



Hjemmeksamen
MET4
Vår 2018

**Kredittkort
og mislighold**

Kandidatnummer: [REDACTED]

Antall sider: 10

**EKSEMPELBESVARELSE
MET4 HJEMMEKSAMEN**

Denne besvarelsen ble levert som hjemmeksamen i faget MET4 vårsemesteret 2018 og er publisert i anonymisert form i samråd med den aktuelle studentgruppen og eksamenskontoret ved NHH. Vi takker for samarbeidet.

Se egetvedlegg for sensors kommentarer til denne oppgaven

Introduksjon

Kredittkortgjeld er en av de vanligste årsakene til at man kommer i alvorlige gjeldsproblemer (Forbes, 2017). Om banker ikke setter høye nok krav til anskaffelse av kredittkort kan det føre til en kredittkortkrise, noe som skjedde i Taiwan i 2006, da mer enn en halv million mennesker ikke kunne betale tilbake lånene sine. Disse fikk kallenavnet "kredittkort-slaver" fordi de så vidt kunne betale det minste månedlige beløpet og Taiwan endte opp med store samfunnsproblemer (Seven Pillars Institute, 2018). Det ble dermed ansett som uforståelig og uetisk å utstede kredittkort til personer med høy sannsynlighet for å mistlyholde lånet. På grunn av dette må banker i større grad avgjøre om kunder vil klare å betale tilbake kredittkortgjelden for at de skal utstede kredittkort. Dataene i denne rapporten er fra 2005 og det er tydelig at det går mot et stort nivå av mistlyholdte lån. I denne rapporten skal vi utforske kunders tilbakebetalingsevnen basert på diverse egenskaper. Formålet er å bestemme sannsynligheten for mistlyholdelse mot betaling av lån neste måned blant kredittkortkunder for å videre kunne gi anbefalinger til banken om fremtidig salgsstrategi for kredittkort.

Spørsmål 1

I analysen benytter vi oss av datasett fra University of California, som inneholder 30 000 observasjoner av kredittkortbrukernes egenskaper og betalingshistorikk de siste seks månedene. Datasettet inneholder beskrivende data om hver enkelt bruker. Vi vil se på data med egenskaper som beskriver bankens kredittkortkunder i forhold til hvor stor sannsynlighet de har til å mistlyholde kredittkortlånet neste måned. Dette skal videre brukes til å formulere retningslinjer til bankens salgsstrategi fremover. Egenskapene vi anser som relevante for vår analyse av kundene er dermed maksimal kreditt, kjønn, utdanning, sivilstatus og alder.

Vi har valgt å inkludere typetall som viser hvilke verdier som forekommer oftest, standardavviket som viser verdienes gjennomsnittlige avvik fra gjennomsnittet og i tillegg median, samt min- og maks verdier.

I dataene om utdanning og sivilstatus var det flere observasjoner som ikke defineres i datasettet, som vil si at de hadde tall som ikke kan representeres i regresjonen. Disse eliminerte vi før vi fant den beskrivende statistikken for datasettet. I utgangspunktet vurderte vi å inkludere disse i kategorien "annet", men konkluderte med at disse tallverdiene også kan inkludere feildata.

Tabell 1 viser den deskriptive statistikken til egenskapene: Kredittgrense, utdanning, sivilstatus og alder. Kjønn er ikke inkludert siden den er binær, men det kan nevnes at statistikkutvalget består av 11 888 menn og 18 112 kvinner.

	Kredittgrense	Utdanning	Sivilstatus	Alder
Gjennomsnitt	167484,3227	1,8172	1,5547	35,4855
Median	140000	2	2	34
Typetall	50000	2	2	29
Standardavvik	129747,6616	0,7113	0,5183	9,2179
Minimum	10000	1	1	21
Maksimum	1000000	4	3	79

Tabell 1 - deskriptiv statistikk

Tabell 2 viser sannsynlighet og antall misligholdelser av betaling neste måned ut fra hver enkelt forklaringsvariabel uavhengig av hverandre. Vi valgte å inkludere dette fordi det kan korreleres med regresjonene vi skal utføre i de neste oppgavene og brukes som en indikator. Det er viktig å påpeke at tallene blir mindre representative desto færre det er i utvalget. Dette kan vi for eksempel se ved at innenfor et utvalg på 280, har bare 18 personer misligholdt lånet. Prosentverdien 6,43 % (for personer med utdanningsbakgrunn "andre studier") har derfor liten validitet.

