FRISCHSENTERET
Ragnar Frisch Centre for Economic Research

Frischsenterets rapportserie 3/2025

Maskinlæring og prognoser på justisfeltet

Ole Røgeberg, Andreas Kotsadam, Mette Løvgren og Tao Zhang

Rapport 3/25

Maskinlæring og prognoser på justisfeltet

Ole Røgeberg
Andreas Kotsadam
Mette Løvgren
Tao Zhang

Kontakt: www.frisch.uio.no

Rapport fra prosjektet "Prognoser for kriminalitetsutviklingen" (2722), finansiert av Justis- og beredskapsdepartement.

ISBN: 978-82-7988-309-8
ISSN: 2704-047X

Maskinlæring og prognoser på justisfeltet

Rapport 16 oktober 2025

Ole Røgeberg (Frischsenteret), Andreas Kotsadam (Frischsenteret), Mette Løvgren (OsloMet) og Tao Zhang (Frischsenteret)

Innholdsfortegnelse

1 Sammendrag

Dette prosjektet ble gjennomført på oppdrag fra Justis- og beredskapsdepartementet. Formålet var å undersøke om maskinlæringsmetoder anvendt på mikrodata fra administrative registre kunne brukes til å utarbeide pålitelige prognoser for kriminalitetsutviklingen i Norge. Prosjektet hadde to hovedmål: (1) undersøke om prediksjoner på individnivå med ulik prediksjonshorisont kunne aggregeres til informative prognoser for fremtidig lovbruddsutvikling, og (2) undersøke om selvrapporterte data fra spørreundersøkelser kan korrigere for mørketall i kriminalstatistikken.

Prosjektet brukte tre robuste og godt etablerte maskinlæringsmodeller (LASSO, Random Forest og XGBoost) og konstruerte prediktorer ved hjelp av et bredt sett med administrative registre. Registrene dekket blant annet demografi, utdanning, inntekt, familieforhold, barnevern og kriminalhistorikk. For mørketallsanalyser ble data fra Nasjonal trygghetsundersøkelse (NTU) 2022-2023 kombinert med registerdata dekket av deltagersamtykke for rundt 37 000 besvarelser.

Maskinlæringsmodellene lyktes godt i å avdekke forskjeller i fremtidig siktelsesrisiko. Selv fem år i forveien kunne modellene identifisere den tiendelen av befolkningen som ville stå for 75% av alle siktelser, og predikerte og faktiske siktelser korrelerte stabilt med rundt 0,25 uavhengig av tidshorisont (2-5 år). Dette er på nivå med eller bedre enn internasjonale studier av tilbakefall blant tidligere straffede. Samtidig bør det fremheves at selv i den mest lovbruddstilbøyelige tiendedelen av befolkningen var 90% eller mer uten observerte siktelser. Dette illustrerer faren ved å bruke modellene operativt på enkeltpersonsnivå: målretting av politiinnsats mot «høyrisiko-grupper» kan gi et uforholdsmessig kontrollpress og overvåkning av lovlydige borgere («profilering»), og kan forsterke eventuelle skjevheter i politiets avdekkingsvirksomhet og strategier mot ulike grupper og lovbruddstyper som ligger bakt inn i historiske data.

Til tross for god identifikasjon av risikogrupper, klarte ikke modellene å produsere pålitelige aggregerte prognoser for kriminalitetsutviklingen. Predikerte trender samsvarte dårlig med faktisk utvikling, noe som gjør metoden uegnet for strategisk planlegging og kapasitetsvurderinger.

Analysene av mørketall tydet på betydelige systematiske forskjeller mellom registrert og selvrapportert utsatthet for kriminalitet. Spesifikt tyder analysene på at unge (15-17 år) er mest utsatt for lovbrudd og at denne risikoen faller med alder og er nokså lik på tvers av kjønn. Dette avviker klart fra administrative data, der voksne over 20 er hyppigere registrert enn unge – trolig grunnet forskjeller i rapporteringstilbøyelighet. En metodisk utfordring i mørketallsanalysene var den begrensede mengden data tilgjengelig fra NTU som gjorde det krevende å kalibrere modellene presist.

Den viktigste innsikten fra mørketallsanalysene er at grupper som er mer utsatt for lovbrudd (selvrapportert i NTU) og grupper der lovbruddsutsatte har høyere rapporteringstilbøyelighet (selvrapportert i NTU) systematisk har flere registrerte oppføringer som lovbruddsoffer i administrative registre (observert i registerdata). De systematiske forskjellene i registreringssannsynlighet kan i betydelig grad forklares

statistisk av de systematiske forskjellene i utsatthet og rapporteringstilbøyelighet som hentes ut fra NTU-analysene. Dette styrker NTU-dataenes troverdighet, og tilsier at disse kan være en god kilde til analyser av befolkningens utsatthet for lovbrudd av ulike slag.

2 Innledning[1]

2.1 Bakgrunn for prosjektet

Denne rapporten dokumenterer metodikk, datagrunnlag og resultater fra prosjektet "Prognoser for kriminalitetsutviklingen". Prosjektet har bakgrunn i en utlysning fra Justisdepartementet (Doffin-referanse 2022-929436). Kjernen i utlysningen var spesifisert i konkurransegrunnlaget som følger:

> Prosjektet skal benytte statistiske metoder og stordata for å lage prognoser for kriminalitetsutviklingen. Under overskrifter som predictive policing, crime forecasting og big data policing, har statistisk metodeutvikling bidratt til nye måter å benytte kriminalstatistikken på. Dette prosjektet skal anvende tilsvarende tilnærming og metoder for å si noe om framtidig kriminalitetsutvikling.

Maskinlæring er et begrep som omfatter en rekke nyere statistiske verktøy og arbeidsmetoder (James et al. 2021). Metodene har til felles at de søker å avdekke innfløkte men stabile sammenhenger og mønstre i store og uoversiktlige datasett, og deretter bruke disse for å forutsi hva vi vil finne i nye/fremtidige data.

Slike metoder kan potensielt ha høy verdi på justisfeltet. Statistiske modeller er bedre enn selv erfarne eksperter på å integrere og vekte ulike informasjonskilder (Grove et al. 2015), og på justisfeltet har prediksjoner fra maskinlæringsmodeller utkonkurrert erfarne eksperter (Kleinberg et al. 2018) og vært tildels oppsiktsvekkende presise (Goel et al. 2021).

Mye av arbeidet med maskinlæring på justisfeltet har dreid seg om å identifisere områder som i bestemte tidsrom har høy sannsynlighet for å være åsted for lovbrudd («hot spots») ved hjelp av statistiske modeller som tar inn over seg blant annet tid og geografi («spatiotemporal models»). Disse modellene har særlig blitt utviklet i amerikanske storbyer, og kan bidra til bedre planlagte patruljeringsrunder og vaktlister - men «*den direkte overføringsverdien til andre samfunn er nok usikker. […] For å si det slik: politiet i Oslo vet utmerket godt hvor de skal være når pubene stenger; ikke bare områder, men utenfor hvilke konkrete puber*» (Skardhamar 2019).

Et annet hovedspor i bruken av maskinlæringsmodeller har vært å predikere hvilke individer eller grupper som står i særlig risiko for å begå lovbrudd. Vi vet, for eksempel, at registrerte lovbrudd forekommer langt hyppigere i grupper med visse kombinasjoner av kjønn, alder, familiesituasjon og sosioøkonomiske kjennetegn (Oslo Economics et al. 2022). Maskinlæringsteknikker gjør det mulig å utarbeide prognoser og identifisere grupper som bør være i fokus for forebyggende tiltak. Bruk av slike modeller for å

[1] Denne teksten trekker prosjektskissen Frischsenteret leverte under anbudsrunden.

identifisere hvor politiet bør fokusere sin innsats er derimot betydelig mer omstridt av flere grunner.

For det første tilsier personvernhensyn at politiet ikke skal ha fri tilgang til detaljert informasjon om hver borgers privatliv – som utdanning, helsehistorie, diagnoser, medisiner, familieforhold og -bakgrunn, eller yrkesaktivitet. Slike administrative data er i dag lagret i adskilte «datasiloer» eid av ulike offentlige instanser.

For det andre tilsier rettssikkerhet at det er problematisk å innrette politiets innsats mot enkeltpersoner på bakgrunn av innfløkte statistiske analyser som bygger på ugjennomtrengelige tekniske beregninger.

Og for det tredje er det en reell risiko for at en slik bruk ville forsterke eventuelle skjevheter i publikums *rapportering* av lovbrudd og politiets *prioritering* av lovbruddsavdekkende innsats. Årsaken til dette er kort fortalt at modellene alltid vil fortelle politiet at de bør lete etter flere lovbrudd de samme stedene man har sett lovbrudd i historiske data. Dersom noen gruppers lovbrudd i større grad rapporteres eller prioriteres av politiet vil slike modeller forsterke problemet ved å oppmuntre til en ytterligere økning av innsatsen mot disse samme gruppene. Dette har lenge vært en av de mest sentrale bekymringene rundt bruken av maskinlæringsmodeller på justisfeltet (O'Neil 2016). Slik profilering vil føre til ytterligere økning i *registrerte* lovbrudd i disse gruppene, og fremtidige maskinlæringsmodeller vil tolke dette som ytterligere forhøyet lovbruddstilbøyelighet og anbefale mer av det samme. Slik fokusert oppmerksomhet og kontrollvirksomhet utgjør i tillegg en forskjellsbehandling og mistenkeliggjøring av lovlydige borgere i gruppen på bakgrunn av gruppemarkører (hudfarge, klesstil, bosted). Dette kan tære på publikums tillit til politiet og gi opphav til oppfatninger om «klassejustis» og strukturell rasisme.

For å unngå slike problemer vil vi i dette prosjektet isteden undersøke hvorvidt maskinlæringsteknikker kan brukes til å utforme prognoser for kriminalitets*utviklingen*: det vil si *aggregerte* prediksjoner som sier hvordan antall lovbrudd – av ulike typer og i ulike grupper – vil forventes å endre seg dersom dagens trender og mønstre opprettholdes.

Slike modellbaserte «bottom-up» prognoser kan bistå i den strategiske innretting av fremtidig kapasitet og kompetanse etter fremtidige behov, og de gir en referanseutvikling som faktiske tall kan sammenlignes med. Dermed blir det lettere å avklare hvorvidt nye tall faktisk er uventet, eller om de er i tråd med det vi burde forvente gitt den underliggende aldersdemografiske og sosioøkonomiske strukturen på befolkningen.

I tillegg til slike prognoser undersøker vi om det er mulig å *korrigere* registerbaserte analyser for systematisk underrapportering av lovbrudd – for dermed å gi et mer representativt bilde av den faktiske lovbruddsbyrden i befolkingen. Til dette arbeidet bruker vi survey-baserte data på opplevde lovbrudd (Løvgren et al. 2024) og undersøker i hvilke deler av befolkningen vi ser størst systematisk avvik mellom registrerte og selvrapporterte ofre for ulike typer lovbrudd dekket av undersøkelsen.

3 Prognoser og mørketall

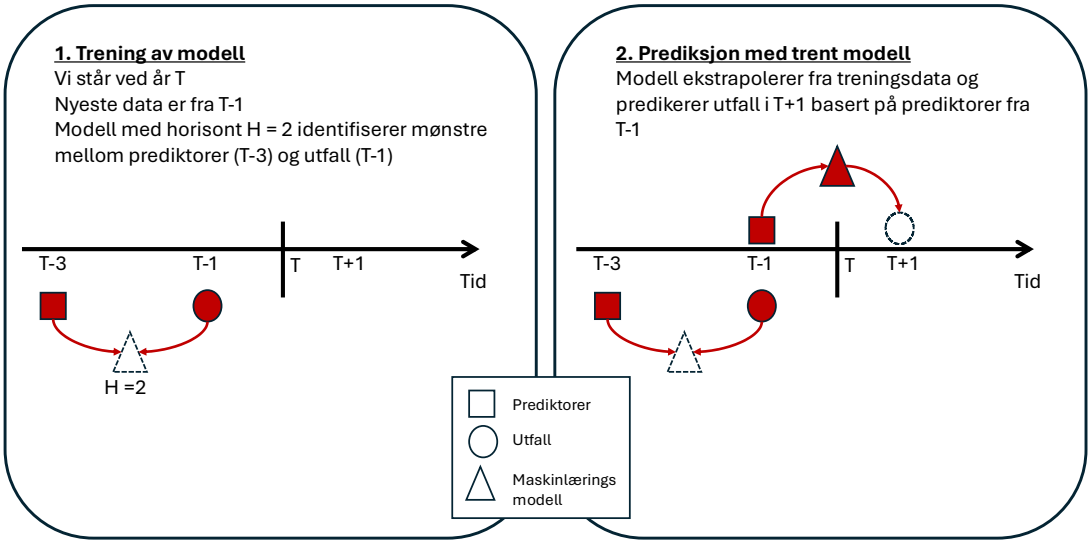
3.1 Analytisk oppsett – prediksjoner (arbeidspakke 1)

Dette er et omfattende og komplekst prosjekt, og i dette kapittelet gir vi en grunnleggende «høynivå» innføring i metodikken. Mer tekniske detaljer er å finne i teknisk appendiks.

Maskinlæringsmodeller er verktøy som avdekker intrikate men robuste sammenhenger i et datasett og bruker disse til å predikere utfall i nye datasett. Vi *trener* modellen ved å mate den med data på *prediktorer* og et *utfall*. Modellen følger deretter en bestemt oppskrift (algoritme) som finner systematiske sammenhenger mellom prediktorene og utfallene. En måte å tenke på dette i vår sammenheng er at modellen leter i data for å finne et finmasket sett av grupper i befolkningen som grupperer personer etter deres utviste lovbruddstilbøyelighet. Deretter tar vi et nytt datasett (f.eks. befolkningen i et senere år) og ber modellen plassere denne nye populasjonen inn i det gamle gruppe-systemet og gjette at hver person vil siktes omtrent så ofte som gruppen deres ble siktet historisk.

Ulike maskinlæringsmodeller som Random Forest, XGBoost og LASSO skiller seg fra hverandre ved at de følger ulike oppskrifter for å finne slike mønstre, og ved at de har forskjellige «justeringsparametre» (tuning parameters) som brukeren må finne gode innstillinger for. Disse påvirker måten modellen leter etter mønstre i data og blir i praksis satt ved å prøve ut ulike innstillinger og se hvilke som fungerer best i en gitt prediksjonskontekst.

I dette prosjektet er tanken å bruke detaljert registerbasert informasjon om enkeltpersoner for å anslå hvor mange siktelser disse enkeltpersonene vil få i et fremtidig utfallsår, for deretter å aggregere dette til prognoser på et høyere nivå (f.eks. region, aldersgrupper).



Figur 1 - Trening og prediksjon med en maskinlæringsmodell. I panel 1 lærer modellen mønstre ved å se prediktordata fra T-3 opp mot utfallsdata fra T-1. I panel 2 bruker modellen disse mønstrene til å gjette på fremtidige utfall i T+1 basert på prediktordata fra T-1.

I sin enkleste form har prediksjonen to skritt (Figur 1). I første trinn skal modellen *trenes* på et datasett (*treningsdata)* med informasjon om *prediktorer* (informasjon vi kan bruke i prediksjonen) og *utfall* (det vi skal predikere). Utvalget kan f.eks. bestå av alle norske innbyggere i et kalenderår, prediktorene kan være en rik og detaljert beskrivelse av hver person basert på informasjon *som er tilgjengelig på prediksjonstidspunktet* (familiebakgrunn, lovbruddshistorikk, utdanningshistorie osv). Utfallet kan f.eks. være "alle siktelser i et kalenderår» - der modellens *prediksjonshorisont* er antall år mellom prediktor- og utfallsår.

Når modellen trenes på dette datasettet identifiserer den sammenhenger mellom prediktorer og utfall. Vi gir så modellen et nytt datasett med prediktorer (f.eks. oppdatert informasjon om alle norske innbyggere fra et senere år), og modellen anslår hva vi kan forvente av lovbrudd fra disse i fremtiden *gitt at utfallet fortsatt henger sammen med prediktorer på samme måte som i treningsdata*.

