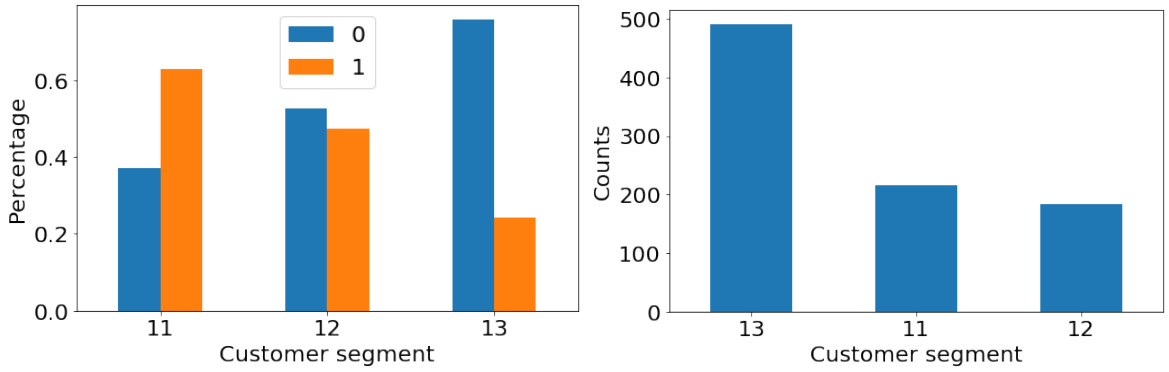
Case til jobsamtale nr. 2

Anna Torp Åkesson

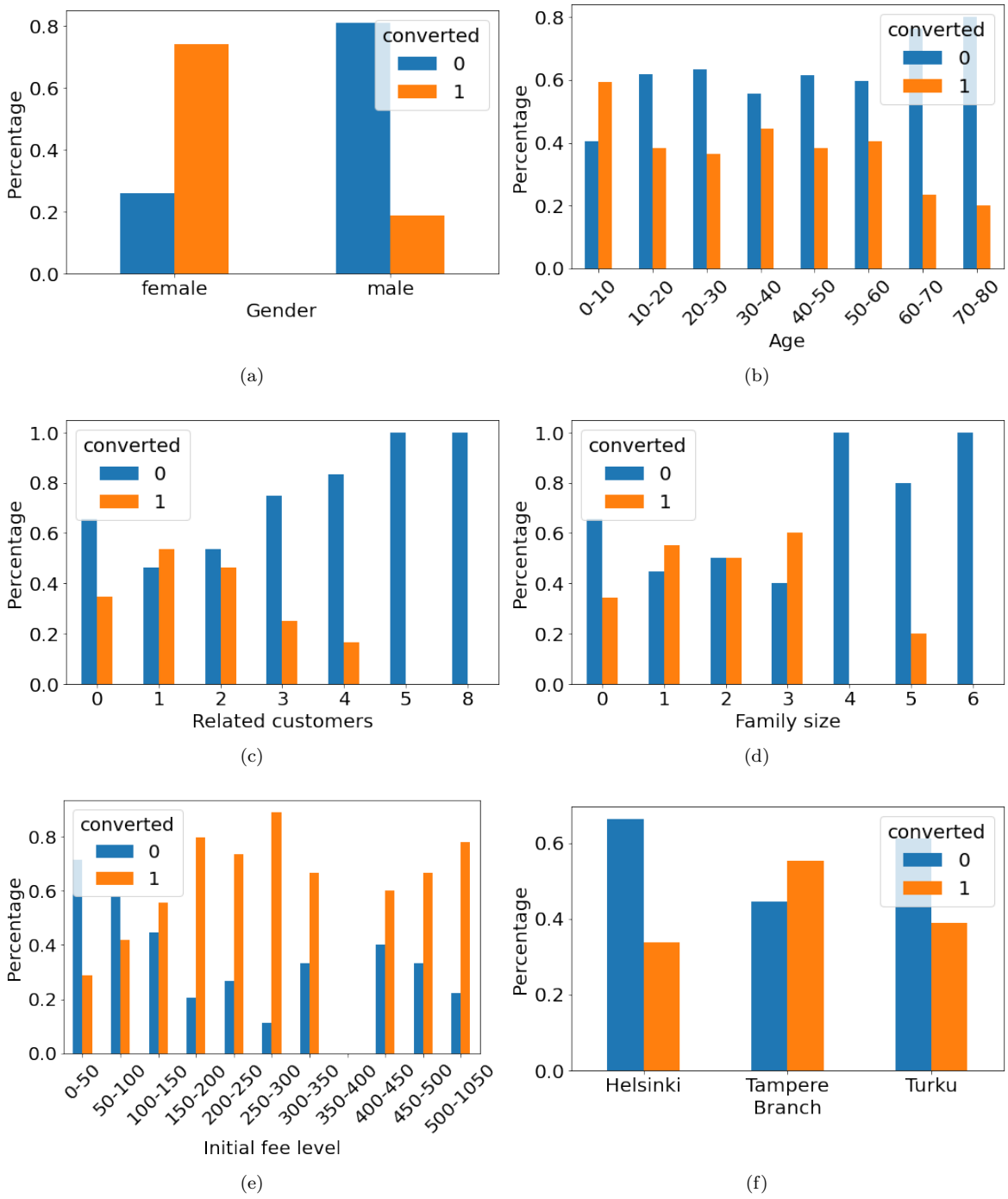
20. april 2022

De givne data inspiceres ved for hver af kategorierne customer_segment, gender, age, related_customers, family_size, intial_fee_level, credit_account_id og branch at lave et barplot, der viser hvor stor en procentdel af en værdi inden for hver parameter der er hhv. konverteret og ikke-konverteret. Det vil sige at for den første kategori, kundesegment (customer_segment), findes der tre forskellige værdier i