Data Scientist - Case

Victoria Quist

October 7, 2022

Intro

Jeg har valgt at undersøge dataet ved hjælp af python. Jeg har valgt at bruge jupyter notebook, da jeg synes det giver et nemt overblik over koden samt figurer, og det er et godt format til at skrive kode der skal overskues hurtigt.

Første analyse med visualisering

Det første jeg gør er at tjekke at kolonnen *customer_id* er unik, og at der ikke er na i kolonnen. Dette er en måde at sikre mig at jeg kan bruge alt dataet som er udleveret.

Jeg kan godt lide at visualisere data, så det næste jeg gør er at dele dataet op i konverteret og ikke konverteret, også plotte de forskellige parameter som histogrammer for at se om der er et tydeligt skæld imellem konverteret og ikke konverteret for de forskellige kolonner. Grundet tidsmangel, er der ikke brugt tid på at gøre figurer pæne. Et plot af de forskellige parametre kan ses i figur 1. Her er det vigtigt at se at *credit_account_id* ikke er plottet. Dette er grunden til den lange "kode" som den kolonne indeholder, og min første tanke var derfor at gemme kolonnen til senere. Dog var der ikke et tydeligt svar på hvilken af parametrene som var de vigtigste for at forudse om en bruger var konverteret, så jeg tænkte jeg heller måtte prøve at kigge på *credit_account_id* også. Denne er plottet på figur 2, hvor *none* er frasorteret.

Ud fra disse figurerne var det stadig endnu ikke helt tydeligt hvilke parametre der forudsagde om en bruger var konverteret eller ej. Dog kan man se på plottet med *gender*, at der mange kvinder er konverteret og mange mænd ikke er konverteret. Derudover kan man se på figur 2 at der også er en forskel i *credit_account_id* på om de er konverteret eller ej. Så disse to kolonner ville være mit bedste bud indtil videre.