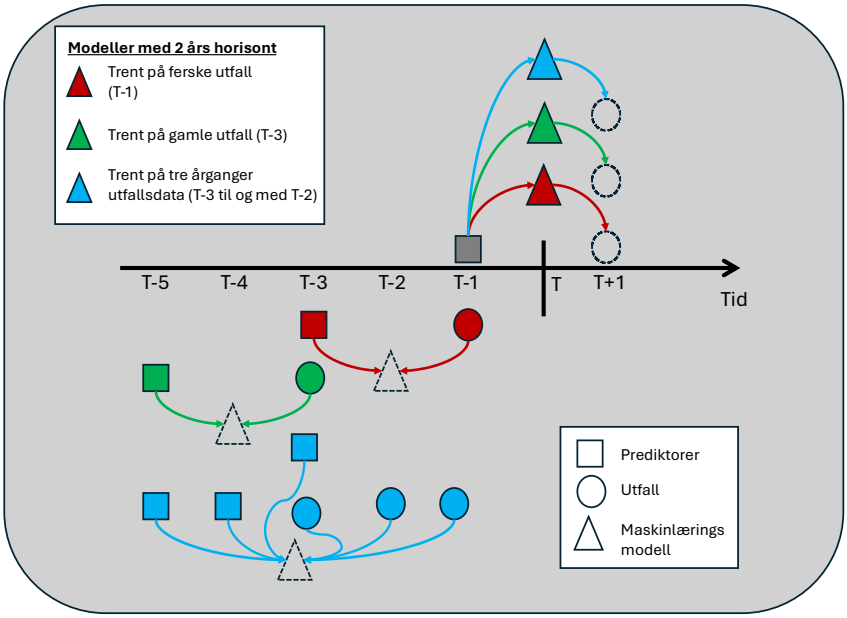
Prediksjonshorisont, treningsår og prediksjonsår

Et sentralt uavklart spørsmål da prosjektet startet handlet om nettopp denne antagelsen: I hvilken grad vil slike prediktive sammenhenger og mønstre i data holde seg over tid? Dette betyr mye for hva slags *prediksjonshorisont* metoden kan egne seg for: Data fra administrative registre vil typisk først være tilgjengelig kalenderåret etter. Ønsker vi å predikere lovbrudd om tre år trenger vi derfor en modell som kan predikere fire år frem, siden vi er nødt til å bruke ett år gamle data til å predikere fra. Siden de ferskeste utfallene vi vil ha data på også er fra i fjor, vil den ferskeste modellen med fire års horisont være en modell som fant mønstre i prediktorer fra *fem* år tilbake.

Mer generelt: En modell med prediksjonshorisont på H år vil i praksis predikere H-1 år frem i tid (siden vi må predikere H år fremover basert på *fjorårets* data), og vil måtte være trent med prediktordata som er H + 1 år gammelt. Vil du predikere 1 år frem betyr det at

den ferskeste modellen vil være en med prediksjonshorisont H = 2 som har lært mønstre ved å se på hva slags lovbrudd befolkningen vi kunne observere i tre år gamle data hadde i fjor.

Dette betyr at vi i prinsippet kan predikere lovbrudd i samme år på mange ulike måter ved å bruke modeller med ulik prediksjonshorisont, treningsdata og prediktordata. Dette er illustrert i Figur 2 (se også i Tabell 1), der vi står i år T og ønsker å predikere lovbruddsutfall i år T + 1. Ønsker vi å bruke *ferske prediktorer* må vi bruke en modell med kort horisont. Ønsker vi å ha en modell trent på de sist tilgjengelige dataene må vi bruke en «*fersk*» *modell* (trent på de siste tilgjengelige utfallsdataene) snarere enn å bruke en gammel (trent på eldre data).



Figur 2 - Ulik treningsdata - gitt horisont. Figuren viser hvordan en modell med to års prediksjonshorisont kan trenes på data fra ulike par med trenings og utfallsdata, eller sett med slike. Ved prosjektets start var det uavklart hva slags betydning slike valg ville ha for kvaliteten på prediksjonene.

Alternativt kan vi lage prediksjoner for T+1 ved hjelp av modeller med lengre horisont og eldre prediktordata, som f.eks. en modell med horisont 5 år (D-F i Tabell 1). Prediksjonen for neste år må da baseres på prediktordata fra fire år tilbake. Utover dette har vi de samme distinksjonene, ved at også denne modellen kan være trent på «nyeste data tilgjengelig» (D), eldre data (E) eller data fra flere år (D-F).

Tabell 1

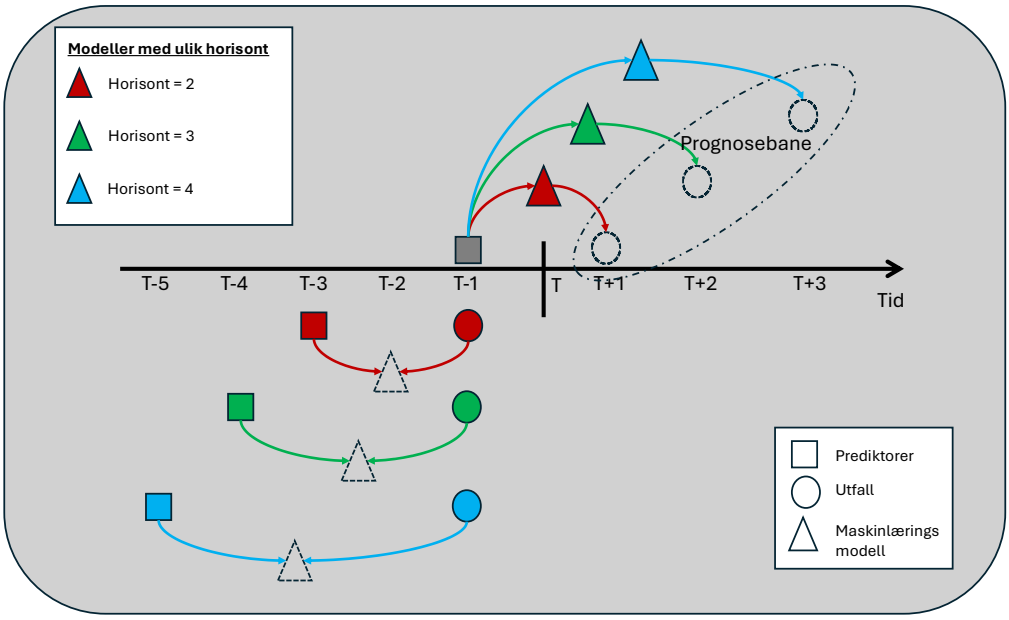
Eksempel	**Modellhorisont**	**Treningsdata**		**Prediksjonsdata prediktorer**
		Utfall	**Prediktorer**	
A	2	T-1	T-3	T-1
B	2	T-5	T-7	T-1
C	2	Fra T-3 til T-1	Fra T-5 til T-3	T-1

D	5	T-1	T-6	T-4
E	5	T-5	T-10	T-4
F	5	Fra T-5 til T-1	Fra T-10 til T-6	T-4

I utgangspunktet er det rimelig å forvente at noen av disse alternativene vil gi bedre prognoser enn andre, men hvilke som fungerer best er ikke gitt på forhånd og må undersøkes i praksis. Generelt vil prediksjonsmodeller være svakere dess høyere prediksjonshorisont vi har (det er lettere å si hva som vil skje imorgen enn hva som vil skje om 10 år) men hvor mye svakere avhenger av kontekst. Mye avhenger også av hvor stabile mønstre maskinlæringsmodellen utnytter: Dersom mønstrene er stabile over tid kan det være tilstrekkelig å trene en modell og bruke denne i flere år uten å trene den på nyeste utfallsdata hvert år. Dersom samfunnet endrer seg raskere er modellene i større grad «ferskvare» som jevnlig må trenes på nye data. Dersom mønstrene er særlig subtile og innfløkte kan det være nødvendig med flere årganger med data for at modellen skal kunne gjenfinne dem. Samtidig kan vi også tenke oss situasjoner der en modell trent på eldre data kan være mer egnet enn en på ferskere data, som da COVID-pandemien ga oss en periode der mønstrene i utfallsdata trolig var svært ulike de som var gjeldende både før og etter.

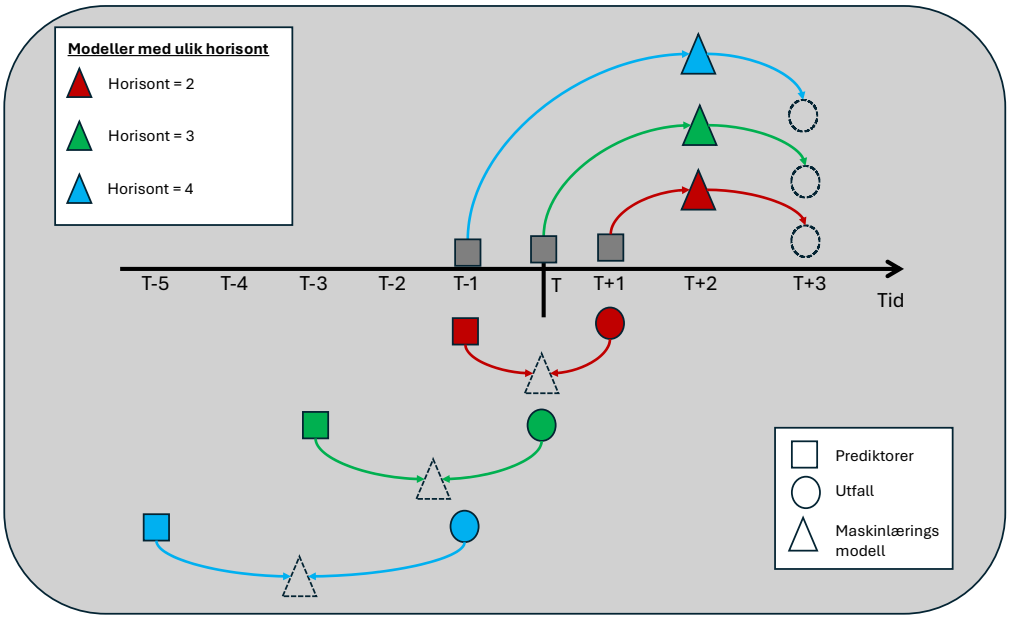
Prognosebaner og utfall

Så langt har vi sett på hvordan vi kan predikere ett bestemt utfall i ett bestemt år, men i dette prosjektet ønsker vi i tillegg å se på prognoser for mer enn et år om gangen samt prognoser for ulike typer lovbrudd. Dette øker kompleksiteten ytterligere.



Figur 3 – Prognosebane. Figuren viser hvordan maskinlæringsmodeller med ulik prediksjonshorisont kan trenes til å finne mønstre mellom utfallsdata i år T-1 og prediktorer fra to, tre eller fire år tidligere. Ved å bruke de trenede modellene til å predikere basert på prediktorer fra T-1 får man en prognosebane med prognoser både ett, to og tre år frem i tid.

Prognosebaner for fremtidige år er det minst krevende, ettersom det gjøres ved å bruke flere modeller - med ulike prediksjonshorisonter - på samme prediktordata (Figur 3). For hvert år som går kan disse prognosene oppdateres og gi en ny prognosebane. Over tid kan disse prognosene oppdateres, og vi får vi en sekvens av ulike prognoser for samme år der nyere prognoser baserer seg på modeller med kortere prediksjonshorisont og nyere prediktordata (Figur 4). Disse er likevel ikke direkte sammenlignbare ettersom de predikerer siktelser i et fremtidig år for *delvis ulike populasjoner*. I figur 4 ser vi f.eks. at modellen med 4 års horisont predikerer siktelser i T+3 for alle vi observerer i registerdata i T-1 mens modellen med 2 års horisont predikerer siktelser i T+3 for alle vi observerer i registerdata i T+1. Siden det vil være dødsfall, utflyttinger og innflyttinger mellom T-1 og T+1 vil dette være en (trolig liten) kilde til diskrepans.



Figur 4 – Langtidsprognoser oppdatert. Figuren viser hvordan en prognose kan oppdateres når nye årganger med prediktordata blir tilgjengelig. Siden det er kortere tid igjen til året man er opptatt av bytter man til en modell med kortere prediksjonshorisont. I figuren er disse modellene i tillegg trent på ferskest mulige treningsdata.

Å lage slike prognosebaner for *ulike typer lovbrudd* er derimot mer krevende, ettersom dette er *ulike prediksjonsproblem*. En spesifikk lovbruddstype som «ordensforstyrrelser» kan være mer eller mindre predikerbar enn «siktelser for seksuallovbrudd», og de kan kreve ulike innstilling av modellens justeringsparametre og ha ulik grad av «ferskvareproblem» i hvor raskt mønstrene endrer seg. Antallet lovbrudd av ulike typer vil også være ulike, som gjør at modellen har ulik mengde informasjon å lære fra selv om de samme årgangene med data er tilgjengelig.

Prediktorsett

Mulighetene for å avdekke robuste og vedvarende sammenhenger mellom utfall og prediktorer avhenger i stor grad av hvor mange observasjoner vi har og hvor mye informasjon vi har om potensielle prediktorer for hver observasjon.

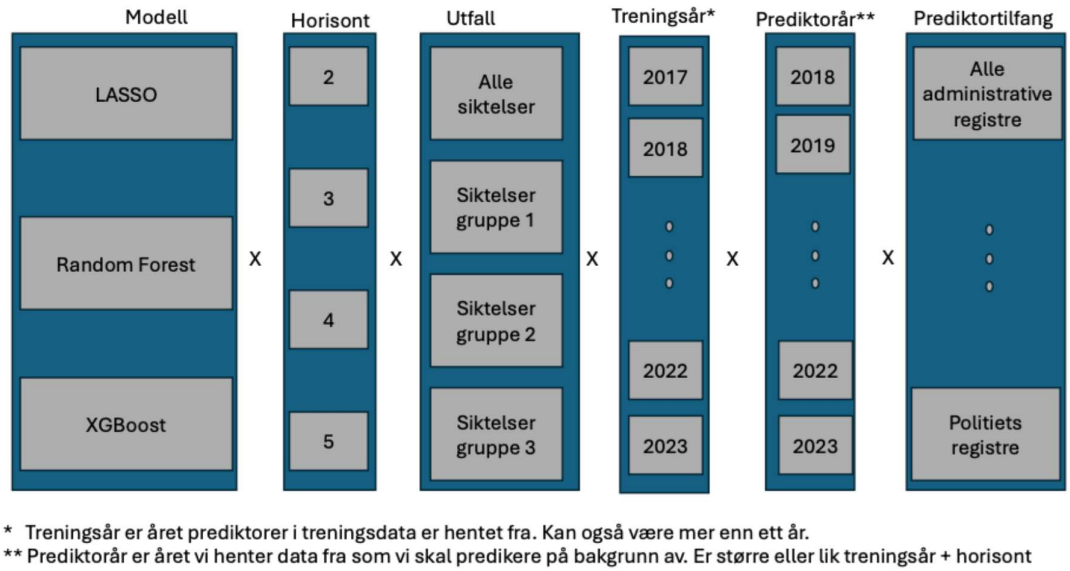
Her står prosjektet overfor to hensyn. På den ene siden er det ønskelig å si noe om *potensialet* for å lage informative prognoser. Dette tilsier at vi bruker mye informasjon fra et bredt sett registerdata. På den andre siden er det ønskelig å si noe om nytten i *praktiske anvendelser.* Her må vi også ta inn over oss at offentlige enheter i praksis ikke vil ha full tilgang til hele settet av prediktorer vi hadde tilgang på i dette prosjektet.

For å belyse dette var prosjektets strategi å først kjøre modeller med full bruk av tilgjengelige registerdata for å undersøke i hvilken grad metoden er egnet for prognose i et best-case, og deretter vurdere i lys av disse resultatene i hvilken grad det var av interesse å avklare ulike registerkilders relative prediktornytte.

Innsnevring

Som det fremgår av disse ulike analytiske valgene er det et eksploderende antall analyser som potensielt kan gjennomføres: ulike modeller med ulike tuning-parameterverdier kan predikere ulike utfall på ulike tidspunkt i fremtiden basert på ulike datakilder fra ulike år. Det totale antallet analyser dette utgjør er ikke praktisk gjennomførbart (Figur 5), og følgende innsnevringer ble gjort.

