NHH

HJEMMEEKSAMEN
MET4

Vår, 2021

Start: 26. april 2021, 0900

Slutt: 28. april 2021, 1400

BESVARELSEN SKAL LEVERES I WISEFLOW

På våre nettsider finner du informasjon om hvordan du leverer din besvarelse:
https://www.nhh.no/for-studenter/eksamen/innlevering-individuelt-og-i-gruppe/

Kandidatnummer blir oppgitt på StudentWeb i god tid før innlevering.
Kandidatnummer skal være påført på alle sider øverst i høyre hjørne (ikke navn eller studentnummer). Ved gruppeinnlevering skal alle gruppemedlemmers kandidatnummer påføres.

UTFYLLENDE BESTEMMELSER OM EKSAMEN

https://www.nhh.no/globalassets/for-studenter/forskrifter/utfyllende-bestemmelser-til-forskrift-om-fulltidsstudiene-ved-nhh.pdf

Antall sider, inkludert forside: 6

Antall vedlegg: 1 (met4_v21.RData)

FOREBYGGENDE MARKEDSFØRING

INSTITUTT FOR FORETAKSØKONOMI

NHH

FRA 09:00 26 APRIL 2021 TIL 14:00 28 APRIL 2021

INNLEDNING

I privatmarkedet for forsikringer har det stor verdi å sikre langsiktige kundeforhold. Når en kunde sier opp kundeforholdet sitt kalles dette «churn». Det kan være vanskelig å hente tilbake kunder som allerede har sagt opp forsikringsavtalene sine. Hvis det er mulig å predikere hvem av kundene som kommer til å si opp forsikringsavtalene sine i fremtiden er det derimot flere tiltak som kan settes i verk for å hindre churn – eksempelvis å kontakte kunden for en gjennomgang av forsikringsavtalene for å sikre at avtalene dekker kundens behov. Å kontakte kunder for en forsikringsgjennomgang har derimot en kostnad, og hvorvidt det er et godt preventivt tiltak er derfor avhengig av hvor presist vi klarer å predikere churn.

Tryg er et av de største forsikringsselskapene i Norden. I denne oppgaven skal dere jobbe med et pseudonymisert datasett fra Tryg, som dekker et utvalg av privatkunder i Norge. Det sentrale i oppgaven er å predikere churn, som i datasettet er målt med variabelen «CHURN_30». Denne er lik 1 dersom kunden sa opp forsikringsavtalen etter 30 dager, og 0 ellers. De øvrige variablene i datasettet måler størrelser relatert til kunden, produktdekning, premier og kontakthistorikk med Tryg.

Ved utvikling av prediksjonsmodeller er det typisk slik at man har bedre prediksjoner på datasettet som er brukt til å estimere modellen, sammenliknet med et «usett» datasett. Av den grunn er observasjonene delt tilfeldig opp i to deler: Treningsdatasettet «`df_train`» består av 80% av kundene, mens testdatasettet «`df_test`» består av de resterende 20%. Treningsdatasettet skal brukes til både deskriptiv statistikk og modellutvikling, og testdatasettet skal brukes til å evaluere prediksjonsegenskapene til modellen vi har bygget.

Den første del av eksamenen dreier seg om å bli kjent med datasettet og vurdere hvilke variabler vi inkluderer i prediksjonsmodellen, og eventuelt om det er nødvendig med noen transformasjoner. Dette kalles gjerne «feature engineering». Deretter skal dere utvikle en prediksjonsmodell. Til sist skal dere vurdere lønnsomheten i å bruke en prediksjonsmodell som del av et preventivt tiltak mot churn, og oppsummere resultatene i en anbefaling om hvordan modellen kan brukes.

Modellene vi anvender i denne eksamenen vil kunne brukes til å predikere *sannsynligheten* for churn for kunde *i*, p_i. Gitt at churn er både vanskelig å predikere og at churn er forholdsvis sjeldent vil typisk de fleste predikerte churnsannsynligheter være lave, og godt under 50%. For å anvende modellen trenger man da en terskelverdi p^*, der vi ringer til kunde *i* dersom predikert churnsannsynlighet er større enn terskelverdien (altså $\hat{p}_i \geq p^*$). Terskelverdien vil da være en konstant mellom 0 og 1. Dere vil analysere denne nærmere i oppgave 4.

