Case study: Klassifiseringsanalyse

Daniel Munch Nielsen

15. august 2022

1 Introduktion og præsentation og rensning af data

Det givne datasæt består af 891 observationer og 10 variable. Et udsnit af data kan ses i tabel 1. Det ønskes at forudsige værdien af converted givet de resterende variable. customer id og credit

customer id	converted	customer segment	gender	age	related customers	family size	initia fee level	credit account id	branch
15001	0	13	male	22.0	1	0	14.5000	9b2d5b...	Helsinki
15002	1	11	female	38.0	1	0	142.5666	afa2dc...	Tampere
15003	1	13	female	26.0	0	0	15.8500	9b2d5b...	Helsinki
15004	1	11	female	35.0	1	0	106.2000	abefcf...	Helsinki
15005	0	13	male	35.0	0	0	16.1000	9b2d5b...	Helsinki
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

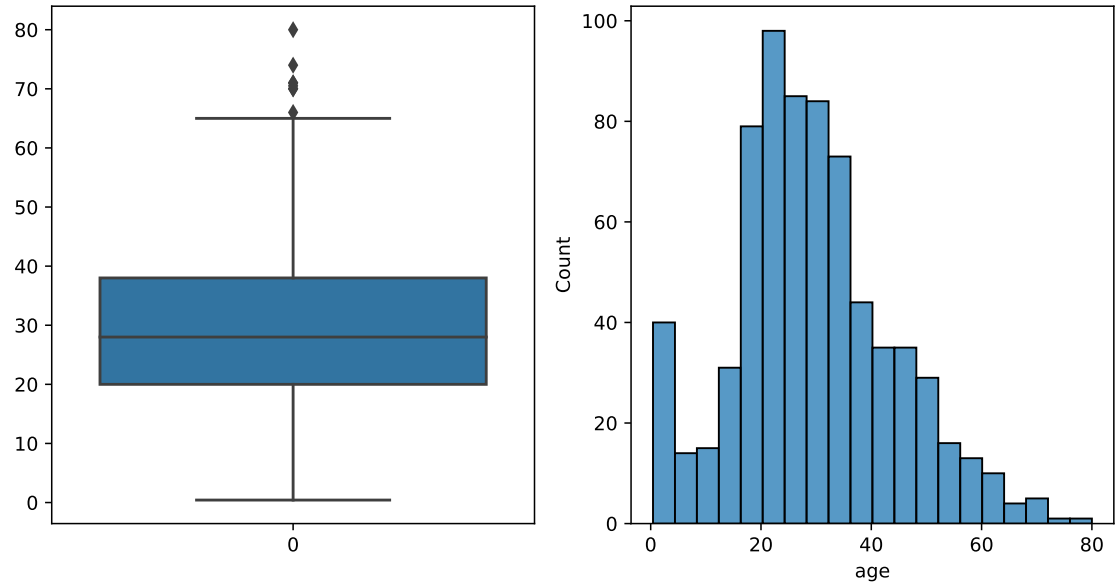
Tabel 1:

account id er irrelevante og fjernes derfor. Vi har de kategoriske variable: 'converted', 'customer segment', 'gender' og 'branch' og numeriske variable 'age', 'related customers', 'family size' og 'initial fee level' som givet i tabel 2.

converted	customer segment	gender	age	related customers	family size	initia fee level	branch
0	13	male	22.0	1	0	14.5000	Helsinki
1	11	female	38.0	1	0	142.5666	Tampere
1	13	female	26.0	0	0	15.8500	Helsinki
1	11	female	35.0	1	0	106.2000	Helsinki
0	13	male	35.0	0	0	16.1000	Helsinki
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Tabel 2:

Klassificeringsmodellen laves med en multiple logistisk regression. For at gøre dette må dataten først renses. 'branch'-info mangler i 2 observation, der derfor fjernes fra datasættet. Yderligere mangler 'age'-info i 177 observationer. Datasættet har en størrelse, hvor det formodentligt ikke ville være er problem at fjerne de 177 observationer, men vi vælger dog alligevel at udfylde. Da fordelingen af 'age' data er let højreskævt, som ses i figur 1, indsættes medianen af age i de manglende felter. Alle observationer indeholder nu værdier for alle variable. I en enkelt observation er der en indtastningsfejl i branch hvor der er angivet 'Turku}'. Denne rettes til 'Turku'. Som et sidste skridt i rensning af data fjernes duplikater, altså identiske observationer der optræder mere end en gang, således at præcision af den endelige model ikke overestimeres. Det resulterer i et datasæt med 8 variable og 773 observationer.



Figur 1: Fordeling af age data.

