Nyttårsforsett-kalkulator

Gruppe 55 - Fredrik Crook, Geir Sætre og Martin Sortland, 15.11.2024

Problemstilling

I denne oppgaven tar vi for oss følgende problemstilling: Hvordan kan vi ved hjelp av maskinlæring motivere personer til å endre mot en sunnere livsstil?

Prosjektets mål er å lage en kalkulator som kan gi en bruker en pekepinne på hvordan deres endringer i livsstil kan påvirke brukerens forventet levealder. Kalkulatoren skal bruke maskinlæring til å ta inn informasjon om brukeren som kjønn, røyking, alkohol, treningsmengde og livsstil. Kalkulatorens "output" skal være et realistisk anslag på hvordan endringer i livsstil kan påvirke brukerens forventet levealder. Prediksjonene er kun en indikasjon for brukeren, og brukeren må selv vurdere i hvilken grad informasjonen brukes for å støtte deres livsvalg.

Eksisterende løsninger

I Canada har Bruyere forskningsinstitutt og Ottawa sykehus utviklet en lignende løsning. Deres løsning er en kalkulator som beregner forventet levealder basert på kjønn, alder i dag, diett, aktivitetsmengde og forbruk av røyk og alkohol. Dataen denne kalkulatoren samler inn er så sammenlignet med canadiske innbyggere og gir en prosentvis sannsynlighet for hvor lenge man har sjans til å leve. [1]

Business objective

Vi skiller oss fra den Canadiske kalkulatoren ved å ikke beregne forventet levealder med den livsstilen brukeren har i dag, men heller beregne forventet levealder med endringer i livsstil. Kalkulatoren vår vil bestå av inputs som "Slutte å røyke", "Drikke mindre alkohol" og "Vær mer aktiv". Ut ifra en viss endring i livsstil skal den kalkulere hvor mange år forventet levealder vil øke eller reduseres med.

Gjennomføring av prosjektet

Til å gjennomføre utvikling og deployment av modellen er vi en gruppe på 3 studenter ved HVL. Det er oss som utviklere som blir stakeholders i prosjektet. Maskinlæringsmodellen lages i Google Colab [2] og nettsiden er laget ved bruk av Gradio[3]. Ved bruk av google colab og Gradio har vi mulighet til å effektivt oppdatere nettsiden, ettersom Colab har innebygget interface til Gradio.

Den største utfordringen ved prosjektet vil være å samle inn data til et akseptabelt datasett. Naturligvis har mesteparten av tiden blitt brukt på dette. Uten et akseptabelt datasett vil kalkulatoren være meningsløs. Det vil alltid være mulig å utvikle en maskinlæringsmodell som kan regne seg frem til en antatt levealder basert på et tilfeldig datasett, men om brukeren skal kunne dra nytte av kalkulatoren må datasettet være realistisk.

Tidslinje

Slik ble tiden vi hadde tilgjengelig brukt:

Uke 42: Drøfting av ulike problemstillinger og muligheter

Uke 43: Fremstilling av valgt problemstilling og tildeling av arbeidsoppgaver

Uke 44: Datainnsamling, kontakt med eksterne aktører

Uke 45: Datafremstilling, finne kilder og ekte korrelasjoner til datasettet

Uke 46: Utvikling av modell og nettside samt ferdigstille rapport.

METRIKKER

Business Metrics, hva skal til for at kalkulatoren lykkes?

Målet er at kalkulatoren skal kunne gi best mulig prediksjon av levealder og være pålitelig for brukere. For å få til dette setter vi en minimum ytelse på kalkulatoren. Minimum ytelse kan for eksempel være et gjennomsnittlig avvik på 5 år. Dette betyr at modellens “mean absolute error” (MAE) er lavere enn 5. MAE er den gjennomsnittlige absolutte feilen mellom de faktiske dataene og dataene modellen har regnet ut [4].