Det finnes mange metoder for å måle kvaliteten på prediksjonsmodeller. I denne oppgaven skal vi bruke Trygs gevinst ved å anvende modellen. Anta at kostnaden ved å ringe til en kunde er $c = 0.025$, og at gevinsten ved å ha en kunde er 1. For hver kunde er det fire mulige utfall, der Trygs gevinst er oppført i kolonnen «Gevinst». Her forutsetter vi altså at å ringe til kunden er et 100% effektivt tiltak for å hindre churn. I virkeligheten finnes det flere mulige utfall (eksempelvis at vi ringer til en kunde som er på vei til å si opp, og kunden sier opp avtalen likevel), men vi skal kun vurdere utfallene i tabellen under i denne oppgaven.

PREDIKERT UTFALL	HANDLING	UTFALL	GEVINST
Kunden kommer til å si opp	Kunden ringes opp	Kunden blir værende	$1 - c$
Kunden kommer til å si opp	Kunden ringes ikke opp	Kunden sier opp	0
Kunden kommer til å bli værende	Kunden ringes opp	Kunden blir værende	$1 - c$
Kunden kommer til å bli værende	Kunden ringes ikke opp	Kunden blir værende	1

Det beste Tryg kan gjøre er å kun ringe til alle kundene som kommer til å si opp. I praksis vil neppe en prediksjonsmodell klare å identifisere alle kundene som kommer til å si opp, men jo bedre vi klarer å predikere hvem vi skal ringe til, jo høyere vil Trygs gevinst være.

EKSAMENSSPØRSMÅL

1. Presenter deskriptiv statistikk av datasettet. Fokuser på egenskaper ved datasettet som er relevant for resten av besvarelsen.
2. Variablene i datasettet er ulike med hensyn til både hva de måler og hvilke verdier de kan ha. Vurder om det er sterke korrelasjoner mellom variablene og om enkelte variabler bør transformeres (dummyvariabler, log-transformering etc) før disse brukes i en prediksjonsmodell, eller om de ikke bør brukes i det hele tatt. Forklar og begrunn valgene dere tar. *Merk: oppgave 1 og 2 bør til sammen ikke utgjøre mer enn 50% av besvarelsen. Prioriter hva dere gjør!*
3. Hvilke hensyn er viktige å ta når vi bygger en logistisk regresjonsmodell for å predikere[1] churn? Estimer minst tre ulike logistiske regresjonsmodeller ved hjelp av `df_train` der dere velger ulike sett med forklaringsvariabler og/eller transformerte forklaringsvariabler. Forklar og begrunn valgene dere tar. I neste oppgave skal dere bruke én av disse modellene for prediksjon.
4. Ved bruk av en prediksjonsmodell for churn vil man i praksis definere en terskelverdi p^*, og ringe til alle kunder med predikert churnsannsynlighet over denne terskelverdien. I denne oppgaven skal dere velge én av modellene som dere laget i oppgave 3, evaluere modellen på testdatasettet, og vurdere om vi forventer at det vil være lønnsomt å bruke modellen til å ringe kunder for å hindre churn. I tillegg har en analytiker i Tryg allerede estimert en prediksjonsmodell ved bruk av en avansert maskinlæringsalgoritme som vi også skal evaluere. Prediksjonene fra denne modellen finnes i `df_test` (se variabeloversikt).
    a. Skriv opp en formel for Trygs gevinst per kunde uttrykt ved terskelverdien p^*. Skriv også opp en formel for Trygs totale gevinst dersom de har *n* antall kunder.
    b. Trygs markedsavdeling vurderer tre ulike terskelverdier[2] p^* for når de skal ringe til en kunde for å forhindre churn: 0.02, 0.04 og 0.06. Beregn Trygs gevinst ved disse terskelverdiene for kundene i testdatasettet `df_test`, både for din valgte modell fra oppgave 3, og for Tryg-analytikeren sin modell. Presenter resultatene i en tabell eller en figur.
    c. Lag en kort oppsummering av problemstillingen der dere bruker en prediksjonsmodell til å velge hvilke kunder man skal ringe til for å hindre churn. Oppsummeringen skal svare på følgende:
        i. Hvilken kombinasjon av modell og terskelverdi er den beste?

[1] Når vi snakker vi om å «predikere» churn, svarer dette til å klassifisere om en kunde sier opp (1) eller ikke (0).
[2] Det er mer realistisk å vurdere *alle* mulige terskelverdier mellom 0 og 1, men det vil kreve programmeringsteknikker som ligger utenfor det vi har jobbet med i MET4.

ii. Er gevinsten ved å benytte prediksjonmodellen på kundene i testdatasettet *større* enn gevinsten ved ikke å ringe til noen kunder?
iii. Hva er din anbefaling? Bør Tryg bruke den beste prediksjonsmodellen i oppgave 4b til å velge ut hvilke kunder som skal kontaktes for å forhindre at de sier opp kundeforholdet?