datasættet, nemlig 11, 12 og 13. For hver af disse tre værdier ses i Figur 1(a) forekomsten af hhv. konverterede og ikke-konverterede brugere (hhv. 1 og 0) i procent. For eksempel er 63% af brugerne, der tilhører kundesegment 11 konverteret, og de resterende 37% er ikke konverteret. Denne type plot viser hurtigt om der er en umiddelbar sammenhæng mellem en parameter og andelen af konverterede brugere, og sådanne plots for de resterende parametre kan ses i Figur 2 og Figur4(a).



(a) (b)

Figur 1

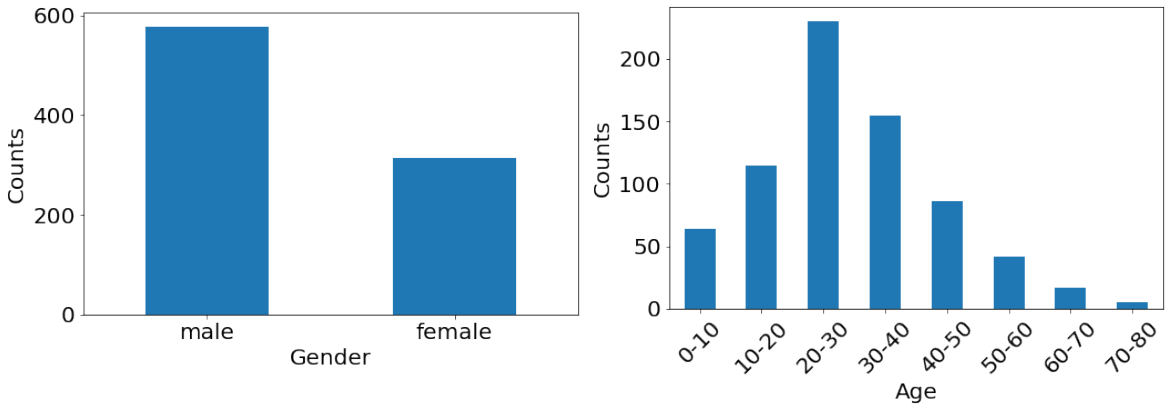


(a) (b) (c) (d) (e) (f)