Ved MAE lavere enn 5 vil kalkulatoren ha en akseptabel feilmargin og brukeren kan stole på at modellens prediksjoner er noenlunde realistiske. Ved innhenting av data kom vi også over andre hindringer som kan gjøre modellens prediksjoner urealistiske, vi kommer mer innpå disse senere i rapporten.

Vi kan forbedre MAE ved å endre valg av maskinlæringsmodell. Ulike modeller egner seg best til ulike oppgaver. Vi finner modellen som passer våre data best ved å teste tilgjengelige modeller.

En annen metrikk som brukes for måling av feil til modellen er “Mean Squared Error” eller MSE. Denne metrikken måler avvik ved å kvadrere feilene i modellen. Dette fører til at større enkelte avvik straffes kraftigere og mindre avvik får relativt mindre relevans [5]. Ettersom vi har satt en øvre grense på 5 år i gjennomsnittlig feil, sikter vi mot en MSE som ikke er for langt over 25.

Den siste metrikken for beregning av avvik er “R-Squared” eller R2. Denne metrikken er best på å forklare variansen i dataene. Tallet på R2 går fra 0 til 1, hvor høyere verdi av R2 gir bedre forklaring av variasjonen av død basert på livsstil i datasettet [6]. Dette er en indikasjon på hvor godt modellen utnytter informasjonen i datasettet. Vanligvis vil en R2 på over 0.8 være et tegn på veldig god forklaringskraft i modellen, vi sikter mot dette.

Andre “business metrics” vi har er “Latency” og “Throughput”. For at kalkulatoren skal lykkes må den gi et raskt svar til brukeren per forespørsel. Et svar innen 10 sekunder per forespørsel er minimum for denne business metrikken. Throughput handler om nettsidens evne til å håndtere flere forespørsler samtidig fra ulike brukere[7].

DATA

En av de største utfordringene med prosjektet har vært innsamling av data. Vi ønsket å opprette et fyldig datasett hvor vi hadde data om ekte personer som har dødd, deres livsstilsfaktorer og levealder. Vi har vært i kontakt med Folkehelseinstituttet (FHI) for å prøve å få tak i disse dataene, men det var ikke mulig for dem å utlevere anonyme data på det vi var ute etter. Selv om personene anonymiseres, vil det fremdeles være en sjanse for å identifisere enkeltpersoner med summen av alle data, såkalt tilbakeveisidentifisering.

FHI har flere statistiske tabeller tilgjengelig på sine nettsider som for dødsårsaker og andre helserelaterte data. Disse kunne vi ikke bruke ettersom vi trenger data fra et utvalg enkeltpersoner. For å få slik data utlevert, må vi søke om personidentifiserbare data, noe som også medfører kostnader for sakbehandling og tilrettelegging. En løsning kunne være å hente data fra andre kilder for å knytte disse opp mot et datasett. Vi estimerte at det ville vært en veldig tidkrevende prosess, og vi valgte derfor å lage et syntetisk datasett med hjelp fra ChatGPT.

Syntetisk datasett
For at verdiene i det syntetiske datasettet skulle være realistisk, hentet vi informasjon fra ulike forskningsartikler. Denne informasjonen ble gitt til ChatGPT, som produserte et datasett bestående av 1000 personer mellom 50-100 år, inkludert deres levealder og livsstilsfaktorer. I tillegg ga vi ChatGPT informasjon om gjennomsnittlig levealder i Norge, gjennomsnittlig BMI, antall røykere, forbruk av alkohol og aktivitet/treningsmengde. Gjennom en prosess med prøving og feiling ble datasettet justert flere ganger, til vi hadde et datasett med verdier som virket realistisk.

Statistikken vi har benyttet for å lage datasettet er følgende:

Forventet levealder
Ifølge FHI sin folkehelserapport var forventet levealder i 2022 84,4 år for kvinner og 80,9 år for menn[8].

Røyking
I en undersøkelse av personer mellom 16 og 74 år var det i 2023 7% av kvinner og 8% av menn som røykte daglig. I tillegg var det 9% som røykte av og til [9].

