



# Hjemmeeksamen

Kandidatnummer:

Norges Handelshøyskole  
MET4 – Vår 2018

Antall sider: 10

## EKSEMPELBESVARELSE MET4 HJEMMEEKSAMEN V18

Denne besvarelsen ble levert som hjemmeeksamen i faget MET4 vårsemesteret 2018, og er publisert i anonym form i samråd med den aktuelle studentgruppen og eksamenskontoret ved NHH.

Se eget vedlegg for sensors kommentarer til denne oppgaven.

## Innledning

Denne rapporten har som formål å belyse sannsynligheten for mislighold av kredittavtale for en bedrift fra Taiwan. Dette vil bli gjort gjennom enkle modeller basert på utlevert datasett. Innledningsvis i rapporten presenteres deskriptiv statistikk av datasettet, da dette presenterer nødvendig informasjon om elementer som brukes gjennomgående i oppgaven.

## Oppgave 1

Det eksisterende datasettet inneholdt fire demografiske variabler; kjønn, alder, sivilstatus og utdanning; samt informasjon om betalingshistorikken for eksisterende kunder. I tillegg har banken gjort en vurdering for hver kunde på om denne vil havne i mislighold ved neste fakturering. Vi har valgt å tolke denne variabelen som at kunden enten er i mislighold eller ikke.

Selgerne har ikke tilgang på betalingshistorikk for potensielle nye kunder, så vi har valgt å fokusere på de variablene som omhandler det demografiske i vår analyse. Figur 1 har som formål å gi en rask oversikt over de nevnte variablene. For enkelhets skyld er de oversatt til norsk.

Tabell 1 Beskrivelse av variabler vi bruker for videre analyser (Yeh & Lien, 2009)

Navn	Kategori	Beskrivelse
Kjønn	Nominal	SEX (1=mann, 2=kvinn)
Utdanning	Ordinal	EDUCATION(1="graduate" utdanning, 2=universitet, 3=VGS, 4= annet)
Sivilstatus	Ordinal	MARRIAGE (1=gift, 2=sløst, 3=annet)
Alder	Ordinal	AGE:(25-30, 30-35, 35-40, 40-50, 50-60)
Mislighold av kredittkort-gjeld	Nominal	default payment next month(1=yes, 2=no)

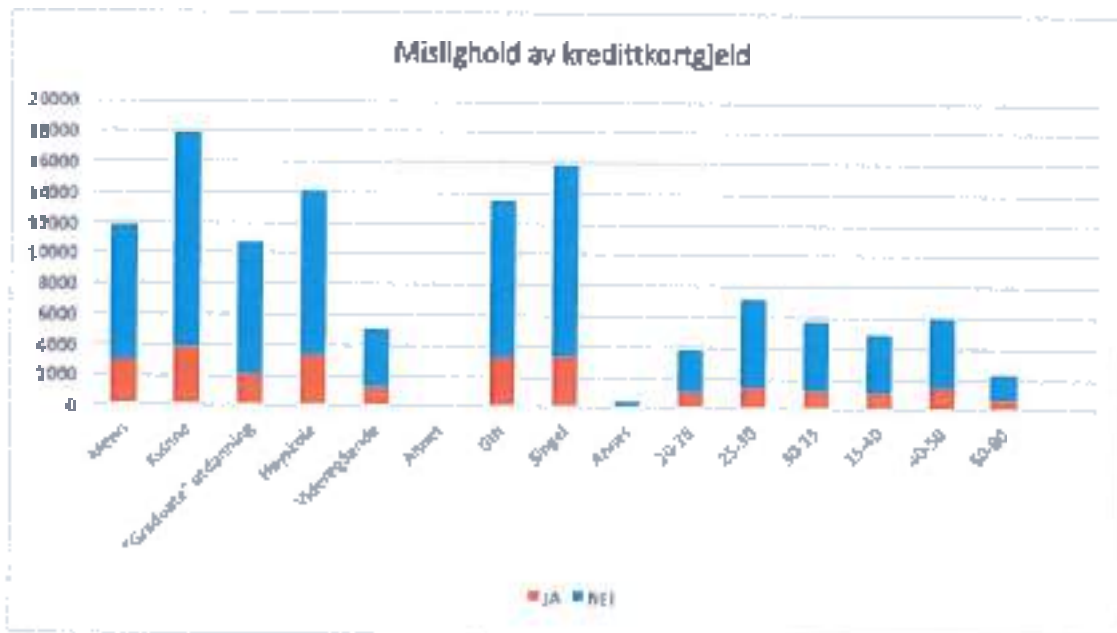
Mislighold av kredittkort er en nominal variabel. Den representerer prediksjonen for betalingsstatusen til en kunde utført med et kredittkort. Mislighold blir oppgavens responsvariabel og alle resultatene vil bli testet opp mot denne. Mislighold er også en dummy-variabel, som vil si en variabel som har to kategorisk mulige utfall. (Keller, 2009)