Prosjektet tok sikte på å sammenligne tre ulike maskinlæringsmodeller (Random Forest, LASSO og XGBoost) for prognoseformål, der ønsket var å lage prognoser for alle siktelser samlet samt for et knippe snevrere lovbruddskategorier. I tillegg ønsket vi å undersøke «ferskvareproblemet» ved å se hvor raskt prediksjonene svekkes ettersom tiden går og treningsdataene som legges til grunn eldes, samt forstå hvor raskt prediksjonene svekkes med prediksjonshorisontens lengde. Dette betyr i praksis at vi må trene prediksjonsmodeller på gamle data (f.eks. utfall i 2016) og «late som» vi står i 2017 og vil predikere lovbrudd med ulike horisonter slik at vi kan sammenligne prediksjonene med hva som faktisk skjedde.



Figur 5 – Antallet mulige kjøringer er høyt. Figuren viser de ulike dimensjonene som må tas hensyn til. Dersom alle mulige kombinasjoner skulle gjennomføres ville dette være mer enn 2000 analyser.

Vi har valgt følgende føringer på arbeidet:

1. Vi begrenser oss til fire horisonter (2, 3, 4 og 5), noe som vil gi oss prediksjoner 1, 2, 3 og 4 år frem i tid når registerdata med prediktorer er tilgjengelig med ett års forsinkelse.
2. Vi begrenser oss til fire utfall
    a. Alle lovbrudd
    b. tre større sett med grupperte lovbruddskategorier:
        i. vinning – lovbruddsgruppe 1 og 2
        ii. vold – lovbruddsgruppe 4 og 5
        iii. eiendom – lovbruddsgruppe 3 og 7
3. Vi undersøker spørsmålene i egne sett med målretteder analyser
    a. Prognoser over flere år vil undersøkes ved å bruke 2014 som utgangspunktår og utarbeide prognoser med ulike horisonter. Prognosene vil gjelde for årene 2016-2019 og kan sammenlignes med faktiske utfall i disse årene. Dette vil gjøres for alle utfallsgruppene.
    b. Behovet for å oppdatere prognosemodeller vil undersøkes ved å predikere 2019-utfall med 2017-data med modeller trent på ulike enkeltår
    c. Verdien av å oppdatere prognoser vil undersøkes ved å predikere for 2019
        i. basert på 2017 prediktor-data med to års horisont
        ii. basert på 2016 prediktor-data med tre års horisont
        iii. basert på 2015 prediktor-data med fire års horisont
        iv. basert på 2014 prediktor-data med fem års horisont
4. Justeringsparametre er tidkrevende å stille inn ved hjelp av data, ettersom dette krever at man trener modellen gjentatte ganger (f.eks. 10) med hver av ulike innstillingskombinasjoner. For to av modellene (XGBoost og Random Forest) er antallet kombinasjoner for høyt til at alle kan prøves ut, og det vanlige er å prøve ut f.eks. ti tilfeldig trukne kombinasjoner og velge den beste, med mulighet for å trekke ytterligere kombinasjoner i «nærheten» av den beste for å se om ytterligere forbedringer er mulig. Preliminære analyser indikerte at de samme innstillingene er best for modeller med ulik horisont, så vi valgte å trene hver modell for hvert utfall for horisont 2.
5. Dersom prognoser med full informasjonstilgang er informative for lovbruddsutviklingen ønsker vi også å undersøke betydningen av informasjonstilgang. I så fall vil denne analysen gjøres på de tre modellene trent på ett bestemt år med *alle siktelser* som utfall.

3.2 Analytisk oppsett – mørketall (arbeidspakke 2)

Siktelser kan gi et misvisende inntrykk av hvor utbredt lovbrudd er i et samfunn og et skjevt bilde av hvem som rammes. Dette skyldes seleksjonsproblematikk. Mer konkret: mange lovbrudd faller fra på veien frem til en siktelse, og det er ikke tilfeldig hvilke. Ikke alle lovbrudd anmeldes, ikke alle anmeldte lovbrudd fører til en siktelse, og for noen lovbruddstyper (trafikk, rusmiddel) styrer politiet i stor grad siktelsesutviklingen selv gjennom kontroll- og avdekkingsvirksomhet.

I denne delen av prosjektet var ambisjonen å bruke data fra registerkoblede spørreundersøkelser til å korrigere siktelsesdata for underrapportering, og slik gi et mer representativt bilde av nivået på ulike lovbrudd og hvem som særlig rammes.

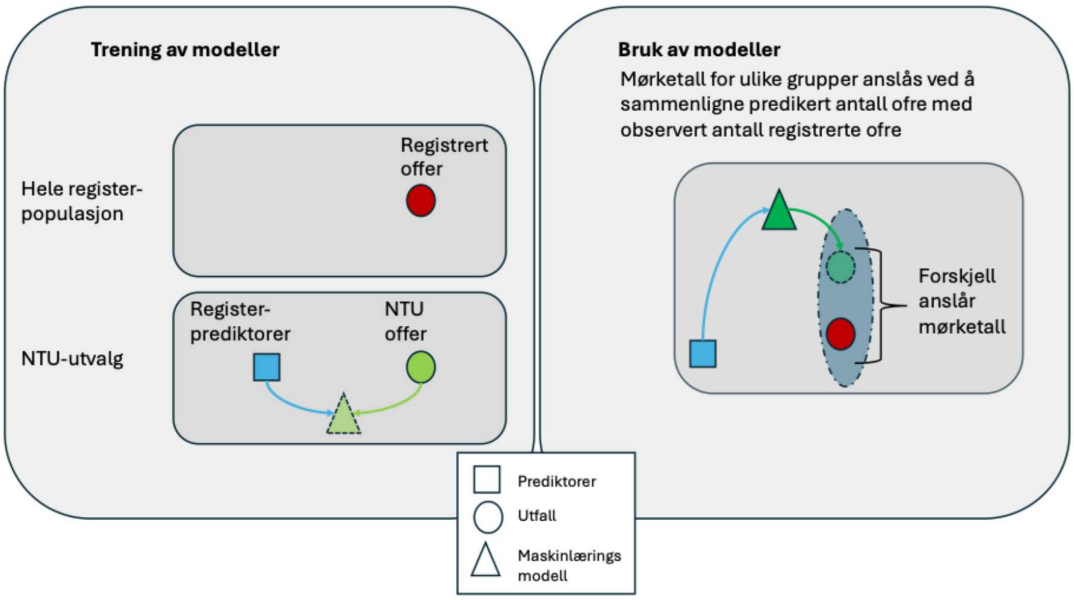
Til dette bruker vi nasjonal trygghetsundersøkelse (NTU). Denne blir beskrevet i mer detalj senere (se kapittel 3.4.2), men i korte trekk er dette en stor spørreundersøkelse finansiert av Justisdepartementet der et flertall av deltagerne har samtykket til at forskere kan koble svarene deres til et (begrenset og spesifisert) sett med administrative registerdata. Dette datasettet kan også belyse lovbrudd et offer har vært utsatt for *uten å forstå at det var en kriminell handling*. For eksempel stilles det spørsmål om respondenten har opplevd konkrete hendelser som å slås med åpen hand.

Data fra nasjonal trygghetsundersøkelse kan kun kobles til et begrenset sett med administrative registerdata som er dekket under det informerte samtykket respondentene har gitt. Disse inkluderer ikke alle de datakildene og variablene vi har tilgang til på arbeidspakke 1 – og mer spesifikt gir de ikke dekning for å koble på lovdata for å undersøke om respondenten faktisk er registrert som offer for ulike typer lovbrudd. Vi er derfor nødt til å undersøke mer indirekte alternative strategier.

I analysene som følger valgte vi å slå sammen NTU-data fra de to rundene i post-pandemi år. Vi antar da at det ikke var store endringer i hvem som ble utsatt for ulike lovbrudd mellom 2022 og 2023, og at rapporteringssannsynligheten var nokså lik i de to årene. Ettersom data fra pandemi-perioden i mindre (og ukjent) grad vil generalisere seg til post-pandemi perioden ble disse utelatt. I tillegg ble det gjennomført store endringer i spørreskjemaet til NTU etter 2020.

Analyse 1: Sammenligne omfang av registrerte og selvrapporterte ofre

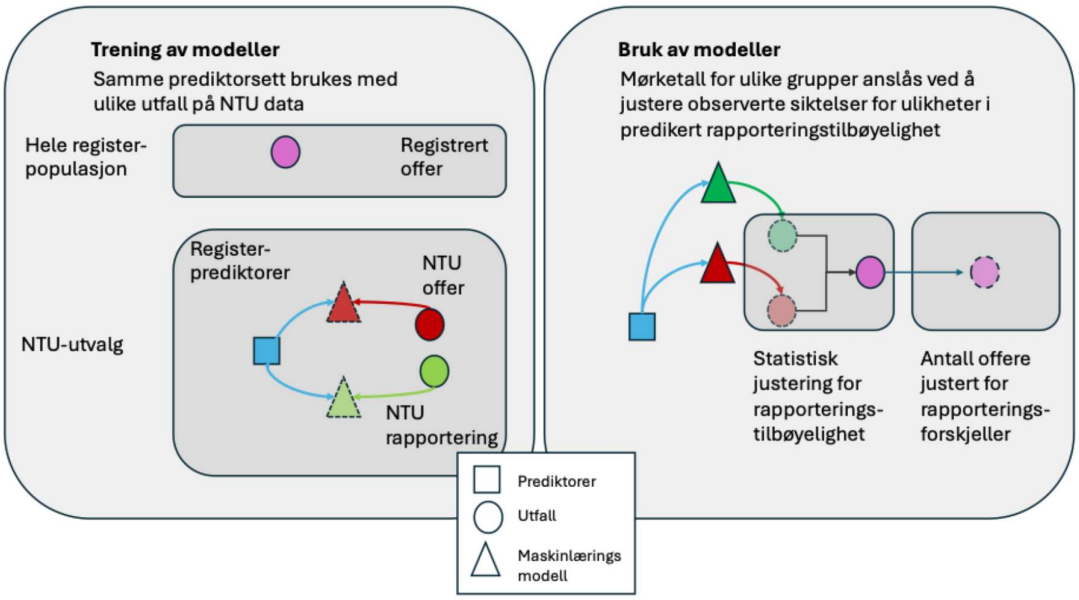
Den første analysen trener to maskinlæringsmodeller og sammenligner prediksjonene de gjør for ulike grupper (Figur 6). Modellene får tilgang til samme registerprediktorer, men bruker treningsdata fra forskjellige utvalg med forskjellige utfall. Den ene modellen trenes på populasjonsdekkende data for å anslå sannsynligheten for å *være registrert som offer* for en gitt lovbruddskategori. Den andre trenes på data fra NTU for å anslå sannsynligheten for å *være offer,* der kategoriseringen er basert på svarene fra spørreundersøkelsen. Deretter brukes begge modellene på registerdatapopulasjonen, og anslagene på antallet forventede ofre og antallet forventede *registrerte* ofre kan sammenlignes innenfor ulike befolkningsgrupper. På denne måten bygger vi opp et kart over hvor denne typen lovbrudd rammer i befolkningen og hvor denne typen lovbrudd *fanges opp* av rettssystemet.



Figur 6– Sammenligning av opplevde lovbrudd og registrerte ofre. Figuren viser hvordan maskinlæringsmodeller trenes på NTU data for å predikere sannsynligheten for å være utsatt for ulike lovbrudd. Modellen brukes til å predikere utsatthet i registerdata og sammenlignes med faktiske registrerte lovbrudd.

Analyse 2: Predikere sannsynligheten for rapportering

Den andre analysen bruker NTU-data mer direkte til å anslå underrapportering (Figur 7). Her vil vi utnytte at respondentene i NTU blir spurt om de har rapportert lovbrudd de har opplevd til politiet. Dette kan vi bruke til å avgrense data til *kun* de som har opplevd et lovbrudd av type X, slik at vi kan trene opp en maskinlæringsmodell til å predikere hvor sannsynlig det er at ulike typer personer ville meldt inn denne typen lovbrudd dersom de opplevde det. Denne sannsynligheten kan deretter brukes til å korrigere andelen vi observerer som ofre i registerdata ved å justere for forskjeller i predikert rapporteringstilbøyelighet.



Figur 7– Registrerte offer justert for ulik rapporteringstilbøyelighet. Figuren viser hvordan maskinlæringsmodeller trenes på NTU data for å predikere sannsynligheten for å være utsatt for ulike lovbrudd og sannsynligheten for å rapportere lovbrudd dersom man er utsatt for dem. Disse modellene brukes til å predikere utsatthet og rapportering for populasjonen som helhet (registerdata). Deretter estimeres sannsynligheten for å være registrert som offer i en logit med utsatthet og rapporteringstilbøyelighet som forklaringsfaktorer. Den estimerte modellen brukes til å anslå hva sannsynligheten for å være registrert som offer ville vært hvis alle hadde samme rapporteringstilbøyelighet.

3.3 Maskinlæringsmodeller

Maskinlæringsmodeller søker etter mønstre i data, og ulike maskinlæringsmodeller representerer ulike strategier for å finne frem til disse. Vi bruker tre ulike modelltyper og gir i dette kapittelet en kortfattet intuitiv forklaring på hvordan disse virker.

Felles for disse metodene er at de står overfor en avveining mellom modellfleksibilitet og overgeneralisering. Dess mer fleksibel modellen er, dess bedre kan den beskrive og ekstrapolere subtile og komplekse mønstre i treningsdata, men dess større risiko er det for at disse mønstrene egentlig er tilfeldigheter som ikke vil være tilstede i nye data. Dersom modellen likevel bruker disse i prediksjonen blir prediksjonen dårligere fordi modellen overgeneraliserer («overfitting»). Grepene som skal redusere dette problemet er til dels felles for alle de ulike maskinlæringsmodellene.

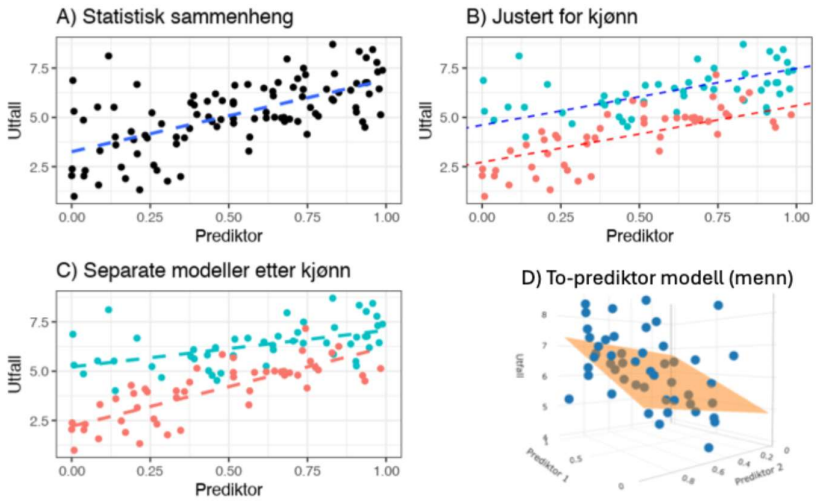
Et sentralt grep er K-fold validering og fininstilling av modellparametre. De ulike maskinlæringsmodellene har «brytere» som lar oss stille inn hvor fleksible og generaliserende modellene skal være – men de beste innstillingene vil variere med hva slags data vi har og hva vi prøver å predikere. For å finne gode innstillinger kan man dele treningsdata i K ulike subsett, for eksempel 10. Deretter deler man data i ti like store deler og bruker ni av delene til å trene en maskinlæringsmodell for å se hvor godt den treffer på de «usette» dataene man utelot. Dette gjentar man mens man roterer hvilken av delene man holder utenfor treningsrunden. Til slutt har man ti ulike målinger på hvor godt modellen predikerer med ett bestemt sett med innstillinger for denne typen data. Deretter gjentar man hele runden med et nytt sett innstillinger – og når man er gjennom

alle man ønsker å sammenligne velger man de innstillingene som i snitt fungerte best over sine 10 tester. Mer konkret hva disse justeringsparametrene er varierer fra modell til modell.