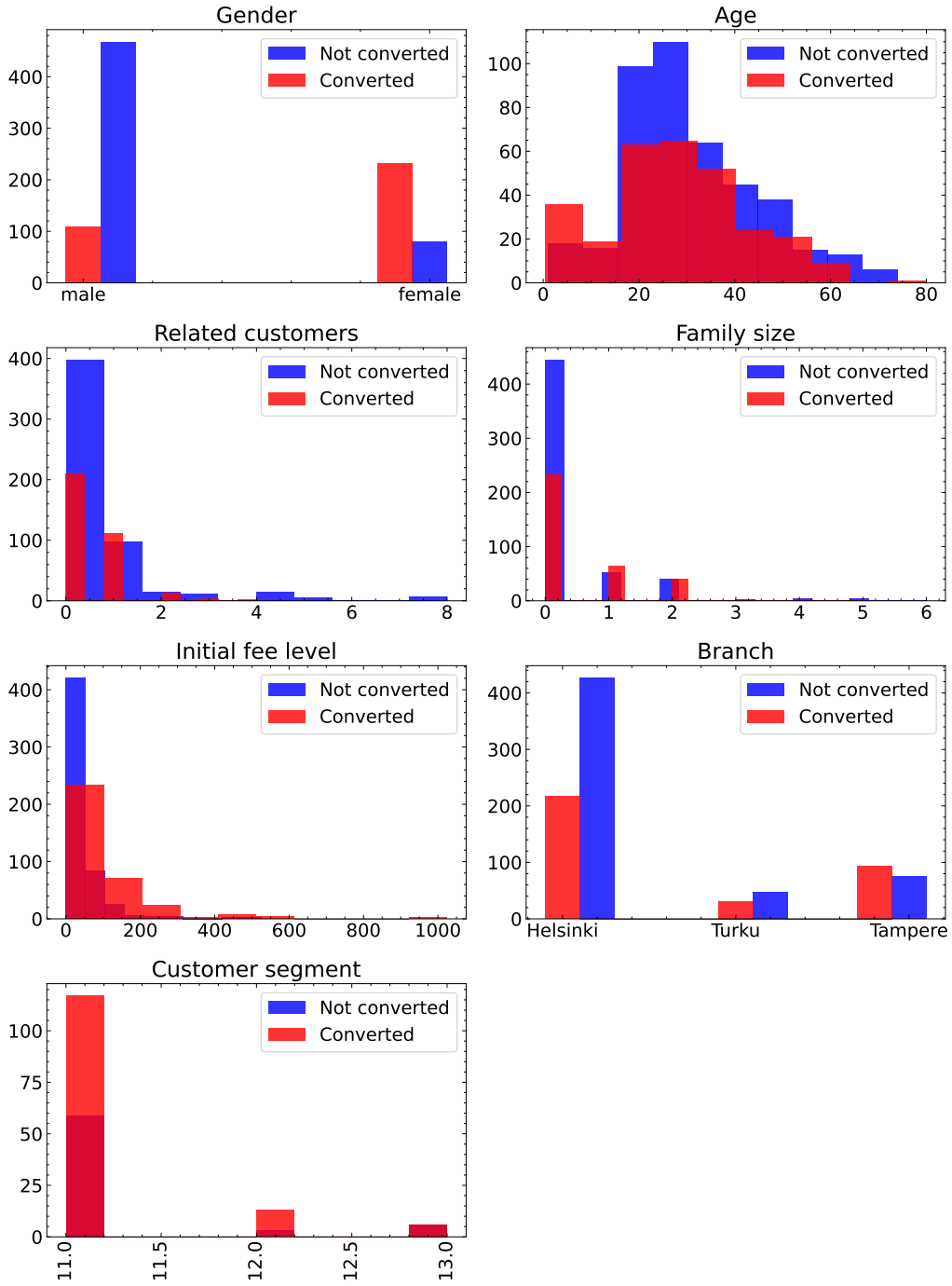

Figure 1: Plots ad de forskellige parameter

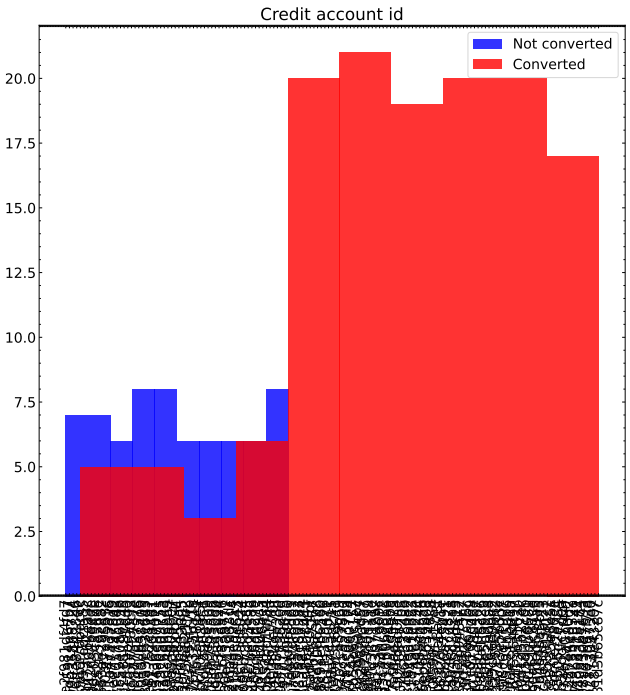

Figure 2: *credit_account_id* plottet

Jeg synes ikke selv, at plotsne var den bedste måde at finde svaret på. Da det er længe siden jeg har haft statistik, var jeg lidt i tvivl om hvordan jeg nemmest skulle gribe opgaven an nu. Dog var der to ting som jeg kom i tanke om. En Fisher's exact test eller finde korrelation imellem kolonnerne. Det skal dog siges at det var disse ting jeg tænkte på uden at have undersøgt hvad de egentlig viser. Jeg havde om disse ting på bacheloren af fysikstudiet, så det er også noget tid siden efterhånden.

Korrelation og Fisher's exact test

Jeg har ikke haft så meget tid til at sætte mig ind i præcis hvad de forskellige typer af korrelation er, og derfor har jeg valgt at regne både *Pearson, Kendall* og *Spearman* korrelation. Vores *converted* kolonnen har kun to mulige værdier, enten 0 eller 1, ved jeg ikke om det var den bedste ide at finde korrelationen. Jeg har dog gjort det alligevel, og de kan ses i tabel 1.

Her er det kønnet som har den største korrelation, så det viser det samme som vi fik fra det visuelle, dog ligner det ikke at credit accont spiller en rolle i, om en bruger er konverteret eller ej. Dette kan dog godt have noget af

	Pearson	Kendall	Spearman
Customer segment	-0.338481	-0.323533	-0.339668
Age	-0.077221	-0.043385	-0.052565
Related customers	-0.035322	0.085915	0.088879
Family size	0.081629	0.133933	0.138266
Initial fee level	0.257307	0.266229	0.323736
Gender	0.543351	0.543351	0.543351
Branch	0.108669	0.134032	0.137869
Credit account	-0.081003	-0.066670	-0.081318

Table 1: De forskellige korrelationer

gøre med den måde jeg har lavet kolonnen numerisk, så jeg ville stadig ikke udelukke dette. Derudover kan vi på tabellen se, at Initial fee level også har en ret høj korrelation, og det samme med customer segment. Dette kan man måske godt ane på plotsne på figur 1, men det er ikke meget tydeligt.