Figur 2

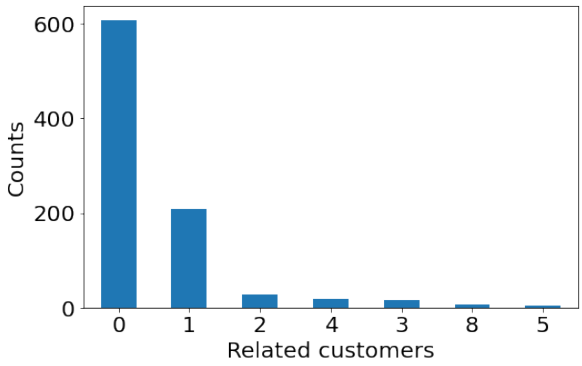
Disse relative forekomster af konverterede/ikke-konverterede brugere kan give et intuitivt billede af hvor stor en betydning hver parameter har. Dog er dette kun et retvisende billede, hvis der i hver kategori er er nok indgange til at man statistisk set kan betragte andelen af konverterede/ikke-konverterede brugere som en sandsynlighed. For eksempel viser Figur 1(a) umiddelbart en tendens til at andelen

af konverterede brugere falder, desto højere en værdi af kundesegment, og i Figur 2(a) ses en langt større andel af konverterede brugere ved kvinder en mænd. I de tre værdier af kundesegmenter (11, 12 og 13) er der hhv. 184, 216 og 491 brugere (se Figur 1(b)), så mens der ikke er lige mange brugere i hver kategori, bør der stadig være nok data til at fastslå en tendens. Samtidigt er fordelingen 314/557 mellem hhv. kvinder og mænd, så Figur 2(a) bør også danne et retvisende billede. Til gengæld er der for eksempel kun 28 eller færre indgange under hver værdi større end 1 i dataene for related customers, og derfor kan man ikke være sikker på at de procentuelle fordelinger hørende til disse værdier er repræsentative. Fordelingen af antal brugere i hver af kategorierne ses i Figur 1(b), og Figur 3 og giver altså en ide om hvor meget man kan regne med den procentuelle fordeling af konverteret/ikke-konverteret. For variablen credit account id, er der 148 forskellige ID-numre i datasættet, og de kan ikke opdeles i intervaller grundet deres natur, hvilket gør det svært at få et klart overblik over den procentuelle fordeling i figur 4(a). Betragtes antal indgange i datasættet (Figur 4(b)), ses det at et enkelt ID-nummer (9b2d5b4678781e53038e91ea5324530a03f27dc1d0e5f6c9bc9d493a23be9de0) står for langt størstedelen af dataene (ID-numrene er ikke noteret på den horisontale akse for overskuelighedens skyld).

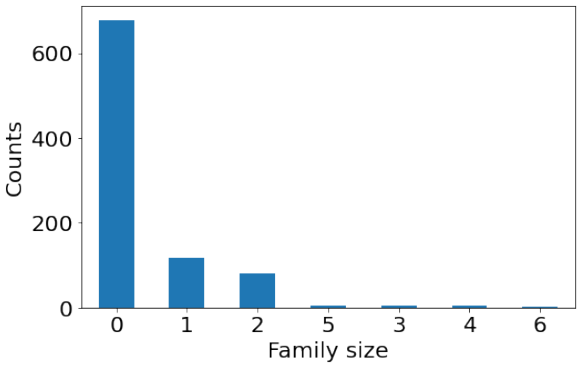


(a)