Personer som røyker av og til øker risikoen for død med 38%. Daglige røykere lever i gjennomsnitt 11 år kortere [10].

Alkohol
Ifølge statistikk hentet fra [11] og [12], gir alkohol følgende utslag på levealder:

0 - 8 enheter/uke: Ingen forskjell i dødelighet
8 - 17 enheter/uke: Forventet levealder redusert med 0,5 år
17 - 25 enheter/uke: Forventet levealder redusert med 1-2 år
25 - 33 enheter/uke: Forventet levealder redusert med 4-5 år

Fysisk aktivitet
Vi gjennomgikk flere studier om fysisk aktivitet, som viste varierende resultat på hvor stor innvirkning trening har på levealder. Basert på disse funnene valgte vi å estimere påvirkningen av fysisk aktivitet på levealderen for ulike aktivitetsnivåer for inaktive, moderat aktive og aktive personer til henholdsvis -5, 3 og 5 år.

MODELLERING

Vi har utforsket følgende modeller:

Random Forest Regressor
Stocastic Gradient Descent
DecisionTree Regressor
Gradient Boosting Regressor
Lineær Regressor
Ridge Regression

Disse ble testet med forskjellige parametre, samt evaluert med Mean Absolute Error (MAE), Mean Squared Error (MSE og R2 score).

Her var det to modeller som pekte seg ut: Random Forest og Gradient Boosting Regressor, og sistnevnte viste seg å være best. Med teststørrelse på 0,2 fikk gradient Boosting Regressor en R2 score på 0,77, mot Random Forest like bak med 0,73. SDG Regressoren viste seg å være spesielt upassende, og konvergerte ikke.

Lineær og Ridge var ikke spesielt effektive, og endte på 0.5 (+- 0.1) i R2 resultater. Vi gikk derfor videre med Gradient Boosting Regressor.

GBR modellen ga en Mean Absolute Error på 4,22. Modellen har altså en gjennomsnittlig feil i predikert levealder på 4,22 år. Dette er innenfor grensen for minimal ytelse gitt tidligere.

Mean Squared Error verdien GBR modellen gir er på 27,0. MAE^2 er 17,8. En MSE verdi på 27 tyder på at modellen har noen få store feil, men ikke nok feil til å forkaste modellen.

Utfordringer med generering av syntetiske data, medførte vanskeligheter for optimal testing av modellene. Dette medførte lite tid til å raffinere modellene videre. Datasettet ble heller ikke normalisert, noe som kan ha redusert kvaliteten på noen av regresorene. Random Forest og Gradient Boosting Regressor synes likevel å ha fungert bra med ikke normalisert data.

DEPLOYMENT

Modellen er klar til å settes i drift, men det er fortsatt forbedringspotensial både når det kommer til forbedring av datasettet, modellen og nettsiden.

For å styrke modellen ytterligere og gi periksjoner som bedre samsvarer mer med virkeligheten, er det essensielt at vi får tilgang til faktiske data som vi kan bygge modellen på. Det vil også være behov for trening av modellen når nye data er tilgjengelig. Det er også mulig å bedre det syntetiske datasettet ytterligere ved å videre studere korrelasjoner mellom ulike livsstilsvalg og levealder. Datasettet kan utvides med nye livsstilsvariabler som for eksempel sukkerinntak, ulike typer treningsformer eller informasjon om sykdommer.

For å forbedre brukeropplevelsen kan vi be om tilbakemelding fra brukerne våre, og oppdatere nettsiden basert på disse tilbakemeldingene. Dette vil være en del av vedlikeholdet, i tillegg til forbedring av modellen.

Vi kan ikke vite med sikkerhet om brukerne faktisk holder deres nyttårsforsett eller ha tilgang til oppdatert informasjon om helsetilstanden og levealderen til brukerne våre. Det er derfor umulig for oss å få en presis forståelse av hvor godt våre prediksjoner samsvarer med virkeligheten. Modellens nøyaktighet kan derfor ikke vurderes basert på dette. Formålet vil i stedet være å gi en indikasjon på potensielle fordeler ved å opprettholde en sunn livsstil, og være et motiverende verktøy.