For hver variabel er det gitt en forklaring på hvilke kategorier en kunde kan havne i. For eksempel er variabelen kjønn delt i kategoriene 1 og 2, forklart med henholdsvis mann og kvinne. Ved gjennomgangen av datasettet fant vi noen kategorier som ikke samsvarer med disse forklaringene. Eksempelvis var enkelte kunder ført opp med "0" for kjønn. "0" dukket også opp i de andre variablene. I tillegg fant vi kategoriene "5" og "6" for utdanning, som kun hadde forklaringer for "1"~"4". Vi vurderte dataene for kunder med slike oppføringer som ugyldige, og fjernet dem fra datasettet for å sikre tilstrekkelig god datakvalitet.

### Deskriptiv statistikk av datasettet

Tabell 2 Sammenheng av de statistiske nominale og ordinale variablene

		Mislighold av kredittkort	
		NEI	JA
Kjønn	Mann	8885	2861
	Kvinne	14111	3744
Utdanning	"Graduate"	8545	2036
	Universitet	10695	3329
	VGS	3640	1233
	Annet	116	7
Sivilstatus	Gift	10285	3192
	Singel	12477	3379
	Annet	235	84
Alder	20-25	2799	1028
	25-30	9636	1436
	30-35	4597	1123
	35-40	3783	1059
	40-50	4521	1385
	50-80	1658	574



Figur 1 Spylvediagram for mislighold av kredittkortgjeld

Tabell 2 og figur 1 viser sammendraget av variablene på bakgrunn i den informasjonen vi hentet ut av R. Histogrammet er med for en mer visuell fremvisning av datasettet. Tabell 2 viser oss at hoveddelen av kredittkortbrukere er kvinner. Selv om kvinner er utgjør en større del av det totale utvalget, viser figur 1 at menn har høyere en relativt høyere andel som er vurdert til mislighold. I tillegg ser vi at flere gifte enn single misligholder kredittavtalen.

## Oppgave 2

I denne oppgaven skal vi estimere regresjonsmodeller som vi bruker som grunnlag i drøftelsen om hvorvidt forklaringsvariablene forklarer sannsynligheten for mislighold.

For å estimere en regresjonsmodell for mislighold, benytter vi R studios.

Forklaringsvariablene vi bruker er som nevnt definerte som kategorivariabler. Altså variabler som identifiserer en gitt gruppe der variabelen hører til. For eksempel vil en person i datasettet vårt som har verdien "1" som utdanning ha oppnådd "graduate"-utdanning. (StataCorp, 2014) Da vi skulle lage modellen trengte vi å isolere radene i datasettet som korresponderte med oppgavens spesifikasjoner. Vi trengte dermed alle som var gift og hadde "graduate"-utdanning, uavhengig av kjønn. For å oppnå ønsket resultat så vi det som enklest,

og tryggest, å utføre dette i Excel. Datasettet ble kalt `data.df` og lastet opp i R studios. For enkelhets skyld lagret vi en variabel kalt "Y" for 'default payment next month'.