Misligholdelse av betaling neste måned	%	Antall
Menn	24 %	2873
Kvinner	21 %	3763
"Graduate"	19,23%	2036
Universitetsutdanning	23,73 %	3329
Videregående	25,16 %	1237
Andre studier	6,43 %	18
Gift	23,47 %	3206
Singel	20,93 %	3341
Annen sivilstatus	26,01 %	84

Tabell 2 - misligholdelse av betaling neste måned (individuell egenskap)

Spørsmål 2

I denne deloppgaven skal det undersøkes hva som øker og reduserer sannsynligheten for et bestemt utfall av en "dummy"-variabel.

Valg av regresjonsmodell

Siden vi skal finne sannsynligheter for misligholdelse av betaling neste måned så er ikke lineær regresjon optimalt da den kan gi negative verdier og større enn én, noe sannsynlighetsprosenten ikke kan. Utfallet fra en lineær regresjon har uendelige tall av mulige verdier i motsetning til en logistisk regresjon der utfallet har et begrenset antall mulige verdier. Logistisk regresjon brukes når responsvariabelen er kategorisk av natur som i denne sammenheng er 0 eller 1, der 1 er predikert mislighold av gjeldsbetaling neste måned og 0 er predikert betaling neste måned. I tillegg vil den lineære regresjonsmodellen ha samme marginaeffekt for økning i en variabel og viser raskt om det er heteroskedastisitet, som vil si at det er underpopulasjoner med ulike variabler enn andre. Derfor vil vi benytte oss av logistisk regresjon.

Både logit- og probit- modeller er passende å bruke når man skal estimere en dikotom avhengig variabel. Logit- og probitmodellen gir ofte svært like resultater da de begge tar en lineær modell og filtrerer den gjennom en funksjon som gir et ikke-lineært forhold.

Den lineære prediksjonsformelen ser slik ut: $\hat{Y} = \alpha + \beta x$.

Logit og probit prediksjonsformelen ser slik ut: $\hat{Y} = f(\alpha + \beta x)$.

Forskjellen ligger i hvordan de definerer $f(\cdot)$. Logit bruker kumulativ distribusjon av den logistiske distribusjonen for å definere $f(\cdot)$ og Probit modellen bruker en kumulativ distribusjonsfunksjon av standard normal distribusjon. Normal distribusjon gir ikke mening i denne sammenhengen fordi vi bruker mange ulike ikke-korrelerte variabler med bestemte kategoriske verdier. Vi bruker derfor logistisk regresjon fordi kumulativ distribusjon eger seg bedre til multivariate tilfeldige variabler (Keller, 2012).

Regresjonen

I denne regresjonsmodellen har vi endret tallverdiene på forklaringsvariabelen «Utdanning» fra et intervall på 1 til 4 til en kategorisk variabel med alternativene 1 for «graduate»-utdanning til 0 for kundene med annen utdanning. Det samme gjelder forklaringsvariablene «sivilstatus» og «kjønn», som vi har oversatt til dummy-variabler med verdien 1 for gift og 0 for de annet, og 1 for mann og 0 for kvinne. Dette ved hjelp av funksjonen «ifelse».

Tabell 3 viser regresjonsresultatet utfra formel `glm (Default ~ Utdanning + Sivilstatus + Kjønn, family = binomial(link = 'logit', data = d))`

Dependent variable:	
	Default
Utdanning	-0.253*** (0.030)
Sivilstatus	0.115*** (0.028)
Kjønn	0.205*** (0.028)
Constant	-1.311*** (0.026)
Observations	30,000
Log Likelihood	-15,778.970
Akaike Inf. Crit.	31,565.950