3.3.1 LASSO

En av de vanligste statistiske analysemetodene er lineære regresjonsmodeller. Et enkelt eksempel er illustrert med fiktive tall i Figur 8, panel A: vi har en prediktor (inntekt) som henger sammen med et utfall (antall siktelser), men sammenhengen er støyfull fordi mange andre ikke-modellerte faktorer også spiller inn. Den lineære regresjonsmodellen regner ut den linjen som i best grad beskriver sammenhengen, definert matematisk som at den minimerer alle «kvadrerte avvik»[2].

I figuren illustrerer vi en modell med kun en prediktor – men slike modeller er fleksible verktøy som kan bruke mange prediktorer på en gang. I illustrasjonen kan vi f.eks. åpne for en nivåforskjell etter kjønn (Figur 8, panel B) som tillater en kjønnsforskjell som er like stor uavhengig av inntekt. Vi kan også tillate en samspills- eller interaksjonseffekt der kjønnsforskjellen endrer seg med inntektsnivået vi sammenligner menn og kvinner ved (Figur 8, panel C). Man kan også ha mer enn en kontinuerlig prediktor. Dette er illustrert for en situasjon med to prediktorer (Figur 8, panel D), der modellen regner seg frem til en geometrisk flate som sier «hvor høyt» utfallet typisk ligger hvis du har en bestemt kombinasjon av de to prediktorene.



Figur 8 – Lineære modeller. Figurene illustrerer lineære modeller. Panel A viser en støyfull men systematisk statistisk sammenheng der er utfall øker med verdien på en prediktor. Panel B lar modellen forskyve linjene for kvinner og menn så de har en nivåforskjell. Panel C lar modellen ha ulik nivå og helning for kjønnene. Panel D bruker kun data fra menn og viser hvordan en lineær modell estimerer en "flate" i et tredimensjonalt rom når man har to kontinuerlige prediktorer.

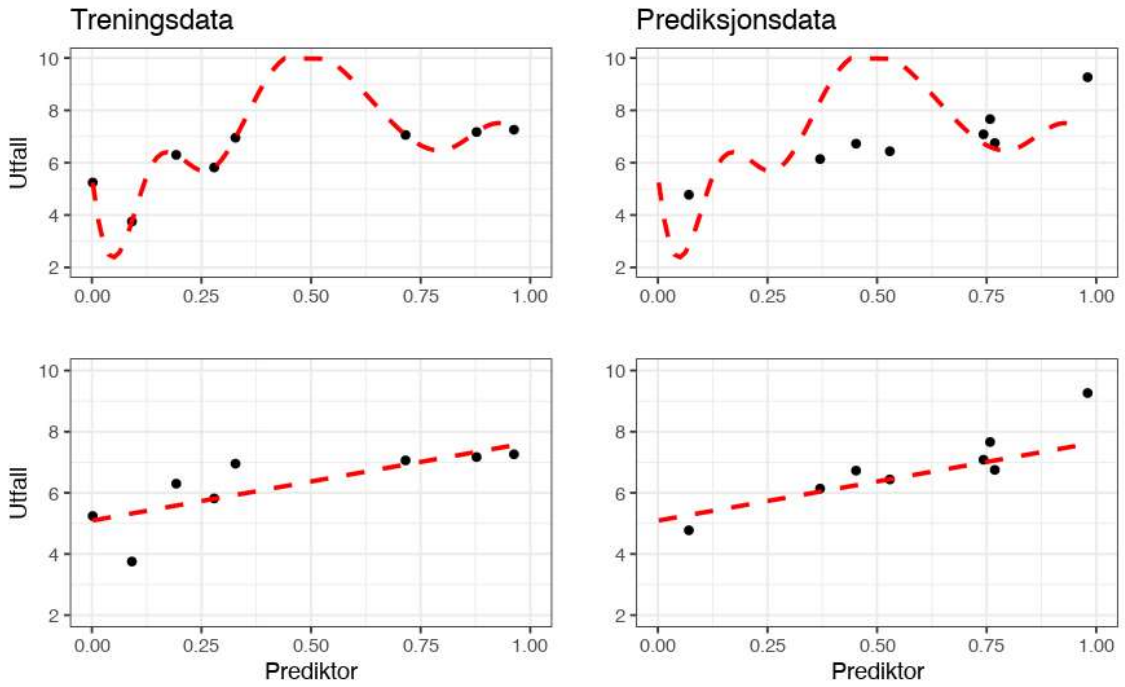
Slike modeller er fleksible og raske å estimere, og ved å legge inn samspillseffekter, terskeleffekter (der den store forskjellen er mellom de over og under en bestemt verdi,

[2] Det vil si at den beregner hvor mye hver observasjon ligger over eller under regresjonslinjen, ganger tallet med seg selv, og summerer over alle punkter.

uavhengig av hvor langt over eller under den de ligger), og polynomer av prediktorene kan man beregne svært komplekse prediksjonsmodeller.

For prediksjonsformål er det begrenset hvor langt det er lurt å øke på med slik kompleksitet. Ettersom vi legger til ytterligere forklaringsledd nøyer vi oss ikke lenger med å avdekke en «glattet systematikk» - vi fortsetter til nærmer vi oss mer og mer en ren deskriptiv beskrivelse av data. Dette er illustrert i Figur 9, der vi tenker oss at vi skal predikere et utfall basert på en enkelt prediktor. I panelet øverst til venstre vises observasjonene vi har gjort som punkter, mens den stiplete kurven viser prediksjonene fra en fleksibel modell som er trent på disse treningsdata-punktene og treffer perfekt, dvs. alle punktene ligger på den stiplete kurven. I panelet øverst til høyre har vi gjort nye observasjoner så punktene er endret, og her viser den stiplete kurven hva prediksjonene våre ville vært hvis vi brukte denne fleksible modellen til å predikere: den treffer relativt dårlig. De to nederste panelene har samme tolkning som de to øverste panelene, men her har man en trent en mye enklere lineær modell. Denne trente modellen føyer seg dårligere til punktene i treningsdata, men denne mer utglattede sammenhengen er mer informativ og treffer betydelig bedre når den skal predikere i nye data.

Problemet er at overfleksible modeller legger seg for tett opp til hva vi ser i treningdata og overvurderer hvor predikerbart utfallet er. Slike «mettede» modeller fanger ikke bare opp *systematiske mønstre* som kan brukes til prediksjon – de beskriver også alle tilfeldigheter som inntraff og antar at disse ville inntreffe på samme måte i fremtidige datasett. Siden det ikke er tilfelle gjør modellen det dårligere i praksis enn en enklere modell på nye data.



Figur 9 – Overgeneralisering. En god maskinlæringsmodell plukker opp robuste og systematiske mønstre på en måte som reduserer forstyrrelsene fra tilfeldigheter i data. Figurene illustrerer hvordan en modell som "treffer perfekt" i treningsdata i praksis er for "godtroende" og tolker tilfeldig støy som systematikk. Det gir kompliserte modeller som ikke passer med nye data.

Dette er bakgrunnen for LASSO (Least Absolute Shrinkage and Selection Operator), en maskinlæringsmodell som «kondenserer» en overkomplisert lineær regresjonsmodell ned til en enkel modell med et mindre knippe prediktorer. I tillegg «regulariserer» den betydningen til disse ulike prediktorene, som vi kan tenke på som en konservativ skepsis: hvis det er et svakt tegn i data på at en bestemt prediktor kan bety veldig mye for utfallet, slutter modellen at denne prediktoren sannsynligvis har *noe* for seg – men neppe *så* mye som det kunne se ut som i testdata – og krymper koeffisienten mot null.

Justeringsparameteren i LASSO avgjør hvor aggressivt LASSO skal forenkle modellen – og dette vil avhenge av kontekst. For å sette et gunstig nivå på denne benytter vi kryssvalidering, som beskrevet ovenfor. Kort fortalt: Vi deler derfor data i flere deler og roterer hvilken del av utfallsdata vi skjuler for modellen før vi ber den gjette på hva de utelatte dataene vil vise. Dette gjør vi for ulike kandidat-innstillinger, og så velger vi den som ser ut til å fungere godt i vårt kontekst.

Fordelen med LASSO er at den er rask å estimere og gir en tolkbar modell i den forstand at vi kan se nøyaktig hvilke prediktorer modellen til slutt bruker og hvor viktige hver av dem er. Den har også bare èn innstillingsparameter, noe som forenkler arbeidet med å justere modellen. Modellen kan ta inn over seg subtile mønstre ved å starte med en modell som inneholder både vanlige lineære effekter, ikke-lineære effekter (polynomer), og interaksjonseffekter mellom ulike variable. Dette gjør at den er en god «referansemodell» som de mer komplekse maskinlæringsmodellene Random Forest og XGBoost kan sammenlignes mot.

3.3.2 Random Forest

Random Forest («tilfeldig skog») er en godt etablert maskinlæringsteknikk som leter etter mønstre på en helt annen måte enn regresjonsmetoden.

Random Forests er bygget opp av mange enklere modeller (kalt trær). Ett enkelt tre deler data sekvensielt på en måte som bryter det opp i små grupper. F.eks.: vi skiller først mellom alle som er over og under 30. Deretter, blant de som er over tretti skiller vi mellom de med og uten tidligere siktelser. Deretter, skiller vi innad blant de med tidligere siktelser etter om de har fullført videregående. Og så videre.

Disse trærne lages av algoritmen på en måte som skaper størst mulig forskjell mellom gruppene (f.eks. høy og lav lovbruddsprevalens). Et tre består av en serie med «grener» der en gruppe splittes opp i undergrupper slik at de ulike grenene samler folk med ulik grad av lovbruddstilbøyelighet. Treet kunne f.eks. oppdage at det er betydelig høyere antall siktelser for de med få grunnskolepoeng, og den leter da frem det antallet grunnskolepoeng som gjør at de to gruppene får mest ulik lovbruddstilbøyelighet. Deretter leter algoritmen videre etter neste variabel den skal bruke til å «kutte» data, og gradvis «gjerder man inn» observasjoner som er like seg imellom. For å predikere plasserer modellen deg i den inngjerdede gruppen du hører hjemme i – og så antar den at du får samme utfall som de gruppemedlemmene den så i treningsdata.

I praksis viser det seg at man får betydelig bedre prediksjoner ved å «dyrke» et stort antall slike trær med litt forskjellige forutsetninger så de ikke blir for like. Ulike trær kan for eksempel gis tilgang til ulike subsett av prediktorene, slik at de finner ulike mønstre og kan utfylle hverandre. Hvert enkelt tre kan være ganske enkelt – men når ulike trær som fanger opp ulike mønstre og nyanser i data kombineres får vi til sammen et detaljert

og robust bilde av mønstrene i data. For å predikere med en slik «skog» av trær (derav navnet Random Forest på modellen) blir hvert tre i skogen bedt om å gi en prediksjon, og så tas gjennomsnittet av disse.

En utfordring med Random Forest sammenlignet med LASSO er at vi har mange flere innstillinger som kan justeres: hvor ulike prediktorer skal hvert tre få? Hvor mange splitter skal hvert tre potensielt kunne inneholde? Hvor mange trær skal vi ha totalt? Hva er det minste antallet observasjoner vi skal tillate i en inngjerdet gruppe av observasjoner? Siden hver av disse (og andre) innstillinger kan gjøres tildels uavhengige av hverandre blir det et stort antall *mulige* kombinasjoner – og i praksis forenkler man ved å angi et sett rimelige verdier for hver innstilling og trekke f.eks. ti tilfeldige slike innstillingskombinasjoner og kjøre disse gjennom den samme type K-fold validering som vi beskrev for LASSO. Dette sikrer ikke at vi finner "beste" innstillinger – men de vil typisk være "OK." Viser resultatene at innstillingene har mye å si for hvor godt modellen predikerer kan det også være nødvendig å angi nye sett med rimelige verdier som ligger "I nærheten" av den beste kombinasjonen vi fant for å se om det er mulig å finne enda bedre innstillinger i dette "området".

En viktig ting å merke seg med Random Forest er at slike modeller alltid vil predikere verdier som er innenfor det som ble observert i testdata siden prediksjonen til et tre alltid er "snittet av det som ble observert for denne gruppen i testdata". Dette betyr at det er noen typer mønstre den ikke vil kunne ta inn over seg. Dersom for eksempel et fødselskull kan observeres til alder 18 i testdata og har betydelig høyere lovbruddstilbøyelighet enn tidligere fødselskull ved både alder 16, 17 og 18, så vil ikke Random Forest kunne ekstrapolere dette videre og predikere mer lovbrudd enn andre fødselskull ved alder 19, 20 og 21.

3.3.3 XGBoost

XGBoost er en finurlig variant av Random Forest som over tid har etablert seg som en av de mest suksessrike modellene i prediksjonskonkurranser[3].

Enkelt sagt er forskjellen at en Random Forest kobler trær «i parallell» mens XGBoost kobler trærne sekvensielt som en lyslenke på et juletre.

Med Random Forest dyrker vi mange trær som hver for seg gir en uavhengig prediksjon basert på det samme testdatasettet. Hvert tre har sine ting det fanger opp og sine ting det overser eller utelater, men i snitt gir skogen oss et godt bilde av helheten.

Med XGBoost bruker vi også trær – men her kobler vi trærne sammen i sekvens. Først trener vi opp et tre som splitter testdata i undergrupper på samme måte som et tre i en Random Forest modell. Målet med gruppene er å skille ut personer med systematiske høye utfallsverdier fra de med systematisk lavere. Deretter sjekker vi hvor godt denne modellen treffer for hver observasjon, og lar neste tre forsøke å finne systematikk i hvordan den første modellen tok feil. Tre nummer to vet med andre ord ikke hva det første treet gjettet – bare hvor feil det tok – og vil lete etter systematikk i disse feilene: hvor de er og hvor store de er. Dermed bruker vi dette andre treet til å *justere*

[3] Dette er konkurranser der man gir fri tilgang på et treningsdata sett og lar team konkurrere i å lage den beste modellen for å predikere et utfall. Modeller som leveres inn får prestasjonen sin målt etter hvor godt de treffer på et (for dem ukjent) sett av utfallsdata, og kan deretter velge å forbedre modellen sin. Når fristen er over blir alle modellene i konkurransen kjørt opp mot nok et usett datasett for å kåre vinnerne.

prediksjonen fra det første treet. Så lager vi oss et nytt tre og forteller dem hvor det forrige treet gjorde feil, slik at dette tredje treet kan justere den justerte prediksjonen vi fikk fra tre nummer to. Og så videre.

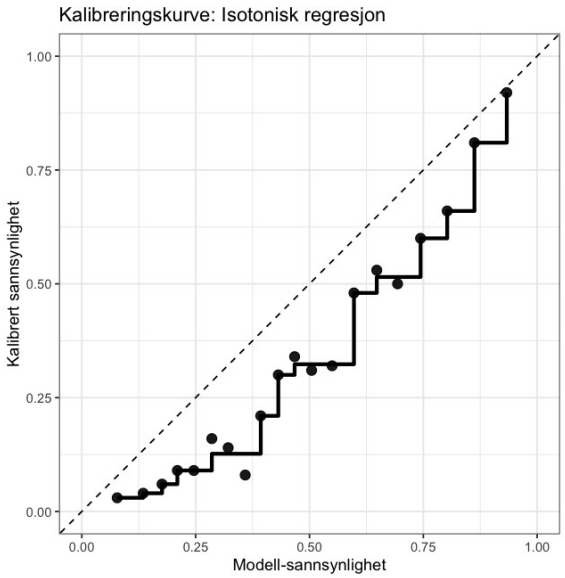
På samme måte som Random Forest vil også XGBoost ha mange innstillingsbrytere, som antall trær, hvor kompliserte hvert tre skal være, hva slags informasjon de skal få tilgang til, hvor mye vekt vi skal legge på hvert tre, og så videre. Til forskjell fra Random Forest vil vi derimot nå kunne få ut prediksjoner som ligger utenfor observerte verdier i test-datasettet. Ta eksempelet der et ungt fødselskull har markant høyere lovbruddsfrekvens enn tidligere fødselskull: hvis vi passer på at treningsdata dekker mer enn ett år slik at ulike fødselskull observeres ved samme aldere, vil XGBoost kunne oppdage at prediksjoner basert på alder må boostes opp hvis de kommer fra ett bestemt fødselskull. De vil da kunne booste opp dette fødselskullet også ved høyere aldere enn det fødselskullet var observert med i testdata.