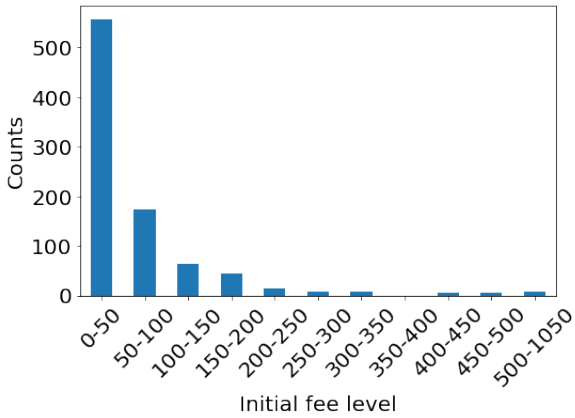
(b)



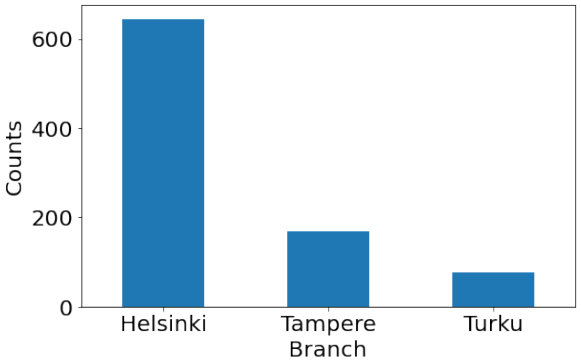
(c)



(d)

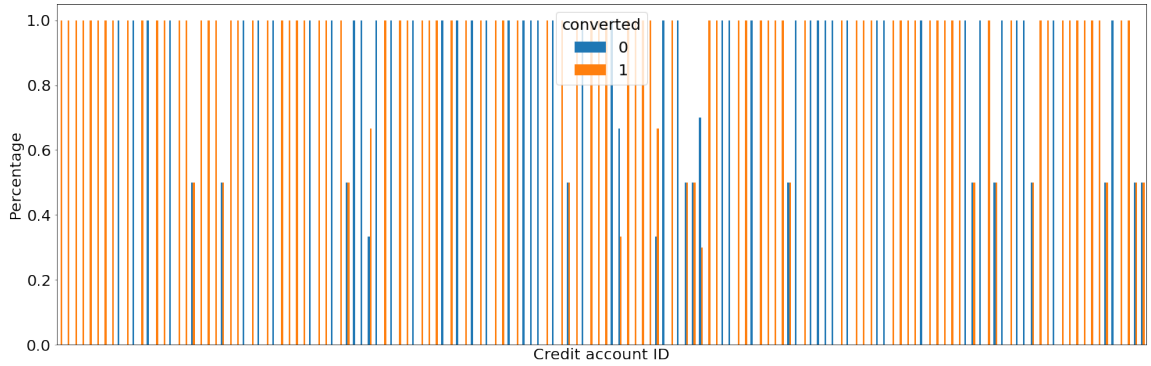


(e)

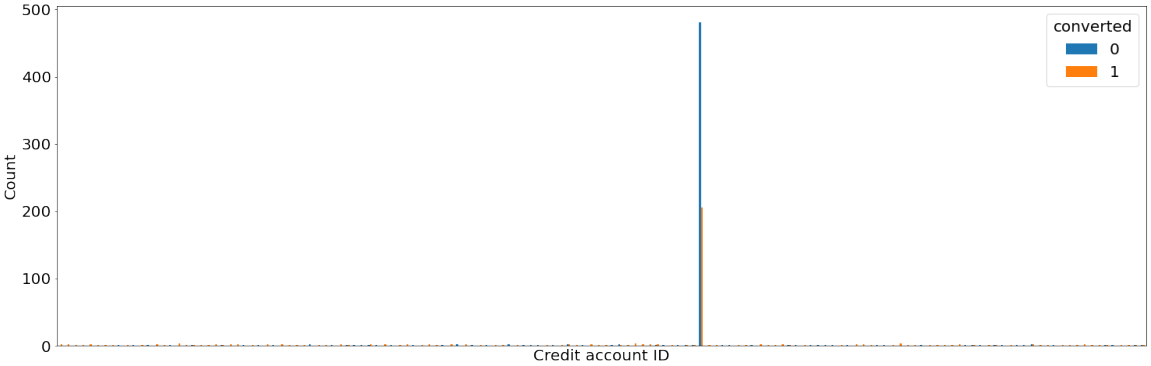


(f)

