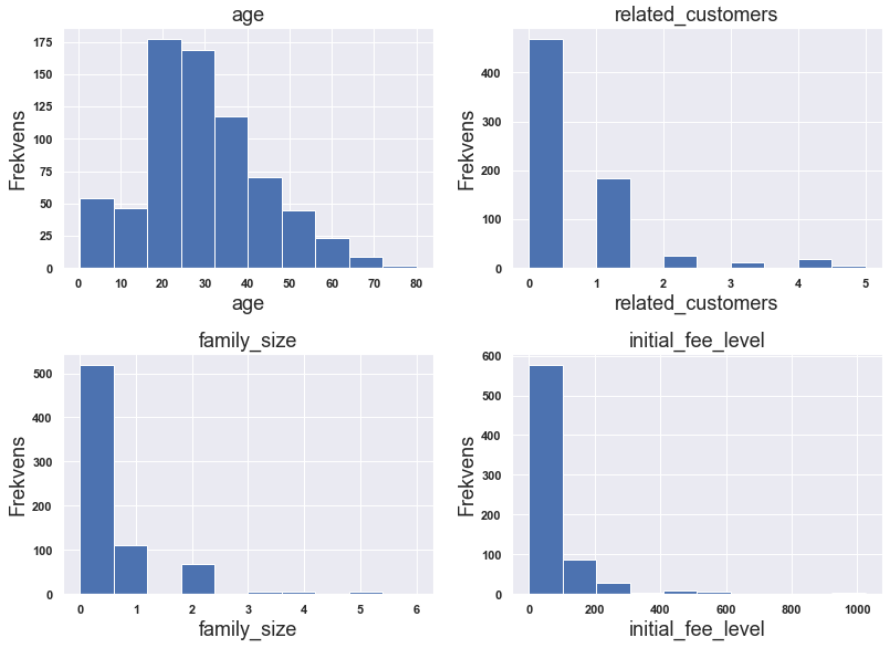
- **Customer segment:** Der ses klare forskelle i fordelinger af konverterede ved de tre kundesegmenter, og denne kategori er værd at undersøge yderligere.
- **Gender:** Her ses også markante forskelle mellem de to kategorier, og denne kategori vurderes ligeledes værd at undersøge.
- **Existing accounts:** Den store mængde af manglende konto-ID'er udgør cirka 80% af datasættet, og denne del har cirka samme fordeling af konverterede som resten af datasættet. Omvendt er brugere med konto-ID hyppigere ikke-konverterede og denne parameter kan således også vise sig at være interessant.
- **Branch:** Selvom der er variationer mellem de forskellige kategorier synes de umiddelbart ikke så store som ved de andre kategoriske parametre.

De indledende observationer for de kategoriske data tyder på, at kundesegment og køn er de betydeligeste faktorer til bestemmelse af om en person er konverteret eller ej, men en endelig konklusion kræver enten mere statistisk analyse, evt. i form af t-test eller ved implementering af en ML-model. Dog har vi fået et overblik over de forskellige kategorier, og kan se at der er markante forskelle for konverterede i de forskellige kategorier.

Numerisk data

For at få et overblik over fordelingen af observationer i de numeriske variable, plotter jeg deres fordelinger i histogrammer som vist i figur 3. Det primære

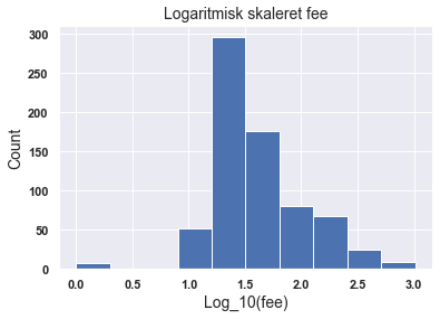
formål er at se, om de fordeler sig som forventet, og om der er skæve fordelinger der skal skaleres ved en eventuel implementering af en ML-model.



Figur 3: Overblik over numerisk data.