3.4 Kalibrering

I dette prosjektet er formålet å undersøke om vi kan få gode prognoser på aggregert nivå basert på prediksjoner gjort på individnivå. For at dette skal lykkes må modellanslag stemme overens med virkeligheten. Kalibrering er en prosess som bidrar til dette.

Kalibrering forstås enklest med et konkret eksempel. Vi tenker oss at en maskinlæringsmodell finner 100 personer som hver har 50% sannsynlighet for å oppleve tyveri. Vi forventer da at rundt 50 av disse personene faktisk vil oppleve tyveri. Modellen finner også 100 personer med 25% sannsynlighet. Her forventer vi rundt 25 ofre. Vi henter inn faktiske tall og sjekker, og finner at det var 35 (ikke 50) ofre i den første gruppen og 12 (ikke 25) i den andre. Modellen har da lykkes i å *rangere* risiko og identifisere systematiske forskjeller, men den bommer på det faktiske risikonivået i de to gruppene.

For å håndtere dette problemet estimerer vi kalibreringskurver ved hjelp av historiske data. Vi grupperer personer i data etter hvor ofte modellen tror de vil utsettes for lovbrudd og undersøker hvor ofte de faktisk utsettes for lovbrudd. Hvis alle med risikoanslag rundt 50% typisk er ofre 25% av tiden så justerer vi modellanslagene ned tilsvarende. Dette gir oss en trappelignende kurve som oversetter modellens gjetninger til kalibrerte sannsynligheter (teknisk bruker vi en isotonisk regresjon, se illustrasjon i Figur 10). Disse kalibrerte sannsynlighetene burde i større grad la seg aggregere til pålitelige prognoser på gruppenivå.



Figur 10 – Kalibreringskurve – illustrasjon. Figuren viser hvordan vi justerer modellens prediksjoner for å matche virkeligheten bedre. Den stiplede linjen viser perfekt samsvar. Prikkene viser hva vi faktisk observerer. Dersom modellen traff perfekt ville de svarte prikkene ligge rundt den stiplede linjen. I eksempelet ligger prikkene konsekvent under den stiplede linjen, som betyr at modellen overvurderer faktisk risiko. For eksempel ser vi at der modellen anslår 50% (0.50 på x-aksen) så ser vi i praksis rundt 30% som opplever utfallet (0.3 på y-aksen). Den svarte trappelignende kurven viser den isotoniske regresjonen som oversetter modellprediksjonene til mer presise sannsynligheter. Modellen stiger alltid mot høyre ettersom den isotoniske regresjonen legger til grunn at risikorangeringen er korrekt.

Siden justeringsbehovet kan variere over befolkningsgrupper lager vi kurvene på det nivået vi ønsker å utforme prognoser. Vi lager en nasjonal kurve for totalutviklingen, separate kurver for aldersgrupper, og separate kurver for geografiske område som brukes i senere analyser. Vi beregner videre kalibreringskurver for både sannsynligheter og for siktelsestall, ettersom det også er modelltekniske grunner til at kalibrering kan være viktig også for modeller som ikke anslår sannsynligheter. For analysene på NTU-data er det for få observasjoner til at en fleksibel isotonisk regresjon gir stabile resultater. Her brukte vi en enklere logit-modell til kalibreringen. Denne identifiserer en kalibreringskurve som er S-formet (flater ut når den nærmer seg null eller 1).

3.5 Datamateriale

Maskinlæringsmodeller leter etter robuste systematiske mønstre som kan brukes til prediksjonsformål. Dess mer presist og nyansert vi kan beskrive hver person i data, dess større potensiale har vi for å avdekke subgrupper med særlig høy eller lav lovbruddsrisiko. Dette betyr at vi forventer bedre prediksjoner dess flere datakilder vi kan trekke på. Samtidig er det ikke nok å «kaste alle data inn i modellen»: en viktig kilde til å forbedre prediksjoner er å konstruere bedre prediktorer fra et gitt datamateriale («feature engineering»).

3.5.1 Registerdata

Et viktig formål med dette prosjektet er å undersøke i hvilken grad maskinlæringsmodeller kan gi gode prognoser på kriminalfeltet. Prosjektet la derfor opp til en omfattende samling av ulike registerkilder som dekker både utdanning, stønader,

inntekt og formue, bosted, demografi, lovbruddshistorikk og helse. I praksis viste det seg vanskelig å få tilgang til helsedata gitt prosjektets formål, ettersom registereiere her la til grunn at slike data kun kan gjøres tilgjengelig for *helse*forskning. En oversikt over registerinformasjonen som ble brukt for å predikere siktelser gis i Tabell 2. For mørketallsanalysene var vi begrenset til de informasjonskildene som deltagerne hadde gitt samtykke til å koble på fra registere (Tabell 3).

Tabell 2 – Datakilder brukt til prediksjon av siktelser

Informasjon	Variable
Demografiske kjennetegn	Alder, kjønn, innvandringskategori, landbakgrunn, innvandringsalder, antall søsken, antall eldre søsken, posisjon i søskenflokk (eldst, nest eldst osv), kommune/delområde, sivilstand, antall barn, alder ved første barns fødsel, alder til yngste barn.
Lovbruddsdata	Siktelser: Totalt, per lovbruddsgruppe, og per lovbruddskode: Antall siktelser totalt, tid siden siste, tid siden første, antall siktelser inneværende år/siste to/siste tre/ siste fire/siste fem. Soning: Antall dager sonet (lukket og åpen anstalt separat)
Barnevern	Separat for hjelpetiltak og omsorgstiltak: Antall år med hjelp, alder første hjelp, hjelp inneværende år/tre siste
Inntekt og støtteordninger	Inntektsrang relativt til andre av samme kjønn og alder, mottak av AAP/Uføretrygd, antall år med observert sosialhjelp
Utdanning	Under utdanning, høyeste oppnådde etter BU-kode siffer en og en-og-to, skår på nasjonale prøver, karakterer grunnskole, karakterer videregående, generell studiekompetanse, fravær videregående,
Verneplikt	Avtjent verneplikt (måneder), sesjonstest (kognitivt)
For hver av egne foreldre	Alder ved egen fødsel, alder ved forelders død, forelders

	innvandringsalder og innvandringskategori, inntektsrang (snitt over alder 35-45 relativt til andre av samme kjønn og alder), utdanning (BU-kode 1 og 2 sifret), mottak av AAP, uførepensjon, sosialhjelp, avtjent verneplikt, antall siktelser observert per lovbruddsgruppe, sonet tid på åpen/lukket anstalt

Tabell 3 - Variable brukt for å predikere offerstatus (selvrapportert i NTU og i register)

Informasjon	Variable
Individinformasjon	Kjønn, alder, innvandringskategori/-alder/-landbakgrunn, bostedskommune, sivilstand, antall barn, barn i husholdningen, parstatus, inntekt (KPI-justert til 2015-kroner), under utdanning, høyeste fullførte (første og andre siffer BU-kode), støtteordninger: AAP, uføretrygd, år med sosialhjelp observert
For hver av egne foreldre	Foreldres utdanning (BU-kode 1 og 2 sifret), sosialbakgrunn

3.5.1.1 Konstruksjon av prediktorer

En viktig del av arbeidet med å utvikle en maskinlæringsmodell er å utvikle gode prediktorer. Kort fortalt: skal informasjonen i registerdata kunne utnyttes av en maskinlæringsmodell må vi uttrykke informasjonen i et språk modellen kan forstå og på en måte som gjør det lettest mulig for modellen å se relevante forskjeller mellom folk.

Som en illustrasjon på dette kan vi tenke oss ulike måter å bruke tidligere siktelser. Tidligere siktelser er karakterisert med blant annet en registreringsdato, gjerningsdato, lovbruddsgruppe, lovbruddskode, saksnummer osv. De fleste har null siktelser, noen har veldig mange.

For at modellen skal bruke informasjon om tidligere siktelser må denne gjøres om til numeriske variable. Ved å uttrykke tildels den samme informasjonen på mange ulike måter gjør vi det enklere for modellen å finne frem til informative oppsplittinger av data. F.eks. kan man legge inn informasjon om «antall tidligere siktelser totalt», «antall siktelser siste år», «antall siktelser siste tre år» osv. Man kunne i tillegg angi slike tall for spesifikke lovbruddsgrupper eller for hver lovbruddskode, konstruere en variabel for

«antall ulike lovbruddsgrupper man har vært siktet for», telle «antall medsiktede totalt» eller «andel av tidligere siktelser som var med medsiktede», og så videre. Hver slik variabel gjør det mulig for modellen å sammenligne data langs nye retninger – litt som å se data gjennom ulike «linser» for å finne ut hvilke som synliggjør klare forskjeller.

I praksis er det selvsagt begrensninger på hvor mange variable man kan legge inn, ettersom arbeidet med å tenke ut og implementere prediktorer tar tid. I tillegg kommer implementeringshensyn: prosessorkapasitet og internminne på maskinene begrenser hvor omfattende data det er mulig å bruke og hvilke analyser det er teknisk gjennomførbart å utføre i en gitt tidsramme.

I dette prosjektet har vi valgt en pragmatisk tilnærming. Vi tar utgangspunkt i prediktorene som ble utviklet for å predikere hvilke lovbrytere som sto i fare for å bli skutt de neste 18 månedene og følger i stor grad strukturen på disse for datakilder av en type som begge prosjektene har tilgang til (Heller et al. 2024).

3.5.1.2 Utfallsdata: lovbruddsinformasjon

Vi bruker to typer utfall i prosjektet: antall siktelser (totalt og innen ulike sett av lovbruddsgrupper), og hvorvidt (ja/nei) en person er registrert som offer for et lovbrudd (på tvers av alle lovbruddsgrupper og innen sett av lovbruddsgrupper). Merk at lovgruppene vi sammenstiller er ulike i de to analysene, ettersom offer-analysen benytter lovbruddsgrupperinger som kan gjenfinnes som selvrapporterte lovbruddskategorier i NTU.

Siktelser brytes ned på ulike subgrupper med lovbruddstyper for å undersøke om det er noen subsett av lovbrudd der prognosemetodikken fungerer bedre eller dårligere enn andre. Særlig har vi en bekymring rundt rusmiddelrelaterte siktelser, der omfanget av siktelser i særlig høy grad er et utslag av politiets egne prioriteringer. Ettersom dette er en av de største lovbruddskategoriene i data forventer vi at prognoser potensielt kan treffe bedre i mer spissede lovbruddskategorier (som f.eks. vold).

Lovbrudd i registerdata har ulike dateringer. Et viktig skille i vår sammenheng er forskjellen mellom registreringsår og gjerningsår. Ettersom noen lovbrudd først anmeldes eller avdekkes i ettertid kan disse to avvike.

I prognosearbeidet tar vi utgangspunkt i registreringsår, siden dette er det *kjente* omfanget av lovbrudd ved utgangen av et kalenderår. I arbeidet med mørketall, derimot, ønsker vi å sammenligne predikert *utsatthet* (fra selvrapporterte opplevde lovbrudd i NTU) med predikert *registrering noensinne* av lovbrudd *fra det kalenderåret*. Vi inkluderer derfor alle lovbrudd med dette gjerningsåret så langt frem våre data går.

Når det gjelder data på registrerte ofre opplyste forøvrig SSB via epost under dataleveransen at det er "*viktig å merke seg at kvaliteten på disse dataene er vesentlig dårligere enn for siktelser. Straffesaksregisteret er bygget opp rundt gjerningspersoner og avgjørelser mot dem. For ofre er det ikke bare dårligere kvalitet i fastsettelsen av status («rolle» i saken kan ofte være feil, fornærmet vs. vitne eller pårørende, siktelser baserer seg ikke på rolle med påtaleavgjørelse) men også i utfyllingen av status når det er flere lovbrudd i samme sak. Så når vi snakker om «alle» lovbruddene ofre er utsatt for er det med forbehold om at dette ikke alltid er fullstendig registrert. Vi publiserer ikke statistikk på denne enheten per i dag. Men det vil gi et mer utfyllende bilde enn hvis man bare tar med hovedlovbruddet.*"

3.5.2 Nasjonal Trygghetsundersøkelse

Nasjonal trygghetsundersøkelse (NTU) er et viktig verktøy i arbeidet med å kartlegge befolkningens uro og utsatthet for kriminalitet i Norge. Mens offisiell kriminalstatistikk først og fremst omfatter lovbrudd anmeldt til politiet, gir NTU et bredere bilde ved å inkludere selvrapportert utsatthet – også for lovbrudd som aldri blir kjent for myndighetene.

Historikk og formål

NTU ble gjennomført for første gang i 2021, som et svar på behovet for mer detaljert kunnskap om trygghetsfølelse og skjult kriminalitet i den norske befolkningen. Målet med undersøkelsen er å gi et kunnskapsgrunnlag for beslutningstakere, politi, forskere og sivilsamfunnsaktører, og den supplerer eksisterende kriminalstatistikk ved å fange opp mørketall (Løvgren et al. 2024). Siden ble undersøkelsen gjennomført ytterligere tre ganger, i 2023-2025, og neste gjennomføring blir i 2026.

Undersøkelsen kartlegger utsatthet for ulike typer lovbrudd, uro for kriminalitet, og om de lovbruddsutsatte valgte å anmelde hendelsen. Spesifikke typer lovbrudd som måles er blant annet vold, seksuell vold, digital kriminalitet, og diskriminering.

Organisering og ansvar

NTU ledes av Velferdsforskningsinstituttet NOVA ved OsloMet, i samarbeid med Frischsenteret og ideas2evidence. Prosjektet gjennomføres på oppdrag fra Justis- og beredskapsdepartementet. Dataene samles inn og analyseres av forskerne, og videreformidles gjennom Sikt – kunnskapssektorens tjenesteleverandør. Prosjektledelsen for undersøkelsene er ivaretatt av Mette Løvgren, med en bredt sammensatt prosjektgruppe bestående av forskere fra de tre institusjonen (Høgestøl and Stokke 2024).

Utvalg, representativitet og metodikk

Datainnsamlingene blir gjennomført som en web-basert spørreundersøkelse og spør om utsatthet for lovbrudd foregående kalenderår. 68 000 personer i alderen 16 til 84 år blir trukket ut fra Folkeregisteret og invitert til å delta.

Utvalget blir trukket ved hjelp av stratifisert tilfeldig utvalg, basert på alder, kjønn og geografi. I tillegg blir det gjort et målrettet uttrekk av personer med innvandringsbakgrunn for å sikre representasjon fra grupper med kjent lav svarvillighet. Utvalgene er representative på den måten at de deler samme demografiske karakteristikker som uttrekkene fra Folkeregisteret, men det er verdt å påpeke at det trolig er en skjevhet i hvem som velger å delta. For eksempel har trolig personer som har blitt utsatt for lovbrudd høyere tilbøyelighet til å delta enn personer som ikke er blitt utsatt.

35-37 prosent av de inviterte deltar, og blant samtykkene som blir innhentet er også innhenting av registerdata. Disse følger samtykket og er: “ (…) historiske og fremtidige opplysninger om: pensjonsgivende inntekt, yrke, bosted, høyeste fullførte utdanning, trygder og familieforhold (sivilstand, antall barn, antall personer i husholdningen). I tillegg vil dette omfatte informasjon om dine foreldres høyeste utdanning og deres pensjonsgivende inntekt.” (Høgestøl & Stokke, 2024, s. 119). Tabell 4 viser antall

deltakere, svarprosent, samtykker til registerdata og paneldata samt datainnsamlingsperiode for hver gjennomføring av NTU.

Tabell 4

År målt	Antall deltakere	Svarprosent (uvektet)	Samtykke registerdata (%)	Datainnsamlingsperiode
2020	24163	35,5	74	Mai – August 2021
2022	25420	37,4	70	Mai – August 2023
2023	24991	37	71	Februar – April 2024

Tidligere bruk og betydning i forskning og forvaltning

NTU har på kort tid etablert seg som et viktig datagrunnlag for både akademisk forskning og offentlig forvaltning. Undersøkelsen har blitt brukt i analyser av kriminalitetsutvikling, polititillit, ulikhet i trygghetsopplevelse, og for å belyse utsatthet i marginaliserte grupper. Dataene benyttes av Justis- og beredskapsdepartementet i vurderingen av kriminalpolitiske tiltak, og har blitt brukt i politiets prioriteringer knyttet til vold, hatkriminalitet og digital kriminalitet.