Note: *p<0.1; **p<0.05; ***p<0.01

Tabell 3 – Logistisk regresjon - "graduate", sivilstatus, kjønn

Den logistiske regresjonen av datasettet med mislighold som responsvariabel, "graduate"-utdanning, status som gift og kjønn som forklaringsvariabler gir oss følgende funksjon for sannsynlighet av mislighold: $Y = -1,311(\text{konstant}) - 0,253(\text{graduate-utdanning}) + 0,115(\text{gift}) + 0,205(\text{mann})$. Denne funksjonen kan i teorien brukes til å predikere tilbakebetalingsevnen til en gift kredittkunde ved å bruke de relevante variablene. Dersom vi tar for oss en gift mann med "graduate"-utdanning, kan sjansen for mislighold beregnes slik:

$$Y = -1,311 - 0,253(1) + 0,115(1) + 0,205(1) = -1,244$$

$$e^{-1,244} = 0,2237405$$

Dette forteller oss at den gifte mannen med "graduate"-utdanning har en 22,37% sjanse for å misligholde kredittgjelden sin. En kvinne med alt annet like ville henholdsvis hatt en 19,02% sjanse for å misligholde.

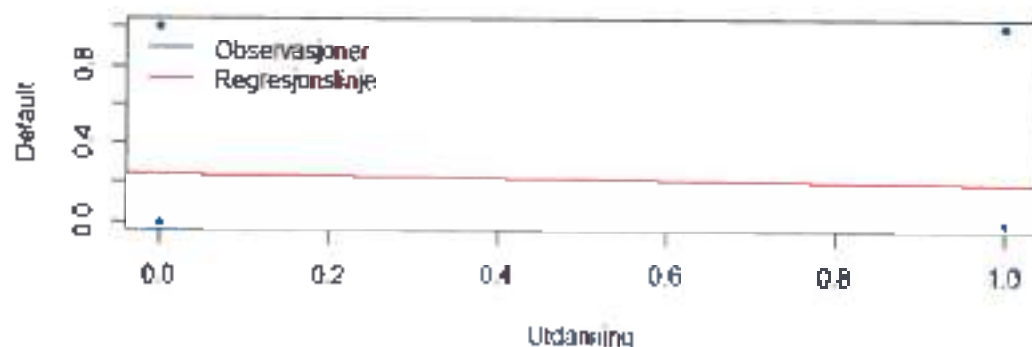
Videre har vi testet modellens reelle egenskap til å predikere mislighold i datasettet. Dette har vi gjort ved å sammenligne alle de predikerte misligholdene med faktiske. Den totale andelen mislighold av alle observasjonene er 0,2212. Derfor velger vi å bruke det som en estimert sannsynlighet for mislighold, og bruker det videre til å bestemme at modellen skal predikere mislighold for alle $p > 0,2212$.

	Faktisk	
	0	1
Predikert	0 11541	2798
	1 11823	3838

Tabell 4 - Predikert mislighold av bering

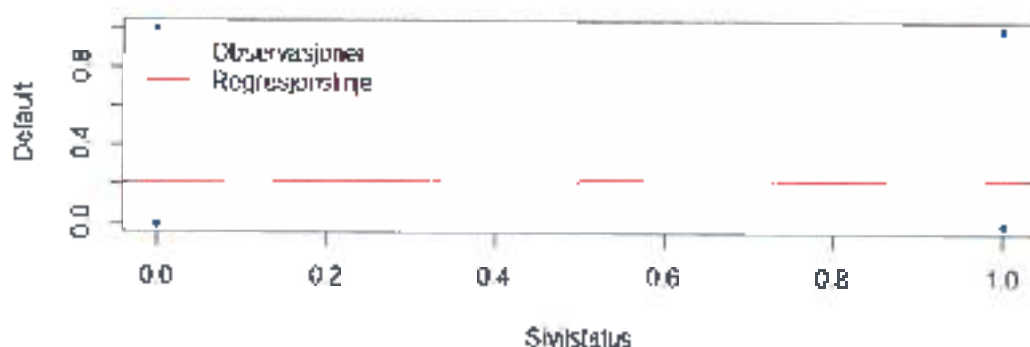
Tallene her forteller oss at ved å bruke regresjonsfunksjonen vår og estimert verdi for p , får vi totalt 14621 av 30000 misprediksjoner. 11823 av disse ligger hvor funksjonen har predikert mislighold men de faktiske observasjonene viser ikke mislighold. Dersom vi øker verdi for p til 0,27, som er den høyeste verdien for p som funksjonen kan predikere går andelen misprediksjoner ned til 8224 av 30000. Dette forteller oss at regresjonsfunksjonen vår kan predikere opp til 27% sjanse for mislighold med de gitte variablene, med 72,59% sikkerhet for riktig prediksjon.