Vi så det som mest forklarende å lage to forskjellige regresjonsmodeller for å besvare oppgaven. En logit- og en probit-modell. Grunnen til at vi bruker disse er at vi har en dummy-variabel som responsvariabel. Å bruke en lineær regresjon ville ikke vært hensiktsmessig, da vi ville fått verdier som gikk utenfor intervallet  $[0,1]$ . Vi trenger da en skårfunksjon, eller en indeks, som kan anta alle verdier i tallinjen vår. (Otneim, 2018) En slik funksjon på generell form ser slik ut:

$$Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_r X_r$$

Videre trenger vi en funksjon  $F(Z)$  som kan omgjøre Z-verdier, slik at en høy Z vil vise en høy sannsynlighet for å misligholde kredittkorgjelden sin (1), og omvendt. Det er her våre to regresjonsmodeller kommer inn. Modellene er relativt like, men skilles ved måten de definerer  $F(Z)$ . Probit-modellen har opprinnelse fra normalfordeling, mens logit-modellen har opphav i den logiske fordelingen. (Albright, 2015)

Videre bruker vi R studios for å estimere regresjonsmodellene. Resultatet er vist i Tabell 3.

Tabell 3 Regresjonsanalyser

REGRESJONSANALYSE		
	Dependent variable:	
	Y	
	logistic (1)	probit (2)
SEX2	-0.21*** (0.03)	-0.12*** (0.02)
EDUCATION1	-0.23*** (0.03)	-0.15*** (0.02)
MARRIAGE1	0.11*** (0.03)	0.07*** (0.02)
Constant	-1.11*** (0.03)	-0.68*** (0.02)
Observations	30,000	30,000
Log Likelihood	-15,778.97	-15,779.12
Akaike Inf. Crit.	31,565.95	31,566.25
Note:	*p<0.1, **p<0.05, ***p<0.01	

Før å starte drøftingen om hvorvidt forklaringsvariablene forklarer sannsynligheten for mislighold kan vi trekke fram p-verdiene. P-verdiene vi leser ut av tabellen viser sannsynligheten for å få resultatet vi får, i et utvalg der forklaringsvariablene ikke har noen effekt. (Princeton University Library, 2007) Vi ser nederst i tabell 1 at p-verdiene er kategorisert og vises bak koeffisientene som antall stjerner(\*). Tre stjerner viser til en p-verdi under 1%, med andre ord viser det et høyt signifikansnivå. (Keller, 2009) Selv om dette viser med høy sannsynlighet at koeffisientene påvirker responsvariabelen, er det derimot ikke en begrunnelse for omfanget av effekten. Vi kan altså ikke utelukke at det er andre variabler som påvirker responsvariabelen til tross for lav p-verdi. Noe som kan være nærliggende å tenke i dette tilfellet.

Modellen viser, som nevnt, at en lav verdi for Y vil gi en lavere sannsynlighet for å misligholde kredittlån. Koeffisientene til forklaringsvariablene kan gi oss et bilde på hvem som har lavest sannsynlighet for mislighold. Det vises i begge regresjonsmodellene at kvinner har mindre sannsynlighet for mislighold. Videre kan vi se at det å ha "graduate"-utdanning, ikke overraskende, gir lavere sannsynlighet for mislighold. Det motsatte gjelder om personen er gift, altså at gifte personer har høyere sannsynlighet for mislighold.

Under hver koeffisient kan vi lese av et tall i en parentes. Dette er standardavviket til koeffisienten. På bakgrunn av at vi har 30000 observasjoner vil denne naturlig bli veldig lav.

Avslutningsvis kan vi konkludere med at forklaringsvariablene tilsynelatende forklarer sannsynligheten for mislighold på en god måte, men at det er flere variabler som også spiller inn. Forklaringsvariablene forklarer noe av sannsynligheten, men det er ikke grunnlag for å hevde at de gir en fullstendig forklaring.

### Oppgave 3

Denne oppgaven går ut på å analysere profittpotensialet til forbipasserende, slik at selgerne i kredittkortselskapet vet hvem de bør henvende seg til. For å gjøre dette har vi estimert en

regresjonsmodell, og brukt denne til å predikere ulike misligholdssannsynligheter basert på ulike variabler.

Først ønsket vi å finne ut hvilke variabler som faktisk påvirker misligholdssannsynligheten. Variabelen for mislighold er som nevnt en dummy-variabel. Derfor satte vi opp enkle, logistiske regresjoner i R studios med hver av de demografiske variablene som forklaringsvariabler og misligholdsvariabelen som responsvariabel. (Ouncim, 2018)

Alle kategoriene innen kjønn og utdanning viste seg å ha lave p-verdier og derfor en signifikant påvirkning. Utdanningskategorien "annet" har niktignok et mye mindre utvalg og er en mindre praktisk kategori uten en klar definisjon, men den kan omfatte potensielle kunder med en utdanning over VGS, som ikke er akademisk. Siden den var signifikant har vi derfor valgt å ta den med. Disse variablene vurderte vi som relevante for modellen vår.