Medieoppmerksomhet og offentlig debatt

NTUs resultater omtales jevnlig i nasjonale medier, spesielt funn som viser høye mørketall for seksuell vold og digitale overgrep. Særlig stor oppmerksomhet har vært rettet mot unges utsatthet, kjønnede forskjeller i voldserfaringer, og tilliten til politiets evne til å håndtere anmeldte saker. Tallene har bidratt til økt politisk bevissthet og debatt rundt forebygging, strafferett og hjelpetjenester. Dataene brukes også i flere pågående forskningsarbeider og i en nylig publisert artikkel (Kotsadam and Løvgren 2025).

3.6 Resultater

3.6.1 Prognoser for siktelser

3.6.1.1 Hvor godt avdekker modellene systematiske forskjeller i lovbruddstilbøyelighet?

Har vi grunn til å tro på modellen vår når den forutsier hvem som vil bli siktet ofte og hvem som ikke vil siktes i det hele tatt?

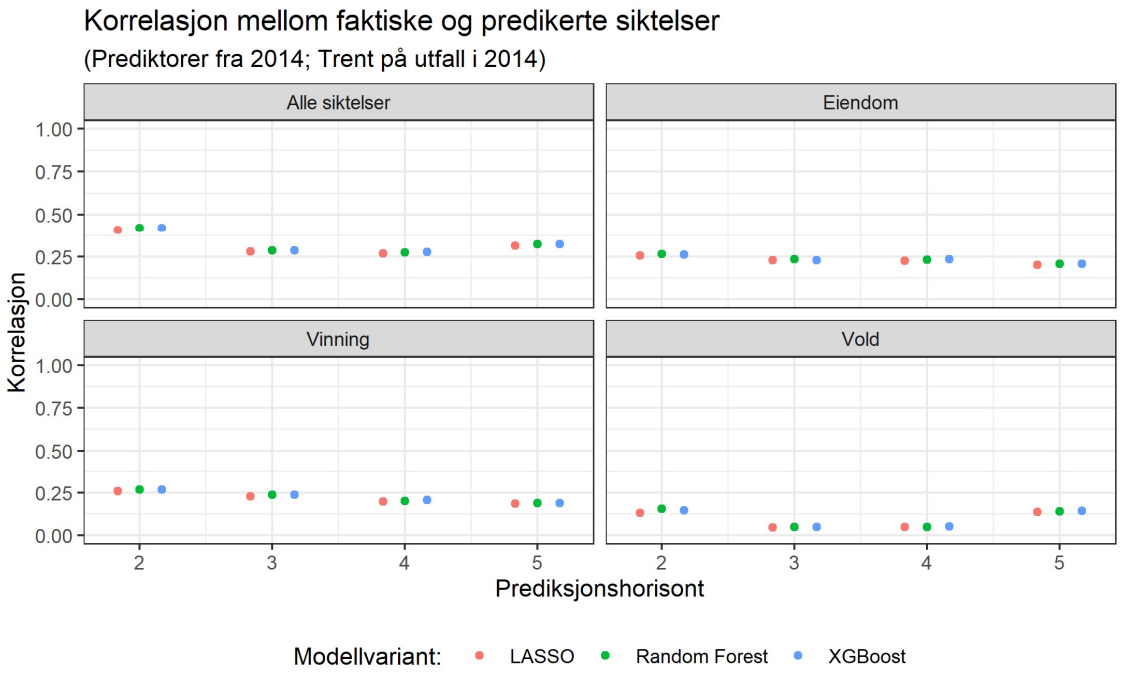
Dette er et spørsmål om samsvar mellom modellens prediksjoner og de faktisk realiserte utfallene: Hvor mye av forskjellene i antall siktelser som vi faktisk ser i en befolkning kunne vi fanget opp i prediksjonene våre allerede 2 til 5 år i forkant? Et mål på dette får vi ved å beregne korrelasjonskoeffisienten mellom prediksjonene og de faktiske utfallene.

Korrelasjonskoeffisienten ligger alltid mellom -1 og 1, og sier oss hvor sterk lineær sammenheng det er mellom to sett med tall[4]. Er tallet null er det ingen lineær sammenheng mellom tallene. Da har vi ingen grunn til å tro at de med høyere predikert antall lovbrudd vil ha flere lovbrudd i fremtiden. Er tallet 1 vil det *alltid* være slik at A *faktisk* har flere lovbrudd enn B hvis dette er prediksjonen. Korrelasjonskoeffisienten kan også bli lavere enn 0, men dette ville bety at modellen gjettet i systematisk feil retning.

I dette prosjektet ønsket vi å sammenligne hvordan ulike maskinlæringsmodeller var i stand til å forutsi, på medium og lengre sikt, både det totale antallet siktelser en person ville bli registrert med og antallet siktelser innenfor ulike subgrupper av lovbrudd. For å undersøke dette antok vi at vi stod i 2015 og hadde tilgang på data fra 2014 som vi kunne predikere på bakgrunn av. Dette året ble valgt siden vi vurderer modeller fra 2 og opp til 5 års prediksjonshorisont. Med 5 års horisont gir 2014-data oss en prediksjon for 2019 – det siste året som er upåvirket av COVID19 og pandemitiltakene. Vi antok videre at vi brukte den *ferskeste* modellen vi kunne for hver horisont. Det vil si at en modell med to års prediksjonshorisont ville bli trent på utfall i 2014 med prediktor-data fra to år tidligere, og denne modellen vil deretter bli brukt på 2014-prediktorer for å predikere utfall i 2016. Modeller med 5 års horisont vil da bli trent med utfall i 2014 og data fra 5 år før det (2009), og deretter brukt på 2014-data for å predikere utfall i 2019.

Etter å ha trent en modell og predikert faktiske utfall i «fremtiden» kobler vi på de faktiske utfallene som inntraff og beregner korrelasjonen mellom predikert og faktisk kriminalitet. Dette gir oss punktene i Figur 11, som har ett panel for hver av de fire lovbruddsgruppene, og som innad i hvert panel angir prediksjonshorisont på x-aksen og modelltype med punkt-farge.

[4] Vi snakker her kun om lineære sammenhenger ettersom det er mest relevant for en sammenligning av predikert og faktisk antall siktelser. Med ikke lineære sammenhenger vil korrelasjonskoeffisenten kunne være null selv om vi nøyaktig kan forutsi det ene tallet basert på det andre. F.eks. kan man tegne en sinuskurve som bukter seg over og under en horisontal linje og i snitt ligger på den horisontale linjen. Korrelasjonskoeffisienten ville være null siden du ikke entydig får høyere eller lavere verdier ved å bevege deg bortover kurven, men kurven ville være prefekt predikerbar.



Figur 11 – Korrelasjon mellom predikert og faktisk kriminalitet – Etter horisont, modelltype og lovbruddsgruppe.

Det første vi kan merke oss er at *modell-typene gjør det omtrent like godt:* For hver prediksjonshorisont har vi tre fargede prikker (LASSO, Random Forest og XGBoost), men man skal tett inn på figuren for å finne forskjeller. I det store og hele ser modellene ut til å fange opp omtrent like mye av den genuine variasjonen i data.

Dernest ser vi at *prediksjonshorisont har begrenset betydning*. Vår gjetning ville vært at modellene skulle gjøre det betydelig dårligere dess lengre frem i tid de skulle predikere, men dette kan ikke påstås å være noe slående og konsistent mønster. For «alle siktelser» treffer alle tre modellene bedre på 2 års horisont enn høyere horisonter, men i det store og hele er forskjellene over prediksjonshorisonter nokså små og usystematiske.

En ytterligere sammenligning vi kan gjøre er mellom *ulike lovbruddsgrupper*. Her ligger korrelasjonene stort sett rundt 0.25, men for vold ser de ut til å ligge betydelig under.

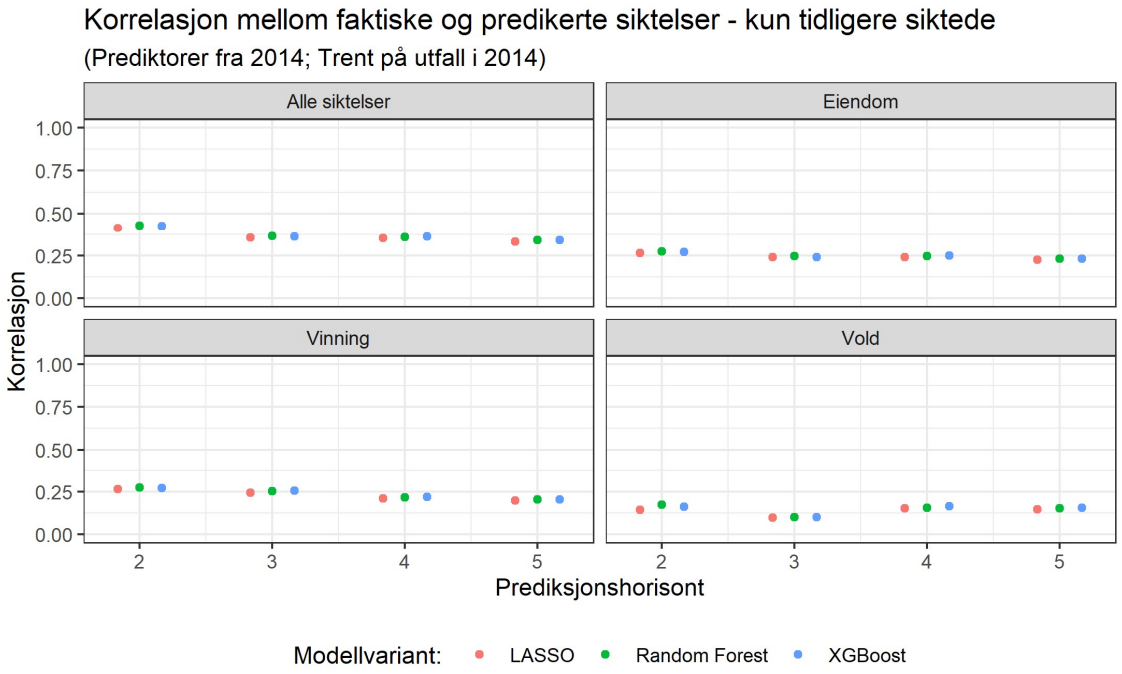
Slike korrelasjoner kan også sammenlignes med hva vi ville forvente basert på tidligere forskningslitteratur. Tidligere meta-analyser av prediksjon av kriminalitet viser konsekvent at prediksjonskraften ligger på et moderat nivå, men det må påpekes at disse analysene ser på prediksjoner for tidligere straffede individer. For generell tilbakefallskriminalitet finner Bonta et al. (1998) en gjennomsnittlig effektstørrelse på r = .27 (64 studier), mens Gendreau et al. (1996) rapporterer r = .26 basert på 131 studier av voksne lovbrytere. Også for voldelig tilbakefall ligger korrelasjonene på samme nivå (r = .24; Bonta et al. 1998). I en nyere meta-analyse finner Fazel et al. (2012) at risikovurderingsverktøy predikerer voldelig kriminalitet med en median AUC på 0.73 og generell kriminalitet med AUC på 0.68. Dette tilsvarer korrelasjoner på omtrent 0.25–0.30.

Spesifikke oversikter over seksuallovbrudd og vold viser samme mønster. Hanson and Morton-Bourgon (2009) analyserer 118 studier av seksualforbrytere og finner en

gjennomsnittlig AUC på .71 for tilbakefall, mens Yang et al. (2010) rapporterer at ni ulike risikovurderingsverktøy predikerer voldelig tilbakefall med AUC-verdier mellom 0.62 og 0.71. På tvers av utfallstyper og populasjoner finner altså forskningen at prediksjon av fremtidig kriminalitet blant tidligere straffede sjelden overstiger korrelasjoner rundt 0.2–0.3.

Denne graden av prediksjonskraft er ikke unik for kriminalitet. Også i andre samfunnsvitenskapelige domener, som helse, utdanning og arbeidsmarked, viser forskningen at fremtidige utfall sjelden lar seg predikere med korrelasjoner mye høyere enn 0.3 (Meeh 1954; Obermeyer and Emanuel 2016). Kleinberg et al. (2015) omtaler dette som et "prediction policy problem": selv moderate forbedringer i prediksjon kan ha stor praktisk verdi når de anvendes i politiske eller administrative beslutninger. Våre resultater på rundt 0.25 bør derfor forstås i lys av dette bredere mønsteret.

Våre resultater med korrelasjoner rundt 0.25, oppnådd i en bred befolkningskohort der mange aldri blir siktet for noe lovbrudd, fremstår derfor som bemerkelsesverdige. Når vi beregner korrelasjonene mellom prediksjon og faktisk utfall kun for de som allerede hadde siktelser i data får vi tall (for «alle siktelser») rundt 0.35 (Figur 12).

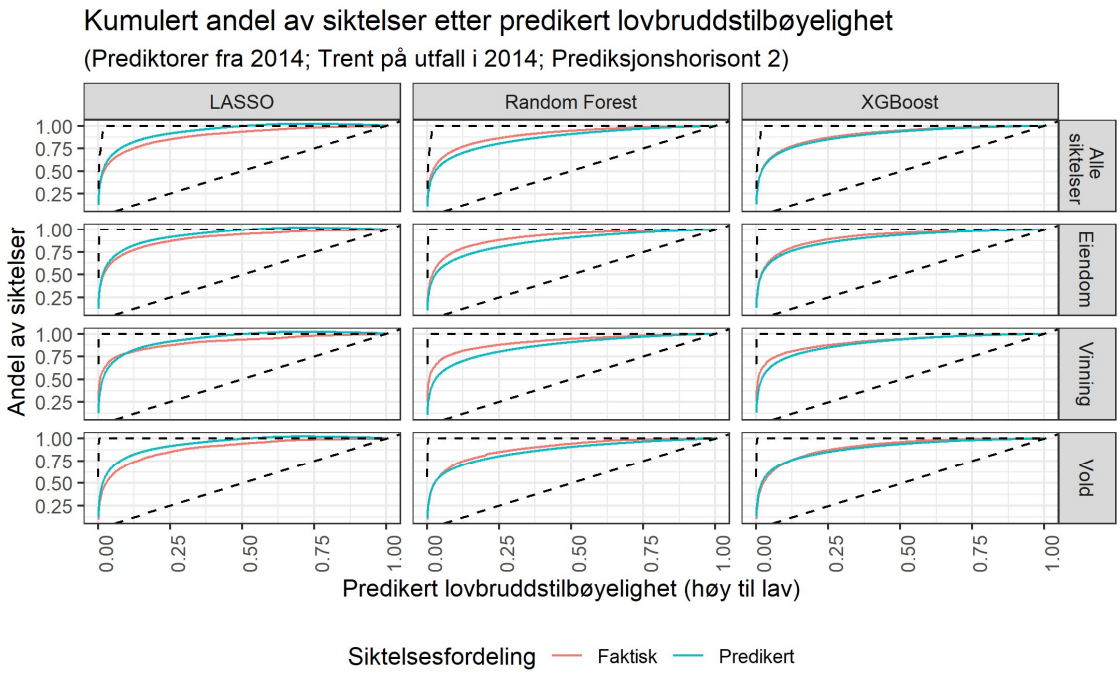


Figur 12– Korrelasjon mellom predikert og faktisk kriminalitet – tidligere siktede – Etter horisont, modelltype og lovbruddsgruppe.

Vi kan også vurdere hvor godt prediksjonene fanger opp forskjellene i fremtidige siktelsestall på måter som ikke avhenger av linearitet. Vi kan f.eks. bruke en trent maskinlæringsmodell til å predikere antall siktelser på person-nivå, og deretter sortere befolkningen fra de med flest til færrest predikerte siktelser. Vi kan deretter tegne en kurve som viser i hvilken grad de *predikerte* lovbruddene er konsentrert blant de med høyest predikert lovbruddstilbøyelighet. Deretter kan vi bruke de faktiske siktelsene som

ble realisert til å vise i hvilken grad de *faktiske* lovbruddene var konsentrert i like stor grad blant disse.