Fordelingen af alder er i overenstemmelse med min forventning og nogenlunde normalfordelt, så den behøver ikke skalering. Parametrene relaterede kunder og familiestørrelse har begge nogle få outliers, men de ligger indenfor samme størrelsesorden som de andre datapunkter og behøver heller ikke skalering. Den sidste parameter, initial fee level, har få, markante outliers, og hvis denne skal implementeres i en ML-model kan den med fordel skaleres. Dette har jeg gjort ved at tage log-10 af initial fee. Resultatet er en jævnere fordeling, som præsenteret i figur 4.

For at få et overblik over mulige sammenhænge mellem numeriske parametre og konverteringer har jeg kigget på gennemsnitsværdierne for de forskellige parametre for henholdsvis ikke-konverterede og konverterede, præsenteret i tabellen i figur 5. Her ser vi en relativt større forskel i gennemsnittet for initial fee level. Der er også en forskel for alder, familiestørrelser og antallet af relaterede kunder, men det kræver mere statistisk analyse at vurdere om de er afgørende. Dog kunne den store forskel ved initial fees tyde på at dette kunne være en afgørende numerisk faktor for bestemmelse af konvertering.



Figur 4: Logaritmisk skaleret initial fee.

converted	age	family_size	initial_fee_level	related_customers
0	30.626179	0.329690	44.235774	0.553734
1	28.343690	0.464912	96.790815	0.473684

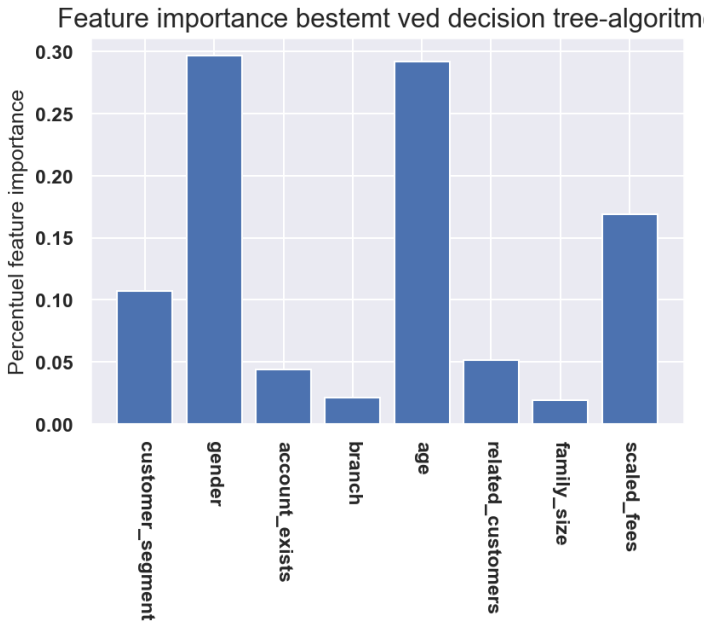
Figur 5: Gennemsnitsværdier for numeriske parametre for henholdsvis ikke-konverterede (0) og konverterede (1).

Vigtigste faktorer ved bestemmelse af konvertering

For at få et estimat på de forskellige faktorers vigtighed ved bestemmelse af konvertering, har jeg anvendt en Decision Tree-model. Denne model kan give såkaldte feature importances, som er procentuelle estimater på, hvor meget de forskellige parametre giver af information ved bestemmelse af målværdien, som i dette tilfælde er konvertering. Jeg har anvendt Scikit-Learn og implementeret modellen for hele datasættet (eksklusive brugere med manglende aldersbestemmelse og branch), og de resulterende værdier for feature importance kan ses i figur 6.

Resultatet af feature importances tyder på, at særligt to faktorer er afgørende for denne models karakterisering af en person som værende konverteret: køn og alder. Ved kategorien køn, kan vi med visualiseringen af konverterede ved de forskellige køn (figur 2 nederst til venstre) se, at der er større sandsynlighed for at være konverteret hvis man er mand. Ligeledes kan vi ud fra tabellen i figur 5 og vores feature importance-estimater anslå, at en lavere alder giver en lavere risiko for konvertering.