ADMINISTRATIVE BESTEMMELSER

- Hjemmeeksamen i Met4 må leveres i grupper på 2, 3, eller 4 studenter.
- Det er ikke tillatt å diskutere eksamen med studenter utenfor din gruppe etter at oppgavesettet er frigitt.
- Besvarelsene vil bli rettet iht. rubrikk postet på Canvas.
- Dere kan besvare eksamen på norsk eller engelsk.
- Send en mail til både Ole-Petter Moe Hansen (s9705@nhh.no) og Håkon Otneim (hakon.otneim@nhh.no) ved spørsmål til oppgaven. Tilleggsinformasjon av betydning vil bli postet på Canvas. Merk: Ingen informasjon utover eksamensteksten vil bli gitt. Kun spørsmål vedrørende eventuelle feil i oppgaveteksten vil bli besvart.
- Rapporten skal ikke være lengre enn 10 sider. Tabeller, figurer og referanser er inkludert i de 10 sidene. Dersom rapporten har en forside uten noen form for svar på oppgavene kan forsiden komme i tillegg til de 10 sidene. Innholdsfortegnelse teller med i sidetallet, men er ikke nødvendig. Prioriter hva dere tar med i rapporten!
- Rapporten skal skrives med fonten Times New Roman, størrelse 12 og linjeavstand 1.15. Tekst i figurer og tabeller kan ha font ned til størrelse 9.
- Eksamen administreres i Wiseflow. Besvarelsen må leveres som en enkelt pdf-fil. Andre format (f.eks. .doc, .docx eller .R) er ikke akseptert.

DATASETT OG RÅD OM R

I datasettet «`df_train`» finnes følgende variabler:

VARIABELNAVN	FORKLARING
CHURN_30	Hvorvidt kunden sier opp innen 30 dager eller ikke. Lik 1 dersom kunden sier opp, 0 ellers.
CTL_TENURE_DAYS	Lengde på kundeforhold i dager.
CUSATTR_GENDER	Kjønn til forsikringstaker.
AGE	Kundens alder.
OBJ N_OBJ_TOT	Totalt antall objekter kunden har forsikret.
NET_PREM_TOT	Kundens totale nettopremie – dvs. den årlige avtalte premien kunden skal betale, inklusive rabatter.
TAR_PREM_TOT	Kundens totale tariffpremie – dvs årlig premie *før* avtalte rabatter.
N_OBJ_AU_400	Antall biler kunden har forsikret.
N_OBJ_FB	Antall fritidsbåter kunden har forsikret.
N_OBJ_PK_HUS	Antall hus kunden har forsikret.
N_OBJ_PK_INNBO	Antall innbodekninger kunden har.
N_OBJ_PK_FRHUS	Antall fritidshus (hytter) kunden har forsikret.

I datasettet «`df_test`» finnes de samme variablene som i «`df_train`», men i tillegg også følgende:

VARIABELNAVN	FORKLARING
PRED_GBM	Prediksjon, sannsynlighet for churn fra en avansert maskinlæringsalgoritme (se oppgave 4).

Dere importerer datasettet ved å kjøre kommandoen under, gitt at working directory i R er i samme mappe hvor filen `met4_v21.RData` ligger.

```
load("met4_v21.RData")
```

I denne oppgaven kan det vise seg nyttig å bruke logiske vektorer til å for eksempel finne ut hvilke verdier i en vektor som tilfredsstiller visse betingelser (for eksempel er større enn en terskelverdi). Her er noen kodelinjer som kan være nyttige:

```
# Vi lager en testvektor x som består av tallene 1 - 10:
test <- 1:10


# En logisk vektor som indikerer hvilke verdier i test som er større enn 6:
x1 <- test > 6


# Vi kan regne ut hvor mange TRUE vi har i en logisk vektor ved hjelp av
# sum():
sum(x1)


# De motsatte verdiene i en logisk vektor finner vi ved hjelp av "!":
!x1
```

```r
# Vi kan bruke "&" til å finne ut om tilsvarende verdier i to logiske
# vektorer begge er TRUE:
x2 <- test < 9
x3 <- x1 & x2

# Vi kan bruke en logisk vektor til å hente ut verdiene i test som svarer
# til TRUE. For eksempel:
test[x1]
test[!x1]
test[x1 & x2]
```