For at udfører logistisk regression skal de kategoriske variable erstattes med dummy variable. Resultatet ses i tabel 3. De ny variable skal forstås sådan at er en 'customer' placeret i segment 11, vil der være et 0 i 'customer segment 12' og 'customer segment 13' søjlerne.

converted	age	related customers	family size	initial fee level	gender male	customer segment 12	customer segment 13	branch Tampere	branch Turku
0	22.0	1	0	14.5000	1	0	1	0	0
1	38.0	1	0	142.5666	0	0	0	1	0
1	26.0	0	0	15.8500	0	0	1	0	0
1	35.0	1	0	106.2000	0	0	0	0	0
0	35.0	0	0	16.1000	1	0	1	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Tabel 3:

En logistisk regression kræver flere antagelser om data, som vi må tjekke. Det kræves at: logit funktionen er linear i de uafh. variable, at der ikke stærkt influeredende outliers, at der ikke er multikolinnearitet og slutteligt at observationerne er uafhængige. Alle disse ses at være opfyldt, (se pdf med notebook for detaljer), men der er dog en lille andel af observationerne, 0.6%, der er stærk influerende outliers. Disse kan ses i tabel 4. Med den givne viden om datasættet, er der ingen tegn på at der skulle være fejl i disse data, og derfor intet argument for at fjerne dem. Det vil formodentlig gøre modellen en smule mindre præcis, men til gengæld vil den fungere på et bredere sæt af data.

Vi er nu klar til at udføre regression med Pythons sklearn og statsmodels biblioteker. Uafhængige variable, der indgår i modellen med en p-værdi mindre end 0.05 vurderes ikke-signifikante og fjernes fra modellen en af gangen. Det resulterer i en model bestående af et konstant led og de uafhængige variable

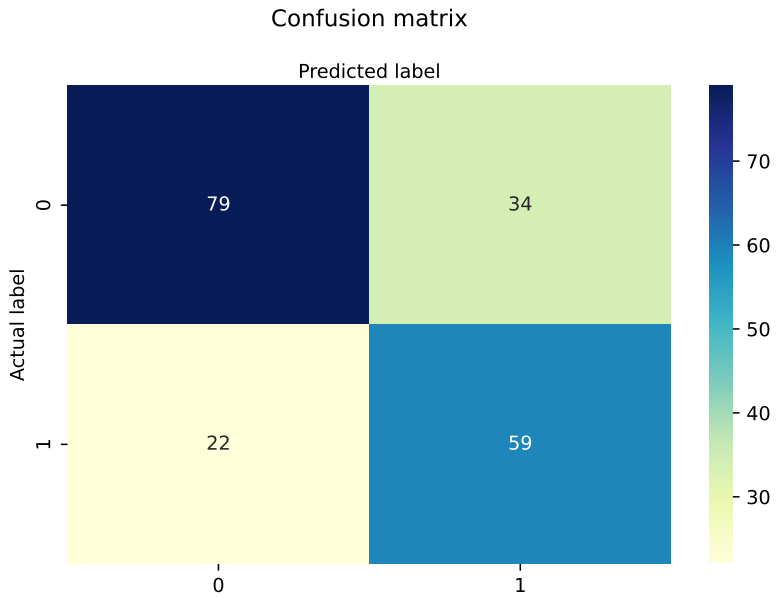
age, related customers, gender male, customer segment 12 og customer segment 13

Der er ingen simpel måde at evaluere kvaliteten af den fundne model. For at gøre dette laves en regression på 75% af date med de 5 uafh. variable foroven og den resulterende model benyttes til

Index	converted	age	related customers	family size	initial fee level	gender male	customer segment 12	customer segment 13	branch Tampere	branch Turku
281	0	28.0	0	0	15.7084	1	0	1	0	0
322	1	30.0	0	0	24.7000	0	1	0	0	1
327	1	36.0	0	0	26.0000	0	1	0	0	0
372	0	19.0	0	0	16.1000	1	0	1	0	0
562	0	28.0	0	0	27.0000	1	1	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Tabel 4:

at klassificere de resulterende 25%. Resultatet af klassificeringer kan visualiseres i en "confusion matrix i figur 2. Det ses at 79 observationer er korrekt klassificeret som ikke 'converted', 59



Figur 2:

korrekt klassificeret som 'converted' og 22 og 34 forkert klassificeret som hhv. ikke-'converted' og 'converted. Dette kan opsummeres i et 'accuary score', der angiver at modellen klassificerer korrekt i 71% procent af tilfælde. Dette er ikke fantastisk, men umiddelbart acceptabelt. Som et sidste skridt ses på odds ratio for the uafhængige variable, der indikerer hvordan hver variabel påvirker sandsynligheden for at en observation er 'converted'. Disse ses i tabel 5. Fra værdierne af odds

	5%	95%	Odds Ratio
age	0.946341	0.975774	0.960945
related customers	0.571794	0.880639	0.709608
gender male	0.055869	0.121554	0.082408
customer segment 12	0.213974	0.606056	0.360112
customer segment 13	0.063856	0.169453	0.104022
const	20.953058	99.895247	45.750529

Tabel 5:

ratio ses det, at højere alder og flere related customers begge medfører en mindre sandsynlighed

for at en customer er converted. Det ses yderligere, at det er meget mindre sandsynligt at en mand er converted end en kvinde og ligeledes for customer segment 12 og 13 ifht 11.

2 Resultat og Konklusion

Det kan konkluderes at de vigtiste parametre for at forudsige om en customer er converted er: age, related customers, gender og customer segment, og at hvis converted er målet, bør man fokusere på yngre kvinder i customer segment 11 med få eller ingen related customers.