Eksamen SOS2901 våren 2023

Om eksamen

Eksamenssettet består av følgende oppgaver:

- 11 flervalgsoppgaver + 1 fyll inn tall
- 2 korte tekstsvar
- 1 sett oppgaver med flere deloppgaver knyttet til analyse av datasett

For *alle* flervalgsoppgaver får du 1 poeng hvis du svarer riktig, og minus 0.25 poeng hvis du svarer feil. Du får 0 poeng for oppgaver som er ubesvart.

Det er også noen andre typer oppgaver der du skal fylle inn et tall eller tekst. På disse er det ikke minuspoeng ved feil.

Tekstoppgaver har gjerne flere elementer og du bør passe på å besvare hver del da det påvirker poenggivningen på disse.

For oppgavene knyttet til datasettet er det viktig at du legger ved script med all R-kode. Her bare kopierer du fra Rstudio og limer inn i eget felt for kode. Du får ikke poeng for selve scriptet, men det er dokumentasjon for den jobben du har gjort. Hvis det mangler script får du ikke full uttelling på noen av oppgavene selv om svaret skulle være riktig. Ganske riktig kode kan i noen tilfeller gi delvis poeng selv om oppgitt svar er feil.
For å legge til rette for at sensor finner frem i scriptet er det fint om du lager en liten overskrift for den delen av scriptet med oppgavenummer slik:
Oppgave X

- Eksamensoppgaven er på norsk. Du kan besvare eksamenen på norsk, svensk, dansk eller engelsk.

Digital kandidatinstruks

Du finner kandidatinstruks for skoleeksamen som en ekstern ressurs i prøven.

Hjelpemidler

Alle R-scirpt som er blitt brukt i kurset.

Spørsmål under eksamen

Har du spørsmål under eksamen, ta kontakt med eksamensvaktene.

Innsyn i egen eksamensbesvarelse

Flersvalgsspørsmålene på eksamen skal kunne gjenbrukes, og vil derfor ikke bli tilgjengelige i etterkant av eksamen. Dersom du ønsker innsyn i langsvarsoppgaven, kan du få en kopi av denne ved å sende en e-post fra din UiO-epost til SV-info.

13 I maskinlæring vil man ofte justere modellene med ulike parametere eller vekting. Beskriv kort hvilke forhold man bør ta stilling til for å justerer en prediksjonsmodell på denne måten.

Skriv ditt svar her

Maks poeng: 4

14 Beskriv kort hva som er de viktigste forskjeller og likheter mellom random forest og gradient boosting.

Skriv ditt svar her

Maks poeng: 4

Last ned datasettet fra denne lenken: techHealth og les den inn i R. Filen er i .rds format.

Når slike filer lastes ned fra eksamenssystemet Inspera kan det skje ting med filnavnet. Du kan gjerne endre navnet på filen hvis du vil. Det kan hende at filhalen i navnet ".rds" blir borte når det lastes ned. Innlesning skal gå likevel så lenge filnavnet du skriver i R er korrekt.

Datasettet inneholder svar på spørreskjema fra 1251 ansatte i teknologibedrifter i 2014.

Utfallsvariabelen som skal brukes i alle prediksjonene med dette datasettet er "treatment", som er om arbeidstakerne har søkt helsehjelp for mentale helseproblemer. Bruk ellers alle variable i datasettet som prediktorer.

Her er hele variabellisten:

Age
Gender
self_employed: Are you self-employed?
family_history: Do you have a family history of mental illness?
treatment: Have you sought treatment for a mental health condition?
work_interfere: If you have a mental health condition, do you feel that it interferes with your work?
no_employees: How many employees does your company or organization have?
remote_work: Do you work remotely (outside of an office) at least 50% of the time?
tech_company: Is your employer primarily a tech company/organization?
benefits: Does your employer provide mental health benefits?
care_options: Do you know the options for mental health care your employer provides?
wellness_program: Has your employer ever discussed mental health as part of an employee wellness program?
seek_help: Does your employer provide resources to learn more about mental health issues and how to seek help?
anonymity: Is your anonymity protected if you choose to take advantage of mental health or substance abuse treatment resources?
leave: How easy is it for you to take medical leave for a mental health condition?
mentalhealthconsequence: Do you think that discussing a mental health issue with your employer would have negative consequences?
physhealthconsequence: Do you think that discussing a physical health issue with your employer would have negative consequences?
coworkers: Would you be willing to discuss a mental health issue with your coworkers?
physhealthinterview: Would you bring up a physical health issue with a potential employer in an interview?
mentalvsphysical: Do you feel that your employer takes mental health as seriously as physical health?
obs_consequence: Have you heard of or observed negative consequences for coworkers with mental health conditions in your workplace?

15 Fra bedriftens side er det to motstridende hensyn: på den ene siden ønsker de tilby hjelp til de som trenger det. På den andre siden er det kostnader knyttet til slik hjelp som de ønsker minimere. De ønsker derfor færrest mulig falske positive. Det anslås at hvis maksimalt 10% at de som får tilbud om tiltak er falske positive, så er det akseptabelt.

Bruk random forest til å løse oppgaven. Selv om det kan argumenteres for at man bør splitte datasettet, så bestemmer du deg for å bruke hele datasettet. I først omgang gjør du ingen justeringer av modellen og bruker forhåndsvalgene på alle parametre.

Bruk set.seed(42) og bsvar følgende spørsmål om resultatet:

Hva er OOB estimatet på total feil?

Sett inn tall her i prosent: [ ] %.

Maks poeng: 1

16 Se på confusion matrix for å vurdere resultatene. Vil denne modellen tilfredsstille de feilratene bedriften ønsker oppnå?

Velg ett alternativ:

- Ja
- Nei

Maks poeng: 1

17 Hvor stor andel av de som er predikert positive er falske positive?

Skriv inn tallet i prosent: [ ] %.

Maks poeng: 1

18 Du prøver igjen og tilpasser en ny modell. Du skal nå justere modellen for å endre resultatet i ønsket retning. Gjør dette ved å bruke tuning parameter for hvordan det trekkes tilfeldige observasjoner i hvert tre. Du angir antallet som skal trekkes for negative og antall positive. Kjør modellen på nytt og angi riktig parameter med tallen angitt i svaralternativene nedenfor. (Husk å kjør set.seed(42) før hver gang du kjører modellen.

Hvilket av disse alternativene gir det resultatet bedriften ønsker?
Velg ett alternativ:

- 200, 200
- 300, 400
- 619, 632
- 400, 300

Maks poeng: 2

19 Det finnes flere spesifikasjoner som gir ønsket resultat. Kjør nå en modell der samme parametern som i forrige oppgave er spesifisert med tallene 300 og 350. Kjør set.seed(42). Lag et plot som viser hvor mye hver variabel bidrar til prediksjonen.

Hvilken variabel er det som bidrar mest til å predikere utfallet?

Skriv variabelnavnet her:

Maks poeng: 1

20 Alder er ikke den viktigste prediktoren totalt sett, men man forventer at mental helse varierer med alder. Undersøk hvordan prediksjonen varierer med alder. Hvilket av følgende utsagn er riktig?

Velg ett alternativ:

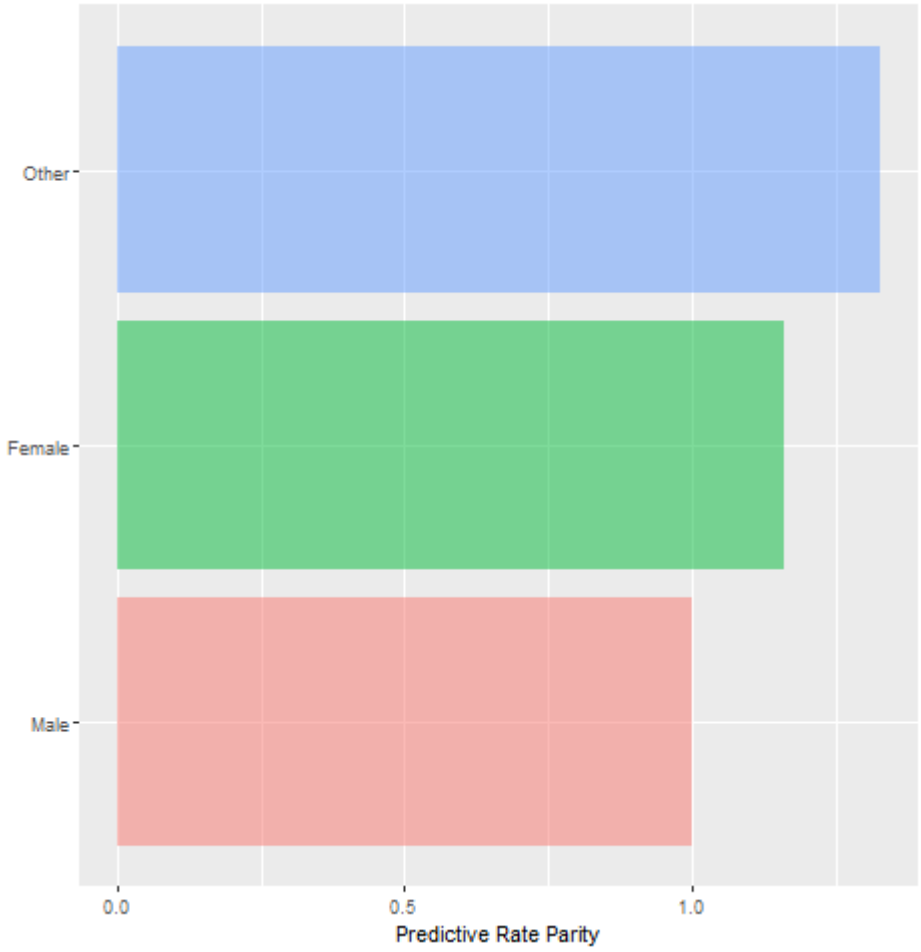
- De som i minst grad er predikert å ha oppsøkt behandling er i alderen 45-55 år
- De eldste aldersgruppene er i minst grad predikert å ha oppsøkt behandling
- De yngste personene er i minst grad predikert å ha oppsøkt behandling
- Det er ikke mulig å si noe om dette basert på modeller som random forest

---

Maks poeng: 1

21 Det er en bekymring for at modellen vil slå skjevt ut for menn og kvinner. Det er velkjent at kvinner er flinkere til å oppsøke hjelp, så det er en bekymring for at det kan slå ut også i prediksjonen.

Figuren nedenfor viser resultat fra beregnet "predicted rate parity".



Kommenter resultatene her. Angi på hvilken måte akkurat dette målet kan være relevant i dette tilfellet. Gi en vurdering av hvorvidt dette kan gi uheldige resultater eller om det er akseptable forskjeller, med en kort begrunnelse.

Skriv ditt svar her

Words: 0/500

Maks poeng: 4

22 Sjefen din har hørt om gradient boosting og synes det høres mye kulere enn random forest. Du blir derfor satt til tilpasse en ny prediksjonsmodel med stochastic gradient boosting.

Gjør analysen helt fra start av med boosting og gjør en nøktern vurdering av hvor god modellen kan forventes å være på nye data.

De samme kravene til asymetriske kostander gjelder. Kommenter resultatene nedenfor som et kort essay. Beskriv hva du har gjort med kort begrunnelse, og vurder de viktigste resultatene.

Husk å gjør om utfallsvariabel til numerisk. Du kan bruke følgende kode: mutate(treatment = ifelse(treatment == "Yes", 1, 0))

Skriv ditt svar her

Format

Words: 0

Maks poeng: 6

23 Legg inn fulstendig script her. Scriptet skal være slik at sensor skal kunne kjøre koden og få samme resultater som deg. Husk derfor å bruke set.seed() før hver prosedyre som har bruker tilfeldige tall på noe vis. (Særlig splitte datasett og random forest).

Skriv ditt svar her

```
1
```

Maks poeng: 0