Et av de største problemene med modellen er at den ikke tar hensyn til brukerens alder. Dette betyr at en person som er 70 år potensielt kan få tildelt like mange ekstra leveår som en person på 20 år, til tross for at biologiske begrensinger og helserisikoer typisk øker med alderen. En vesentlig forbedring vil være å gjøre modellen mer sensitiv overfor alderen, slik at konsekvensen av livsstilsvalgene vil tilpasse seg på en mer realistisk måte. Her vil det også være nyttig å undersøke statistikk som baserer seg på livsstilsendringer i forskjellige stadier i livet, for eksempel hvordan justeringer i kosthold, røykevaner eller fysisk aktivitet spiller inn på individer i ulike aldersgrupper.

REFERANSER

[1] Project BigLife. Life Expectancy Calculator. projectbiglife.ca. Hentet fra: https://www.projectbiglife.ca/life-expectancy. Lastet ned 28.10.2024.

[2] Google. Google Colab. colab.research.com Hentet fra:https://colab.research.google.com/. Lastet ned 15.10.2024.

[3] Gradio. gradio.app. Hentet fra:https://www.gradio.app/. Lastet ned 15.10.2024.

[4] Xu, Z., van Donkelaar, A., Wu, Y., Lu, H. (2022) Mean Absolute Error Hentet fra:https://www.sciencedirect.com/topics/engineering/mean-absolute-error. Lastet ned 27.10.2024.

[5] J, Frost. (2024) Mean Squared Error. Hentet fra: https://statisticsbyjim.com/regression/mean-squared-error-mse/. Lastet: 27.10.2024.

[6] S, Taylor. (Ukjent) R-Squared. Hentet fra: https://corporatefinanceinstitute.com/resources/data-science/r-squared/. Lastet ned: 27.10.2024.

[7] T, Keary. (2024) Latency vs Throughput. Hentet fra: https://www.comparitech.com/net-admin/latency-vs-throughput/. Lastet ned: 27.10.2024.

[8] K, Bævre. (2023) “Forventet levealder i norge” Hentet fra: https://www.fhi.no/he/folkehelserapporten/samfunn/levealder/?term=. Lastet ned 10.11.2024.

[9] Helsedirektoratet. “Statistikk og historikk om røyking, snus og e-sigaretter”. helsedirektoratet.no. Hentet fra: https://www.helsedirektoratet.no/tema/tobakk-royk-og-snus/statistikk-om-royking-bruk-av-snus-og-e-sigaretter#:~:text=Rundt%207%20prosent%20av%20befolkningen,%C3%A5r%2C%20f%C3%A6rrest%20blant%20de%20yngste Lasted ned: 10.11.2024.

[10] . Vollset, S. E., Selmer, R., Tverdal, A., Gjessing. H. K. “Hvor dødelig er røyking?” Folkehelseinstituttet. Hentet fra: https://www.fhi.no/globalassets/dokumenterfiler/rapporter/2009-og-eldre/rapport-20064.pdf. Lastet ned: 10.11.2024.

[11] Shmerling, R. H., (2020) “Sorting out the health effects of alcohol”. Hentet fra: https://www.health.harvard.edu/blog/sorting-out-the-health-effects-of-alcohol-201808 0614427 Lasted ned: 11.11.2024.

[12] Sivertsen, Ø. S, “Mye alkohol gir kortere levetid” (2018) Hentet fra: https://tidsskriftet.no/2018/06/fra-andre-tidsskrifter/mye-alkohol-gir-kortere-levetid#:~:text=Ved

%20alkoholkonsum%20p%C3%A5%20100%E2%80%93200,3%20alkoholenheter%20etter%20norsk%20standard Lastet ned: 11.11.2024.