For sivilstatus fant vi at kategorien "3", for "annet", ikke hadde signifikant påvirkning. Med denne kategorien ville modellen vår blitt svakere. Det hadde ikke vært mulig å fjerne den alene, eller innlemme den i en av de andre kategoriene, "singel" eller "gift". I praksis ville man da ikke være i stand til å benytte modellen for kunder som verken var gifte eller single. Vi valgte derfor å ikke benytte oss av variabelen sivilstatus i modellen.

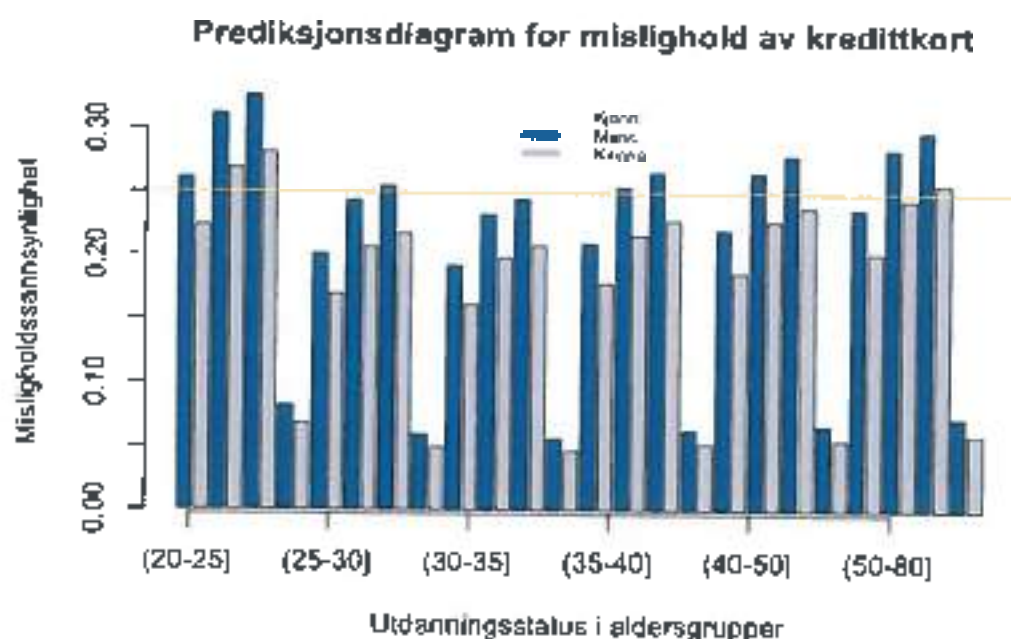
Variabelen for alder var i utgangspunktet ikke kategorisert. Hver kunde var ført opp med sin eksakte alder. Vi kom fram til at man ikke kan anta en lineær sammenheng mellom høyere alder og høyere eller lavere forekomst av mislighold. Av den grunn valgte vi å dele opp "alder" i intervaller. Ved å kjøre regresjonen med forskjellige grupperinger fant vi det mest hensiktsmessig med nokså korte intervaller på fem år for kundene mellom 20 og 40. Dette henger sammen med at de utgjorde majoriteten av utvalget. Vi valgte å plassere alle kundene over 50 i samme intervall, da de ikke viste noen signifikant påvirkning ved mindre oppdelinger. Til slutt endte vi opp med seks alderskategorier som vi inkluderte i modellen vår.

For å vise hvor stor del av den totale påvirkningen av responsvariabelen som skyldes hver enkelt forklaringsvariabel, satte vi dem inn i en multiplert regresjonsmodell med mislighold som responsvariabel. (Statsoft, 2012). For å luke ut eventuelle sammenhenger mellom de ulike forklaringsvariablene, foretok vi en test av multikollinearitet med formelen  $VIF()$  i R

studios, (San Diego State University, 2013). Testen utfører en lineær regresjon for hver variabel opp mot de øvrige variablene. Denne viste lave VIF-verdier (under 5) på mellom 1.01 og 1.09 og dermed at variablene ikke påvirket hverandre i betydelig grad. Modellen var altså av god kvalitet. Da vi kjørte regresjonen, viste den lave p-verdier og betydelige koeffisienter. Den viste med andre ord at variablene påvirket misligholdsforekomsten uavhengig av hverandre, og at påvirkningen var signifikant for alle variablene.