Videre har vi estimert regresjonsmodeller for hver enkelt forklaringsvariabel med mislighold som responsvariabel og plottet grafene for funksjonene inn i spredningsdiagram for hver enkelt variabel fra datasettet. Spredningen vises kun på fire punkter i alle plottene siden begge variablene i alle kategoriene er dummy-variabler, altså 1 eller 0. Derfor er det grafene for regresjonene som viser sammenhengene mellom sannsynlighet for mishold og variablene 1 og 0 på henholdsvis utdanning, sivilstatus og kjønn.



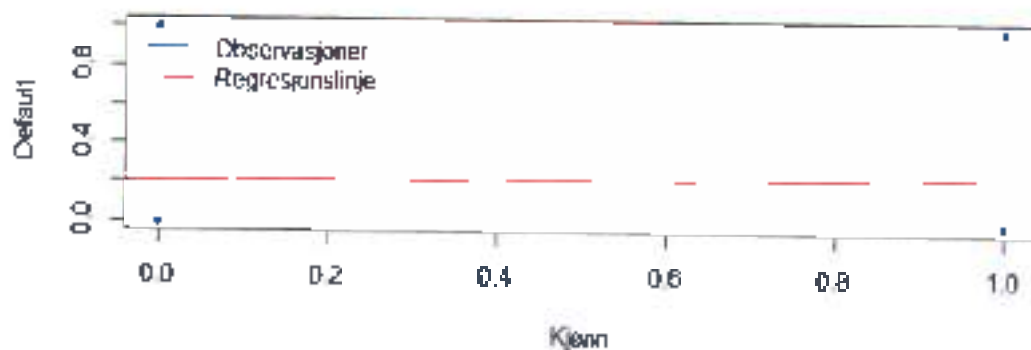
Figur 1 - Regresjonsmodell med variabel Utdanning

Det første diagrammet viser en svak sammenheng mellom oppnådd "graduate"-utdanning og sjanse for å misligholde. Grafen heller svakt mot høyre og forteller oss at en med den utdanningen har noe lavere sjanse for å misligholde.



Figur 2 - Regresjonsmodell med variabel Sivilstatus

Det andre diagrammet viser sammenhengen mellom sivilstatus og sjansen for å misligholde. Grafen stiger svakt mot høyre og forteller oss at en gift person har noe større sjanse for å misligholde enn en ugift.



Figur 3 - Regresjonsmodell med variabel kjønn

Det tredje diagrammet viser sammenhengen mellom kjønn og sjansen for å misligholde. At grafen stiger svakt mot høyre forteller oss at en mann har noe større sjanse for å misligholde enn en kvinne.

Spørsmål 3

Oppgaven forutsetter at bedriften som skal gis rådgivning ikke har noen måte å observere nye kunders betalingshistorikk. Dermed bortfaller alle kategorier i datasettet utenom de direkte observerbare hos et tilfeldig forbigående individ. Det kan være problematisk for selgere å anta sivilstatus og utdannelsesnivå til potensielle kunder på gaten. Sivilstatus kan identifiseres hvis potensielle brukere går sammen som par, men dette varierer og vil ikke alltid være tilfellet. Vi ønsker derfor å velge en sikrere antakelse og holder oss til alder og kjønn da disse kan vurderes raskt av en selger på gaten. De andre kategoriene i datasettet vil dermed falle bort.

Vi utfører videre en regresjon for å anslå hvilken aldersgruppe av ulike kjønn som vil være mest aktuelle for en selger å henvende seg til.

Dependent variable:	
DEFAULT	
Age	0.00047* (0.00033)
Sex	0.033*** (0.0047)
Constant	0.192*** (0.010)
Observations	30,000
Log Likelihood	-16.163.250
Akaike Inf. Crit.	32,332.510

Note: *p<0.1; **p<0.05; ***p<0.01

Tabell 5 - Lineær regresjon

På bunnen av regresjonen ser vi p-verdiene med ulik antall stjerner (*). Når man gjennomfører en hypotesetest i statistikk bruker man p-verdi for å anslå signifikansen til resultatet. Desto mer signifikant et resultat i regresjonen er, desto større betydning har verdien for resultatet av modellen. Hypotesetester brukes for å teste validitet av en såkalt nullhypotese. En liten p-verdi på under 0,05 indikerer at vi kan vurdere å kaste nullhypotesen og tallet er dermed ikke signifikant. P-verdier større enn 0,05 indikerer at det er lite bevis mot nullhypotesen og den kan da ikke forkastes.