Slike kurver er tegnet for ulike kombinasjoner av modeller og lovbruddsgrupper i Figur 13. For å lette tolkningen legger vi inn to stiplede referanselinjer. På den ene siden kan vi tenke oss en modell som ikke klarer å predikere i det hele tatt: Det er *ingen* økt lovbruddsfrekvens for de personene modellen tilskriver høy lovbruddstilbøyelighet. Undersøker vi faktiske lovbrudd vil disse være helt jevnt fordelt utover x-aksen vår, og vi ville fått en rett, skrå linje med 45 graders helning: de 10% «mest» lovbruddstilbøyelige ville stå for 10% av de predikerte siktelsene, siden de i praksis ville være som et tilfeldig utplukk av befolkningen. På den andre siden kan vi tenke oss en modell som predikerer helt perfekt. Da vil de faktiske siktelsestallene være helt identisk med de vi predikerte, så for å tegne denne kurven kan vi sortere befolkningen etter deres *faktiske* siktelser og bruke denne sorteringen når vi tegner fordelingen. Disse to referansekurvene gir oss dermed hele spennet av hva vi kan forvente, og lar oss se hvor modellens faktiske prestasjon ligger i forhold til tolkbare ekstremtilfeller.

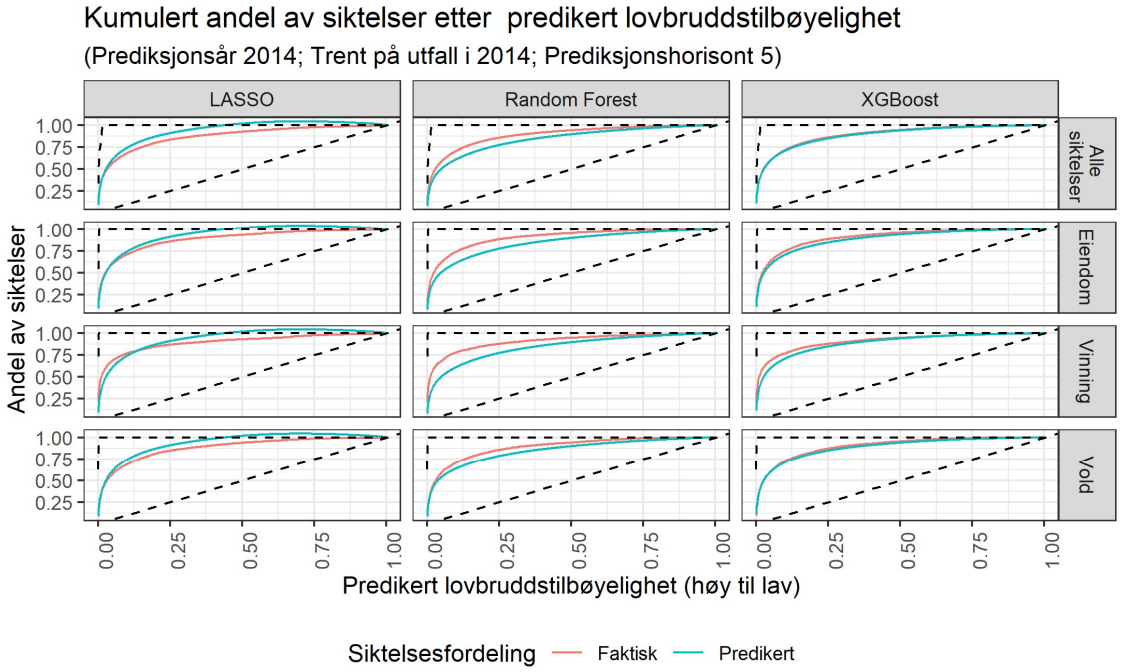


Figur 13 – Konsentrasjon av faktisk lovbrudd etter predikert lovbruddstilbøyelighet. Modellene har to års prediksjonshorisont og er trent på prediktorer fra 2012 og utfall fra 2014 og deretter brukt til å predikere utfall i 2016 basert på prediktorer fra 2014.

Figur 13 viser at modellene gjennomgående fanger opp en betydelig skjevfordeling i lovbruddstilbøyelighet: på tvers av modeller og lovbruddsgrupper ser vi at de 12.5% mest lovbruddstilbøyelige i praksis stod for rundt 75% av alle siktelsene som faktisk ble observert i etterkant, og de predikerte konsentrasjonskurvene ligger godt unna 45 graders linjen – men også et solid stykke unna den vi ville fått med en perfekt prediksjonsmodell.

Sammenligner vi de røde og blå kurvene finner vi noe avvik, som betyr at modellene tar feil i sine anslag på hvor mange flere lovbrudd de med høy predikert lovbruddstilbøyelighet ville stå for. Blant de LASSO modellen tror vil ha flest siktelser ser vi at den tror disse vil stå for en høyere andel av lovbruddene enn de faktisk gjorde i praksis (den blå kurven ligger over den røde). Random Forest gjør det motsatte. XGBoost fremstår her som den best kalibrerte, ved at de røde og blå kurvene sammenfaller i størst grad: fordelingen som ble predikert sammenfaller i størst grad med den fordelingen som ble observert.

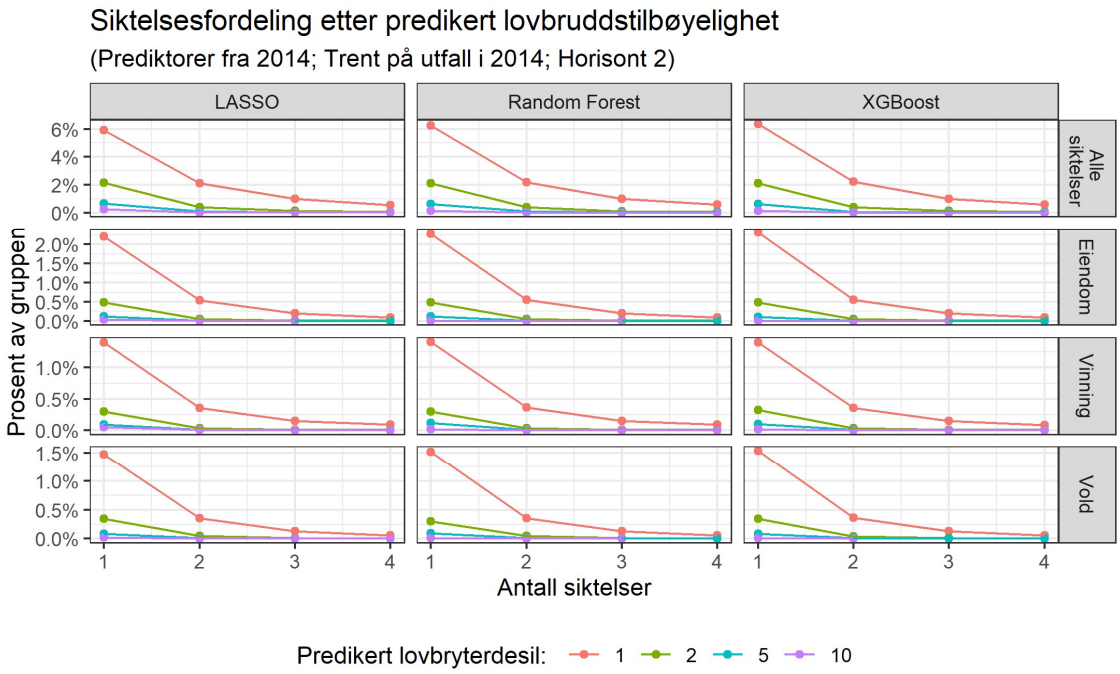
I Figur 14 viser vi de samme typene kurver for modeller med 5 års prediksjonshorisont. Noe overraskende ser vi at modellene fortsatt klarer å predikere hvem som kommer til å stå for brorparten av fremtidige siktelser – de 12.5% med høyest *predikert* lovbruddstilbøyelige står også nå for rundt tre fjerdedeler av alle fremtidige siktelser. Det kan også være verdt å nevne kort at den predikerte konsentrasjonskurven for LASSO her får en pukkelform og faktisk faller mot slutten. Dette skyldes at LASSO-modellen ikke sikrer at de predikerte siktelsene er positive, og i praksis predikerer *negative* siktelser for de mest lovlydige i befolkningen.



Figur 14– Konsentrasjon av faktisk lovbrudd etter predikert lovbruddstilbøyelighet. Modellene har fem års prediksjonshorisont og er trent på prediktorer fra 2009 og utfall fra 2014 og deretter brukt til å predikere utfall i 2019 basert på prediktorer fra 2014.

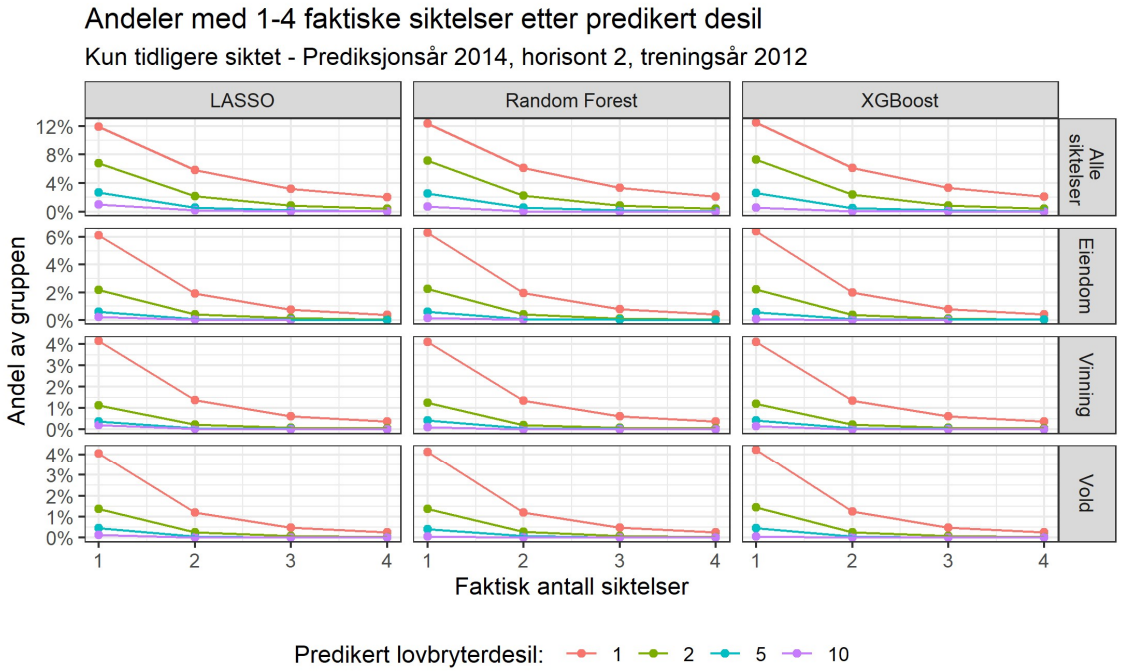
Korrelasjonskoeffisienten og konsentrasjonskurvene viser at modellene fanger opp reell systematikk i hvem som vil og ikke vil få siktelser i fremtiden, men de er samtidig nokså abstrakte. Vi deler derfor også befolkningen inn i ti grupper – desiler – etter deres *predikerte* lovbruddstilbøyelighet. Deretter bruker vi de *faktiske* siktelsene og ser hvor mange som mottok henholdsvis 1, 2, 3 og 4 siktelser i hver desil (Figur 15). Igjen ser vi at alle modellene er i stand til å fange opp mye: Sammenligner vi de to desilene med

høyest predikert lovbrytertilbøyelighet er det gjennomgående tre eller flere ganger så hyppig å ha en siktelse i desil 1 sammenlignet med desil 2.

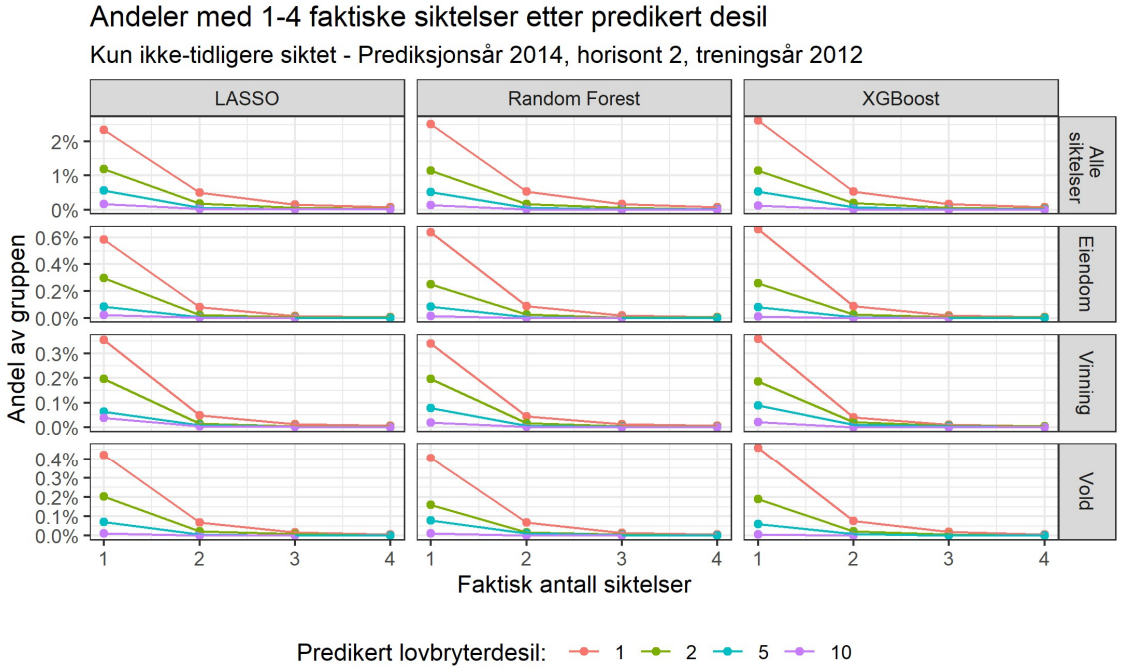


Figur 15 - Antall med 1-4 observerte siktelser etter predikert lovbruddsdesil

En mistanke kunne være at prediksjonene i stor grad er drevet av gjengangere: modellen får tilgang til detaljert informasjon om tidligere siktelser som var observert frem til prediksjonsåret. Vi tegner derfor Figur 15 separat for de med og uten siktelser i prediksjonsdataene. Denne øvelsen viser at modellene også klarer å skille i betydelig grad innad i gruppen både med og uten kriminelle siktelser, selv om andelene med siktelser i den mest lovbruddstilbøyelige desilen er betydelig høyere for de med (Figur 16) enn for de uten (Figur 17) tidligere siktelser.

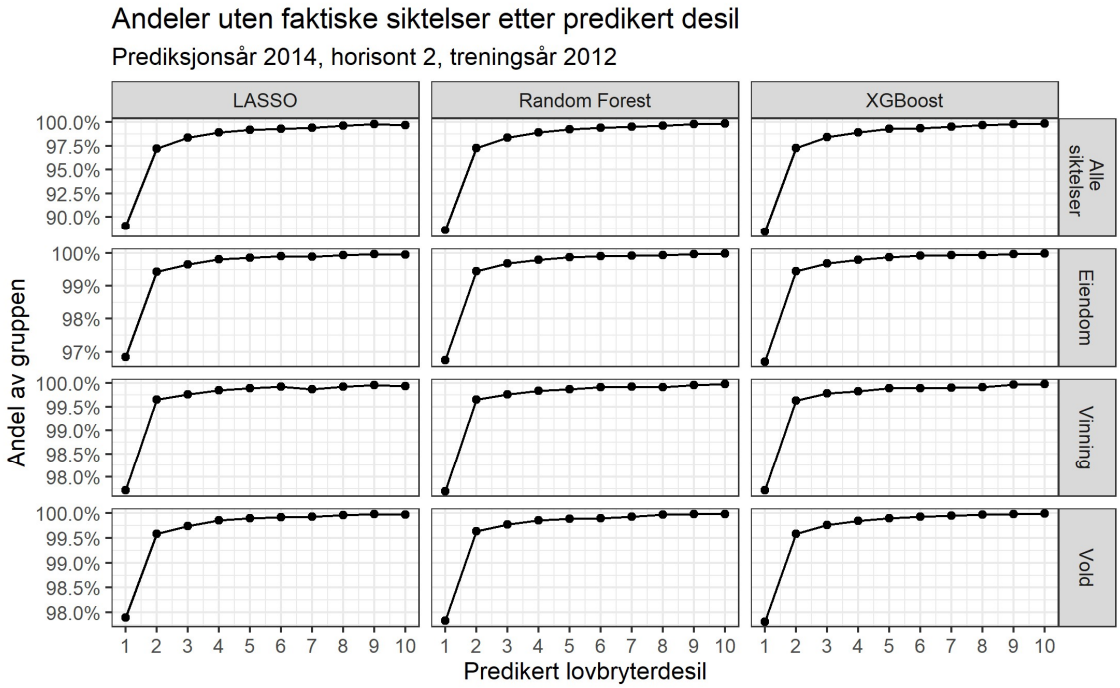


Figur 16 - Andeler med 1-4 siktelser - tidligere siktede.