Regresjonsmodellen i seg selv gir ikke noe tall på sannsynligheten for mislighold, men den kan brukes videre til å predikere hvilke kunder som kommer til å misligholde sitt kredittkort, og hvem som ikke kommer til å gjøre det (Omeim, 2018). Det vil si at vi substituerer for variablene i regresjonsmodellen med de virkelige kategoriene fra datasettet. Ved å utføre en slik prediksjon i R studios, fikk vi tall for hver kombinasjon av variablene som viser sannsynligheten for at en kunde med en gitt kombinasjon vil misligholde kredittavtalen.

Ved å sette sannsynlighetene inn i et diagram, får vi følgende figur:



Figur 2 Søylediagram for misligholdssannsynligheter

Hver søyle representerer en mulig kombinasjon av variabler. For hvert aldersintervall ser vi fire utdanningsnivåer for hvert kjønn fra venstre til høyre, henholdsvis graduate, universitet, VGS og annet. Den gule linjen illustrerer grensen på 25 % misligholdssannsynlighet.



Grovt sett ser vi at den yngste gruppen har den høyeste sannsynligheten, at menn jevnt over har noe høyere misligholdssannsynlighet enn kvinner, og en tendens til at høyere utdanning gir lavere sannsynlighet for mislighold.

Det mest påfallende er unntaket med svært liten sannsynlighet for at en person med utdanning "annet" vil misligholde kredittavtalen. Dette ser ut til å komme av at gruppen har et mindre utvalg, og at det derfor er vanskeligere å predikere en sikker sannsynlighet. 95 %-konfidensintervallet i tabellen under illustrerer dette. Et bredt konfidensintervall vil si at det vil være en større feilmargin for den predikerte sannsynligheten, og motsatt om det er smalt (Store norske leksikon, 2018). For vår del vil det si at man ikke kan være helt sikker på at sannsynligheten under "PredictedProb" er riktig, men at man med 95 % sikkerhet kan si at sannsynligheten ligger mellom den i "2.5%" og den i "97.5%". Vi ser at det er også bredt i tilfeller der utdanningsnivået er 4, altså "annet", for eksempel fra 0,02-0,12, mens det er mye lavere for 3, altså VGS, for eksempel fra 0,30-0,34. Graduate og universitet har tilsvarende lave konfidensintervaller. Dersom vi ser disse tallene i sammenheng med de lave p-verdiene vi fant i regresjonsmodellen, kan vi likevel si med nokså stor sikkerhet at personer i utdanningskategorien "annet" vil ligge under grenseverdien på 25 %. Gruppen er såpass liten at det ikke vil ha mye for seg å søke den ut aktivt, men man trenger da heller ikke forsøke å unngå den.

Tabell 4 Gruppene med flest sannsynlighet for mislighold

SEX	EDUCATION	AGE	PredictedProb	2.5%	97.5%
2	4	5	0,0549	0,0263	0,1111
1	4	6	0,0729	0,0351	0,1452
2	4	6	0,06	0,0287	0,1213
1	3	1	0,3254	0,3029	0,3487
2	3	1	0,2817	0,2622	0,3021

I praksis vil modellen fortelle selgerne at de ikke bør henvende seg til personer under 25 år, med mindre det er snakk om en kvinne med graduate-utdanning. De kan også stort sett trygt

henvende seg til personer mellom 25 og 35 år. Blant disse er det bare menn mellom 25 og 30 år med videregående utdanning som har litt for høy misligholdssannsynlighet. Utover dette vil menn med høy utdanning og alle kvinner være potensielle kunder.

Av videre føringer vil vi anbefale å jevnt over prioritere kvinner over menn og høy utdanning over lav.