Efter jeg havde kigget på korrelationen, kiggede jeg på Fisher's exact test. Jeg skulle lige holde hovedet koldt, for at se mine matricer for mig, men da jeg fandt ud af hvordan de skulle opstilles lavede jeg en Fisher's exact test for gender, intial fee og family size. Grunden til jeg valgte at starte med disse var fordi jeg kunne se en opdeling i hvordan jeg ville dele dataet i to. Disse er vist i matricerne nedenfor:

$$\begin{bmatrix} male - converted & female - converted \\ male - notconverted & female - notconverted \end{bmatrix}$$

$$\begin{bmatrix} family = 0 - converted & family > 0 - converted \\ family = 0 - notconverted & family > 0 - notconverted \end{bmatrix}$$

$$\begin{bmatrix} initialfee <= 100 - converted & initialfee > 100 - converted \\ initialfee <= 100 - notconverted & initialfee > 100 - notconverted \end{bmatrix}$$

De p værdier jeg fik ud af at lave disse tre Fisher's exact test var meget lave, derfor valgte jeg ikke at gå videre med testen, og tænkte at jeg måske var ved at opfinde den dybe tallerken.

Noget tid efter jeg lavede denne test kom jeg i tanke om at Fisher's exact test prøver at vise at der ikke er forskel imellem tingene. F. eks. med den først matrice, her fik jeg en meget lav p værdi, og dette giver også mening i og med at der er forskel på om man er mand eller kvinde og er konverteret. Dette var blot en eftertanke jeg fik, og ville lige inkludere den.

Da jeg ikke rigtig synes jeg havde fået et klar svar fra korrelationen eller Fisher's exact test, valgte jeg som det sidste at kigge på konverterings raterne.

konvertering rater

Konvertering raterne er regnet ud ved at tage det samlede antal konverteret inden for en given kategori/underkategori, og divideret det med det samlede antal i alt. Her har jeg lavet meget manuelt kode, og har jeg burde have lavet en funktion til at regne det i steder for at gøre det manuelt. De forskellige konvertering rater kan ses i tabel 2.

Total konverteret	38.38 %
Male konverteret	18.89 %
Female konverteret	74.20 %
Family size = 0, konverteret	34.37 %
Family size = 1, konverteret	55.08 %
Family size = 2, konverteret	5 %
Family size > 2, konverteret	26.67 %
Initial fee level <= 100, konverteret	31.87 %
Initial fee level > 100, konverteret	68.13 %
Customer segment = 11, konverteret	62.96 %
Customer segment = 12, konverteret	47.28 %
Customer segment = 13, konverteret	24.24 %
Age <= 20, konverteret	45.81 %
Age > 20 eller <= 30, konverteret	36.52 %
Age > 30 eller <= 40, konverteret	44.52 %
Age > 40 eller <= 50, konverteret	38.37 %
Age > 50 eller <= 60, konverteret	40.48 %
Age > 60, konverteret	22.73 %
Related customers = 1, konverteret	53.59 %
Related customers = 2, konverteret	46.43 %
Related customers = 3, konverteret	25 %
Related customers > 3, konverteret	1 %
Branch Helsinki konverteret	33.7 %
Branch Tampere konverteret	55.36 %
Branch Turku konverteret	38.96 %
Credit account id none	29.99 %
Credit account id not none	66.67 %

Table 2: Konvertion rater

I tabellen med konverterings raten er der rigtig meget information, og det kan godt være lidt uoverskueligt. Det man skal ligge mærke til, at der er rigtig mange kvinder der er konverteret, imens der ikke er særlig mange mænd der er konverteret. Yderligere ligner det, at hvis brugeren har en høj inital fee level, så er der 68 % chance for brugeren er konverteret. Her skal man være opmærksom på at der ikke er så mange med så høj et fee level. Til sidst er der ret mange som har en credit account id som er konverteret, hvorimod at der ikke er så mange, som ikke har en credit account id, der er konverteret.

Konklusion

Ud fra de analyser jeg har lavet, vil jeg konkludere at det er kønnet(*gender*) som har størst indflydelse på om en bruger er konverteret eller ej. derudover har credit account (*credit_account_id*) også indflydelse, i og med at hvis brugeren har en (som ikke er none) så er der en stor sandsynlighed for at brugeren er konverteret, hvor imod at hvis brugeren ikke har en, så er brugeren nok ikke konverteret. En videre test kunne være at kigge på at hvis brugeren er kvinde, og har en credit account, hvor høj er sandsynligheden for brugeren er konverteret så, og det samme med hvis brugeren er mand og ikke har en credit account.

Jeg er også kommet ind på, at Intial fee *inital_fee_level* og customer segment *customer_segment* kunne være nogle af de parametre der var med til at forudse om en bruger var konverteret. Her vil jeg sige customer segment godt kunne være en af dem man kunne bruge, da et lav tal her vil betyde at der er større chance for at brugeren har konverteret. For intial fee, har brugeren nok konverteret hvis den har en høj værdi her, dog er der ikke særlig mange i datasættet der har en høj initial fee, og dette gør tallet mere usikkert.