Af de øvrige faktorer kan vi se, at modellen indikerer at også initial fee og kundesegment er af betydning ved karakterisering af konverterede. De sidste datakategorier synes af relativt lille betydning, men her vil jeg bemærke at kategorien 'account_exists' er en kategori konstrueret på baggrund af, om der var en tilgængelig account-ID eller ej. Her ville der muligvis være flere indsigter at hente, hvis der blev foretaget yderligere undersøgelser af account-ID'erne. Man kunne eksempelvis undersøge om der var tendenser ved brugere med account-ID,



Figur 6: Fundne vægte (feature importance) for bestemmelse af konvertering ved Decision Tree model. Til denne model er Scikit-Learn's DecisionTreeRegressor brugt.

som kunne bruges til at bestemme de uidentificeredes konti.

De ovenstående feature importance-estimater skal ses i lyset af at den implementerede model kun er bygget på et datasæt, og ikke testet på et test-datasæt. Der er således en risiko for at modellen er overfittet, og i for stor grad har tilpasset sig de specifikke datapunkter. Måske burde jeg have delt datasættet op og testet modellen på den overskydende del, men jeg vurderede – særligt efter frasorteringen af brugere med manglende aldersbestemmelser – at datamængden var så lille at det ville være mere værd at få mere robuste parametre for feature importance fra den opbyggede model.

Konklusion

Jeg har med den ovenstående analyse forsøgt at besvare spørgsmålet om, hvilke parametre der er vigtigst for at forudse om en bruger er konverteret eller ej. For at kunne gøre dette udførte jeg en rensning af datasættet og præsenterede et dataoverblik i de første tre afsnit. Dette rensede datasæt brugte jeg til at opbygge en Decision Tree-model, som kunne give estimater for de vigtigste faktorer ved dens bestemmelse af om brugere er konverterede. Her fandt jeg, at alder og køn er de vigtigste parametre for bestemmelse af om en bruger er konverteret.

Ved at sammenholde de fundne feature importances med de indledende dataundersøgelser fandt jeg specifikt, at brugere med højere alder og af mandligt køn har større sandsynlighed for at være konverterede.

Selvom bestemmelsen af køn og alder som de vigtigste parametre for konvertering også ser rimelig ud ved sammenligning med den præsenterede data i figur 2 (og i mindre grad ved præsentation af aldersgennemsnit i 5), så ville det være oplagt at kvantisere disse forskelle yderligere. Her ville mit første skridt være at sammenligne fordelingen af konverterede for hvert køn med fordelingen af konverterede for hele datasættet med en t-test. Dette ville give et estimat for sandsynligheden for, at disse udfald ville komme fra den samme sandsynlighedsfordeling. Ligeledes ville en opdeling i aldersgrupper også give mulighed for sammenligning af konvertingsrater og alder ved en t-test og dermed give en øget kvantitativ indsigt i de forudsagte sammenhænge. Uanset hvad er den ovenstående analyse ikke udtømmende, og der er masser af indsigt at hente i datasættet endnu.

Perspektiver til skibbrud

Jeg har tidligere arbejdet med et datasæt over Titanics forlis på en onlineplatform for Machine Learning ved navn Kaggle. Da mit arbejde med det givne konverterings-datasæt skred frem, opdagede jeg ligheder mellem de konverterede brugere og ofrene for det historiske skibbrud. Jeg forsøgte at finde mindre variationer og uoverensstemmelser mellem de to datasæt men jeg er kommet frem til at der er stor korrespondance. Denne sammenhæng fandt jeg, da jeg søgte inspiration til min analyse af det givne datasæt. Er man ligeledes på jagt efter nye input til sine dataanalyser kan jeg varmt anbefale Kaggle.