#### Oppgave 4

Notat til direktør:

Ut i fra forutsetningen om at lav misligholdssannsynlighet gir høy profit, vil vår modell kunne hjelpe dere mot målet om profittmaksimering. Analysen vår viser at variablene i datasettet har en signifikant påvirkning på misligholdssannsynligheten. Koeffisientene fra regresjonsmodellen er høye nok til å ha betydning for om en kunde kommer til å misligholde eller ikke.

Likevel vil vi ikke anbefale å bruke modellen som helhet. Modellen inneholder ikke sikkerhet nok til å foreta definitive økonomiske beslutninger. Selv om vi fant at variablene fra datasettet påvirket, var ikke denne påvirkningen stor nok til å si at ikke andre variabler kan være enda viktigere. Med mer eller annen data i modellen er det derfor ikke utenkelig at personer som nå ser ut til å ha for høy misligholdssannsynlighet til å bli tilbudt kredittkort, under andre omstendigheter kunne blitt vurdert som lønnsomme. Med en for bastant tolking av modellen, kan man derfor risikere å utelukke potensielt gode kundegrupper, eller omvendt.

På samme måte er det også mulig at flere eller andre variabler kunne vist en større forskjell mellom gruppene enn det vi kunne påvise. En slik modell ville gitt klarere retningslinjer og vært mye enklere for selgerne å forholde seg til.

Vi vil derfor anbefale at bedriften kan ta utgangspunkt i modellen for å prioritere hvem en selger skal henvende seg til, istedenfor å bevisst bruke den til å utelukke enkelte demografiske grupper.

## Bibliografi

- Keller, G. (2009). *Managerial Statistics* [8. utgave. utg.]. 45040, OH, USA: South-Western Cengage Learning
- Otneim, H. (2018, Mars 16). Forelesning 16 - Logistisk regresjon.
- Princeton University Library. (2007). Hentet fra [https://dss.princeton.edu/online\\_help/analysis/interpreting\\_regression.htm](https://dss.princeton.edu/online_help/analysis/interpreting_regression.htm)
- San Diego State University. (2013, November 2013). *Logistic Regression (R)*. Hentet fra [http://scg.sdsu.edu/logit\\_r/](http://scg.sdsu.edu/logit_r/)
- StataCorp. (2014, Desember 10). Hentet fra <https://www.stata.com/manuals13/u25.pdf>
- Statsoft. (2012). Hentet fra How To Find Relationship Between Variables, Multiple Regression: <http://www.statsoft.com/Textbook/Multiple-Regression>
- Store norske leksikon. (2018, Februar 20). *Konfidensintervall*. Hentet fra <https://snl.no/konfidensintervall>

### Bedømmelse etter vurderingsskjema for hjemmeeksamen:

---

Presentasjon av tabeller og figurer:	1
Valg av metode:	1.5
Anvendelse av metoder:	1.5
Diskusjon:	2
Etikk:	0
Formalfeil:	0
 Totalscore:	 1.525/3

### Intern sensor gjorde følgende notater under første gjennomlesning:

---

Ok start, gjør et fornuftig variabelutplukk, oversetter alder til kategorier. Bruker mye plass på å fordele observasjoner i kategorier. Oppgave 1 fokuseres utelukkende på mislighold som en responsvariabel, som kanskje er greit nok, okey tabell- og figurbruk. Litt (unødnevndige) detaljer om hva slags programvare som er brukt, forklarer litt om logistisk regresjon. Viser logit og probit. Helt ok diskusjon, men den blir aldri spesielt spennende. Oppgave 3: Tar aldersvariabelen fint, erkjenner ikke-linearitet og deler opp i kategorier. Merkelig referansebruk. Sjekker VIF. Godt forsøk på figur som viser predikert sannsynlighet opp mot 25%-grensen for kjønn, utdanningsnivå og alder, men den blir litt tett og vanskelig å lese. Henger seg litt opp i 'annet'-kategorien. Ikke så vellykket tabell som viser prediksjonsintervall for noen predikerte sannsynligheter. Konklusjonen er: 'Av videre føringer vil vi anbefale å jevnt over prioritere kvinner over menn og høy utdanning over lav.'. Oppgave 4: Nærmest intetsigende.

Intern sensor foreslo følgende karakter for denne besvarelsen:

C

Denne besvarelsen ble bedømt til C ved endelig sensur.