Figur 3



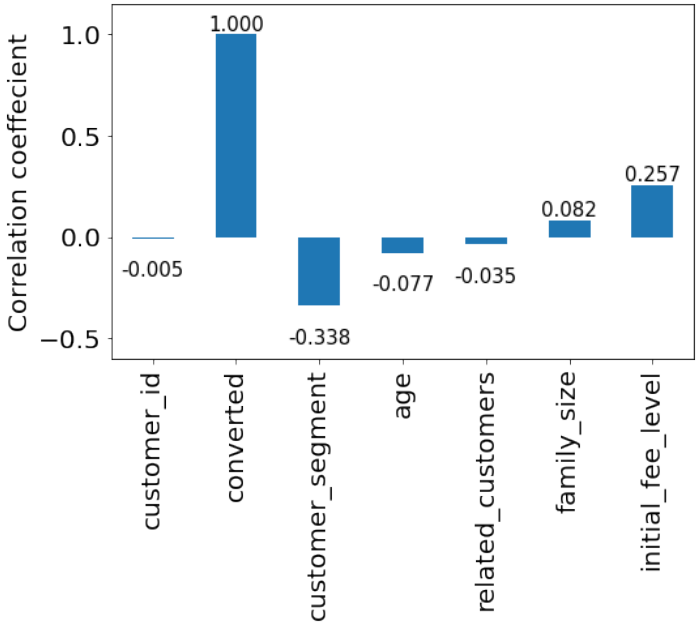
(a)



(b)

Figur 4

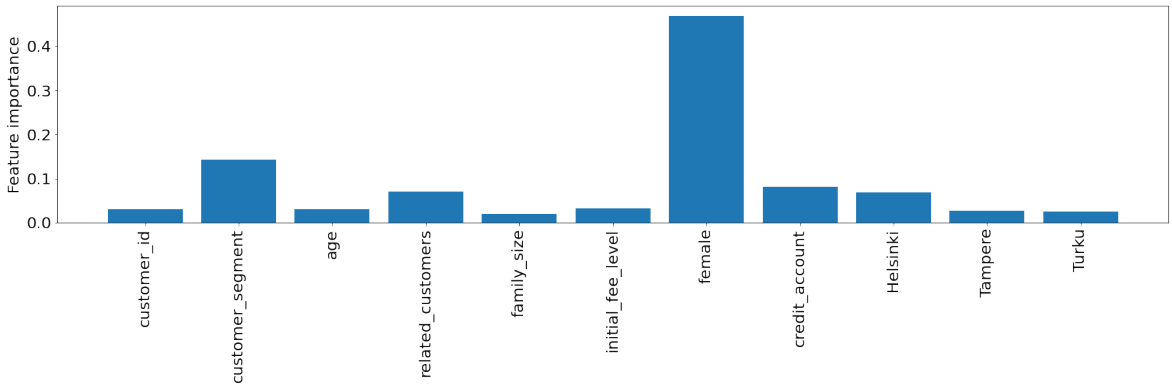
En måde at kvantisere sammenhængen mellem om en bruger er konverteret eller ej med en bestemt parameter, er at undersøge den lineære korrelationskoefficient r (også kaldet Pearsons korrelationskoefficient), som er et tal mellem -1 og 1. Korrelationskoefficienten vil have værdien -1 hvis relationen mellem to parametre x og y er lineær med en negativ hældning, og 1 hvis relationen lineær med en positiv hældning. Alle hældninger herimellem svarer til en korrelationskoefficient mellem -1 og 1, og for en flad, horisontal linje, dvs. for en konstant y-værdi, er $r = 0$, svarende til ingen korrelation mellem x og y. I Figur 5 ses et plot der viser korrelationen mellem converted-dataene og de numeriske variable. Som forventet er der selvfølgelig en fuldstændig, positiv korrelation med converted-dataene, da det er de samme data, og stort set ingen korrelation med customer_id, da denne kolonne med data blot nummererer de andre data. Det ses herfra, at customer_segment ot initial_fee_level har de højeste numeriske værdier for korrelationskoefficienten, og dette peger derfor på, at disse parametre har en vis betydning for om en bruger er konverteret eller ej, selvom det ikke er meget stærke korrelationer. Det må dog også understreges at dette tal kun beskriver lineære sammenhænge, og at en lav korrelationskoefficient derfor ikke udelukker en anden form for korrelation mellem variable.