Ved å estimere en lineær regresjon med mislighold som responsvariabel, og kjønn og alder som responsvariabel får vi følgende funksjon for sjansen for mislighold:

$$Y = 0.192(\text{Konstant}) + 0.033(\text{Hvis mann}) + 0.00047(\text{Alder})$$

Alderen som variabel har et konfidensnivå på 92,58%, noe som betyr at alderen er av relativ lav statistisk signifikant betydning. Derfor velger vi å dele inn i aldersgrupper hvor 25 representerer 21-30 år, 35 for 31-40 år osv. Denne oppdelingen av alder vil også gjøre det lettere for selgere å vurdere alder, siden det er lettere å vurdere alderen for fremmede i tiår enn spesifikk alder. Variabelen for kjønn er derimot mer signifikant og vi oppretter derfor kolonner for begge kjønn i aldersgruppene. I tabellen under har vi brukt regresjonsfunksjonen til å predikere sannsynligheten for mislighold for menn og kvinner i alderskategoriene.

Alder	p Kvinne	p Mann
21-30	0,2038	0,2368
31-40	0,2085	0,2415
41-50	0,2132	0,2442
51-60	0,2179	0,2509
61-70	0,2226	0,2556
71-80	0,2273	0,2603

Tabel 6 - Sannsynlighet for misligholdelse - Alder og kjønn

Ut fra tabellen kan man se at regresjonen gir en sammenheng mellom alder hvor høyere alder gir større sjanse for mislighold. Koeffisienten for menn gir og større sjanse for mislighold. Konklusjonen som kan presenteres for selgerne er dermed at de burde ha størst fokus på unge kvinner blant de forbipasserende kundene når de forutsetter at lavest mulig sjanse for mislighold maksimerer profitt, men alle forbipasserende unntatt menn over 50 år har under 25% sjanse for å misligholde ifølge funksjonen.

Spørsmål 4

Etter å ha analysert datasettene med oversikt over egenskaper til 30 000 kredittkorkunder har vi presentert hvilke grupper som har lavest sannsynlighet for å misligholde kredittlånet. Banker tjener den største andelen av inntekten fra gebyrer på omtrent 2-3% av beløpet det handles for (Investopedia, 2018). Derfor er det fordelaktig for banken at kredittkortet blir brukt jevnlig over tid i stedet for at lånet misligholdes som til slutt kan ende opp med at banken ikke får betalt. Likevel er det et kjent fenomen at banker som oftest vil oppnå høyere profitt av at kredittkundene betaler høye gebyr på sene betalinger og vil bevisst forsøke å selge til personer som gjør feilaktige finansielle avgjørelser (Ru & Schoar, 2016).

Strategien for denne fremgangsmåten er ofte å lokke kunder med lave satser som raskt øker med høye avgifter for sen betaling. Problemet med dette ifølge senere forskning er at personer med høyere utdanning ofte ikke lar i bruk slike løsninger (National Bureau of Economic Research, 2016), som vi kan se av datasettet i at 19,23% misligholder lånet. Dette er mindre enn de andre utdannings variablene (bortsett fra "annen utdanning" da denne er lite signifikant). Dette er et viktig moment mot videre salgstrategi da hele 82,05% av personene i datasettet har universitetsutdanning eller høyere.

I datasettet er det et stort antall personer som ikke kan betale neste måned, nærmere 22,12% av de 30 000 personene i utvalget. Dette er svært høyt sammenlignet med for eksempel i USA der prosentandelen av de med kredittkort som misligholder lånet ligger mellom 2-10%, der alt over 5% regnes som bekymringsverdig (Federal Reserve, 2016). Dette kan tyde på at en kredittkrise er på vei, noe som må tas stilling til.

Det er dermed både etisk og økonomisk forsvarlig å rette salgskampanjen mot grupper som har størst sannsynlighet for å betale kredittgjelden sin. I forhold til profittmaksimering vil dette da på lengre sikt være fordelaktig da svært mange ikke klarer å betale kredittgjelden sin.

Fra resultatet av vår analyse vil vi anbefale å rette fremtidig salgsstrategi mot kvinner i aldersgruppen 21-30 år som har en misligholds sannsynlighet på 20,38%. Det kan også være en fordel å sikte mot personer med høyere utdanning da disse har tilsynelatende mindre sannsynlighet for misligholdelse. Fra regresjonsfunksjonen i oppgave 3 kom det fram at kundegruppen med størst sjanse for mislighold basert på kjønet og alder er menn over 50 år. Det er likevel problematisk at det også er denne aldersgruppen som er desidert minst representert i datasettet, og det kan dermed svekke tilliten til modellen når det gjelder koeffisienten for alder. Dette er også den naturlige forklaringen på det relativt lave konfidensnivået for alderskoeffisienten.

For å få en mer grundig prediksjon burde det bli tatt hensyn til flere faktorer i analysen. Modellen vi har brukt har ikke tatt hensyn til betalingshistorikk, for eksempel om kunden har overholdt tidligere betalingsfrister eller hvor mye kunden har skyldt tidligere. Hvis man estimerer en multipl regressjonsmodell med alle kategoriene fra datasettet ser man at det er flere kategorier som har signifikant statistisk betydning for mislighold som ikke har blitt tatt hensyn til i regresjon fra oppgave 2 og 3. For å kunne ta i bruk en slik modell med mer nøyaktighet anbefaler vi direktøren i selskapet å ta i bruk et register med betalingshistorikk for potensielle kunder.

Andre begrensninger for analysen er at datasettet inneholder for få observasjoner til å trekke klare konklusjoner. For eksempel i aldersgruppene for de som er 60 år og eldre og utdanningsnivå av type "annet" er det for få observasjoner til å kunne anta at sannsynlighetene har validitet.

Bibliografi

- Federal Reserve. (2016). *Credit Card Debt Study: Trends & Insight*. (A. Comoreanu, Red.) Federal Reserve
- Forbes. (2017). *Why 43% Of Adults Have Carried Credit Card Debt For More Than 2 Years*.
- Investopedia. (2018). *Credit Card Debt*.
- Keller, G. (2012). *Managerial Statistics* (9. utg.). South-Western.
- National Bureau of Economic Research. (2016). *Do Credit Card Companies Screen for Behavioral Biases?* NBER.
- Ru, H., & Schuar, A. (2016).
- Seven Pillars Institute. (2018). *Taiwan's Credit Card Crisis*.

Bedømmelse etter vurderingsskjema for hjemmeeksamen:

Presentasjon av tabeller og figurer:	1
Valg av metode:	2
Anvendelse av metoder:	1.5
Diskusjon:	3
Etikk:	3
Formalfeil:	0
 Totalscore:	 2.125/3

Intern sensor gjorde følgende notater under første gjennomlesning:

Fancy forside ihvertfall. Flott innledning, har hentet inn referanser om kontekst og diskuterer etikk allerede her. Gjør noen valg om datarensing som er begrunnet og forklart. Rapporterer gjennomsnitt av utdanning og sivilstatus, som blir litt rart, men tar også med median og typetall. Rapporterer misligholdsgrad hos en del grupper. Finn innledning til logistisk regresjon. Lager dummyvariable og rapporterer regresjonstabell. Oversetter til sannsynlighet. Lager faktisk en ad-hoc klassifiseringsregel, og tester den (in-sample?) og rapporterer confusionmatrise. Artig påfunn, men det blitt litt hjemmesnekret. Lager også noen litt spesielle figurer med enkel regresjon. Er det OLS? Det var ikke så lett å finne ut hva det skulle bety. Oppgave 3: Ser bort fra utdanning (greit nok), kjører alder direkte så får ikke med seg ikke-lineariteten her. Men så: deler opp i grupper. Får så ut en grov tabell som utelukker menn over 50 år. Redder seg greit inn, men ikke helt overbevisende. Oppgave 4: Fine betraktninger. Referanser. Etikk. Alt i alt er selve metodebruken og diskusjonen rundt den ok, men lite spennende. Pluss for å sette ting fint inn i kontekst

Intern sensor foreslo følgende karakter for denne besvarelsen:

B

Denne besvarelsen ble bedømt til B ved endelig sensur.