Figur 5

Korrelationskoefficienten kan umiddelbart kun findes for numeriske parametre, dvs. at køn, credit_account_id og branch ikke kan inspiceres på denne måde. Dog kan køn konverteres til binære data, således at kvinde=1 og mand=0, og herudfra kan der findes en korrelationskoefficient på $r = 0,54$, hvilket indikerer den stærkeste korrelation mellem parametre. Til de resterende parametre (credit_account_id og branch) kan man som første redskab bruge en kvalitativ vurdering ved at betragte ovenstående bar plots, eller lave en χ^2-test med en nulhypotese om at variablene er uafhængige. Gøres dette og vælges et signifikansniveau på 0,05, finder man med værdierne $\chi^2_{branch} = 26,49$, $p_{branch} = 1,77 \cdot 10^{-6}$ og $\chi^2_{credit\_account\_id} = 241,0$, $p_{credit\_account\_id} = 1,62 \cdot 10^{-6}$ at nulhypotesen om at variablene er uafhængige af converted-dataene bør forkastes. Man kan også udforske en størrelse kaldet Cramers V som beskriver korrelationen mellem dataene. Her er $V_{branch} = 0.17$ og $V_{credit\_account\_id} = 0.52$, hvilket foreslår en stærk korrelation mellem branch og converted og en meget stærk korrelation mellem credit account ID og converted (se`https://www.pythonfordatascience.org/chi-square-test-of-independence-python/`).

En anden tilgang til problemet kunne være at anvende machine learning til at udpege de vigtigste parametre. Her kan man bruge en classifier (her XGBoost classifier) og bruge en del (her ca 2/3) af datasættet til træningen og resten til test for at fitte en model ud fra de angivne parametre, og herefter kan man ekstrahere feature importance, altså et mål for hvor vigtigt hver parameter er i klassifikationsprocessen. Siden input-dataene skal være numeriske, har jeg konverteret køn-kategorien til en numerisk, binær kolonne med navnet "female", hvor værdien 1 definerer "female"og værdien 0 "male". Ydermere har jeg i credit_account_id valgt at udelade alle andre ID-numre end det ene, der står for størstedelen af indgangene, siden der er så få indgange inden for de andre numre, at de ikke rummer nok brugbar information. Der laves derfor også en binær kolonne ud fra dette ID-nummer, og samme procedure følges for alle tre branch-værdier (Helsinki, Turku og Tampere). I Figur 6 nedenfor ses et plot af feature importance for hver parameter som resultat af træningsprocessen.



Figur 6

Som konklusion på ovenstående analyser, må køn siges at være den vigtigste parameter for at vurdere om en bruger er konverteret eller ej. Dette ses både ud fra Figur 6, men kan også hurtigt vurderes ud fra Figur 2 og ses ud fra korrelationskoefficienten. Herudover peger både machine learning-algoritmen og korrelationskoefficienterne på at kundesegment er den næstvigtigste størrelse. Herefter bliver uoverensstemmelserne større mellem machine learning-algoritmen og korrelationskoefficienterne, og det tyder på at initial fee level, credit account ID, related customers og muligvis branch bør være de næste parametre der bør undersøges for at vurdere om en bruger er konverteret eller ej.