Figur 17 - Andeler med 1-4 siktelser - tidligere ikke-siktede.

Avslutningsvis er det verdt å minne om at det også er klare begrensninger i hvor godt modellene predikerer. Dette blir tydelig hvis vi snur på flisa og spør hvor mange som *ikke får siktelser* i de ulike desilene vi laget: Selv i desilen med høyest predikert lovbrytertilbøyelighet er rundt 90% uten noen form for siktelser og ser vi for eksempel kun på siktelser for vold er hele 98% i den høyeste desilen uten noen faktisk siktelse for slike forhold i utfallsåret (Figur 18).



Figur 18 - Andeler uten siktelser etter predikert lovbryterdesil

3.6.1.2 Er prediksjonsmodeller ferskvare?

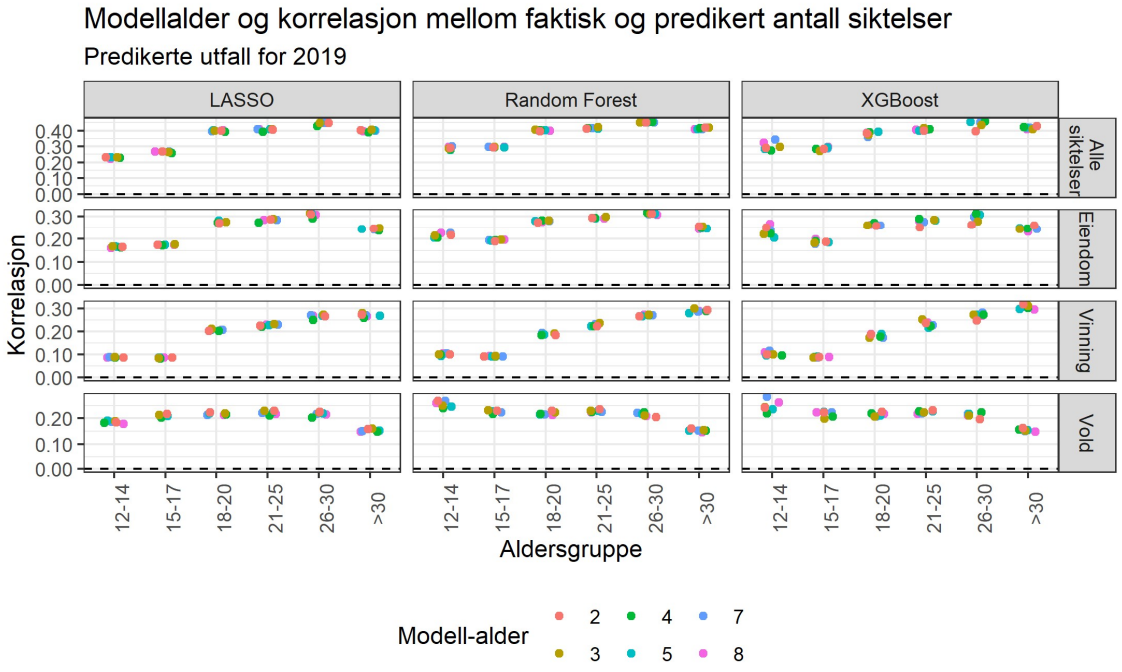
En maskinlæringsmodell trenes ved å lære seg mønstre i et datasett, og den predikerer på nye data ved å finne de utfallsverdiene som ville skape de samme mønstrene der. Dette legger implisitt til grunn at mønstrene er stabile og like på tvers av datasettene. I vårt tilfelle antar vi med andre ord at mønstre vi finner i data fra ett år også vil være tilstede i senere år, men dette trenger ikke være tilfelle.

Det enkleste eksempelet på dette er plutselige omveltninger, som COVID-19 og pandemitiltakene. Når det innføres krav om sosial isolasjon og utesteder og restauranter holdes stengt ville vi anta at det ble færre alkohol-utløste slagsmål og mindre vold i bykjerner på helgene enn det man ville forventet basert på tidligere års data. Men også mer subtile og umerkelige endringer vil gradvis inntreffe over tid, og det er et åpent spørsmål hvor raskt dette skjer og i hvilken grad det påvirker kvaliteten på prediksjonene våre.

For å undersøke dette valgte vi oss utfallsåret 2019 og trente modeller med to års horisont på utfallsdata fra tidligere år (2014-2017) som vi deretter brukte til å predikere på bakgrunn av 2017 prediktorer. Vi kan da se om det er synlig og merkbar forskjell i hvor godt disse prediksjonene treffer.

Igjen kan vi starte med å se på korrelasjonene, denne gangen brutt ned til aldersgruppenivå (Figur 19). Figuren viser ett panel for hver kombinasjon av modell og utfall, og i hvert panel angir vi aldersgruppen på x-aksen og plotter hvor høyt det faktiske utfallet korrelerte med prediksjonene for observasjonene i denne gruppen. Siden vi har laget prediksjoner med modeller av ulik alder har vi flere prikker – en for hver modellalder.

Figuren viser at ferskhet i liten grad var en problemstilling i denne tidsperioden. Modeller av ulik alder predikerer omtrent like godt, og dette er tilfelle i alle aldersgrupper og for alle modeller og utfallskombinasjoner. Derimot avdekker figuren at det er systematisk variasjon i hvor godt vi forklarer lovbrudds-variasjonen ved ulike aldre, og at dette varierer med lovbruddstype. Ønsker vi å predikere alle siktelser gjør vi dette betydelig bedre for de som er 18 eller eldre enn vi gjør for de som er yngre. Det samme mønsteret holder seg for kategoriene eiendom og vinning, mens det ikke er tilfelle for vold – der modellene gjør det omtrent like godt for alle aldersgruppene.



Figur 19 – Korrelasjon mellom utfall og prediksjon etter alder på modellen.

3.6.1.3 Lykkes prognosene i å predikere fremtidig lovbruddsutvikling?

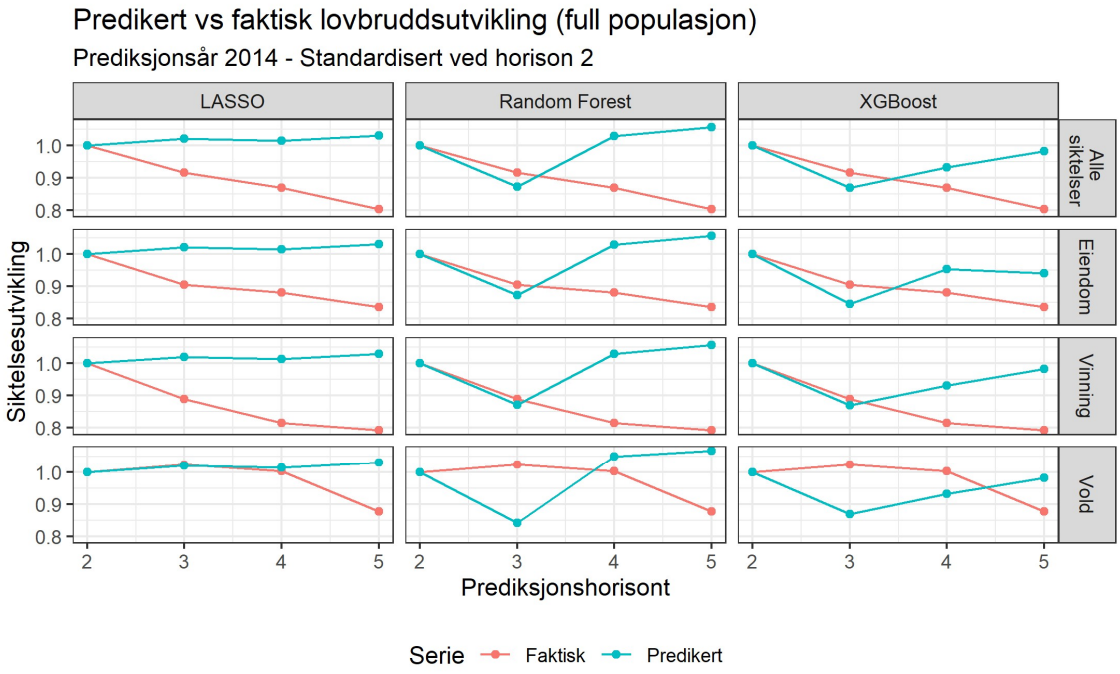
Som vist over fanger modellprediksjonene opp en betydelig del av den faktiske variasjonen i lovbruddstilbøyelighet som finnes i en befolkning, men dette betyr ikke nødvendigvis at modellene vil gi gode prognoser på omfanget og nivået av fremtidige siktelser for en befolkningsgruppe.

For å undersøke dette tenker vi oss igjen at vi står i 2015 og ønsker å bruke prediktordata fra 2014 til å predikere 2, 3, 4 og 5 år fremover i tid. Igjen bruker vi de ferskeste modellene vi kan trene på dette tidspunktet – de som bruker «sist tilgjengelig» utfallsdata som er fra 2014.

På denne måten får vi èn prediksjon for hver person for hvert av fire fremtidige år. For å lage en prognose på gruppenivå følger vi prosjektplanen og beregner – for hvert år - det gjennomsnittlige antallet predikerte siktelser for medlemmer i gruppen. For å fokusere på trend i utviklingen skalerer vi de to seriene (predikert og faktisk) med første tall i serien[5].

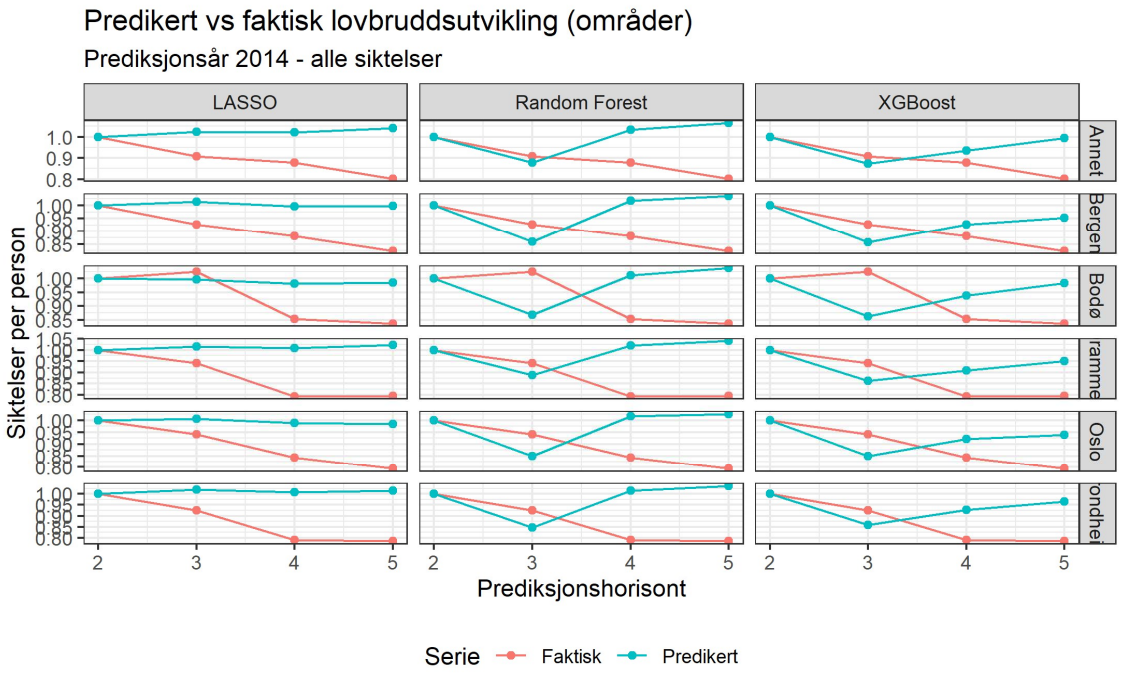
[5] Kurvene kan også tegnes med absolutt nivå uten å skalere etter første verdi, men for de mindre lovbruddskategoriene gir dette en stor diskrepans i nivåer av modelltekniske grunner.

Gjør vi dette for hele populasjonen får vi Figur 20, som er et nedslående resultat: LASSO predikerer et konstant siktelsesnivå gjennom hele perioden, mens den faktiske faller med 10-20% gjennom perioden for alle lovbruddsgruppene. Random Forest og XGBoost gir begge en nedgang fra år 2 til 3 etterfulgt av oppgang, mens de faktiske faller år etter år. Resultatet tilsier at denne prognosemetoden ikke har praktisk verdi som prognoseverktøy.

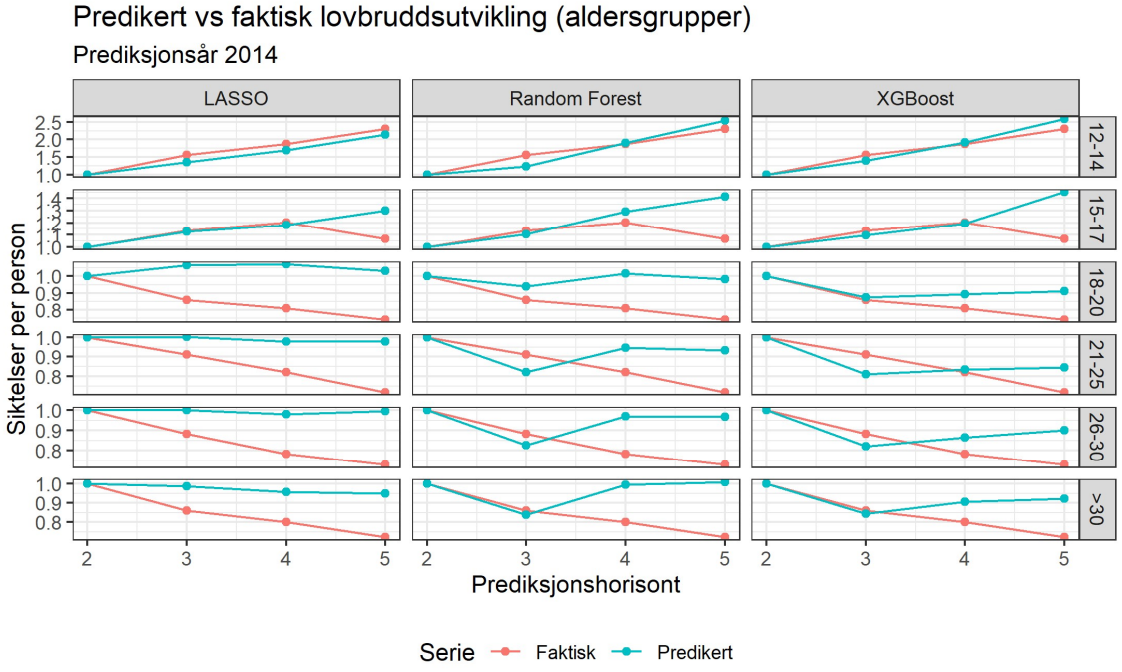


Figur 20 - predikert vs faktisk lovbruddsutvikling - full populasjon – ulike lovbruddsgrupper

I Figur 21 og Figur 22 viser vi også sammenligning av prognoser og faktiske tall for ulike aldersgrupper og større norske byer. Disse resultatene viser – i tråd med de for «alle siktelser» at fremgangsmåten prosjektet skulle undersøke ikke lar seg bruke til å lage informative prognoser for lovbruddsutviklingen.



Figur 21- predikert vs faktisk lovbruddsutvikling - ulike byer – alle siktelser



Figur 22- predikert vs faktisk lovbruddsutvikling - ulike aldersgrupper – alle siktelser

Det kan være ulike grunner til at disse metodene treffer dårlig.

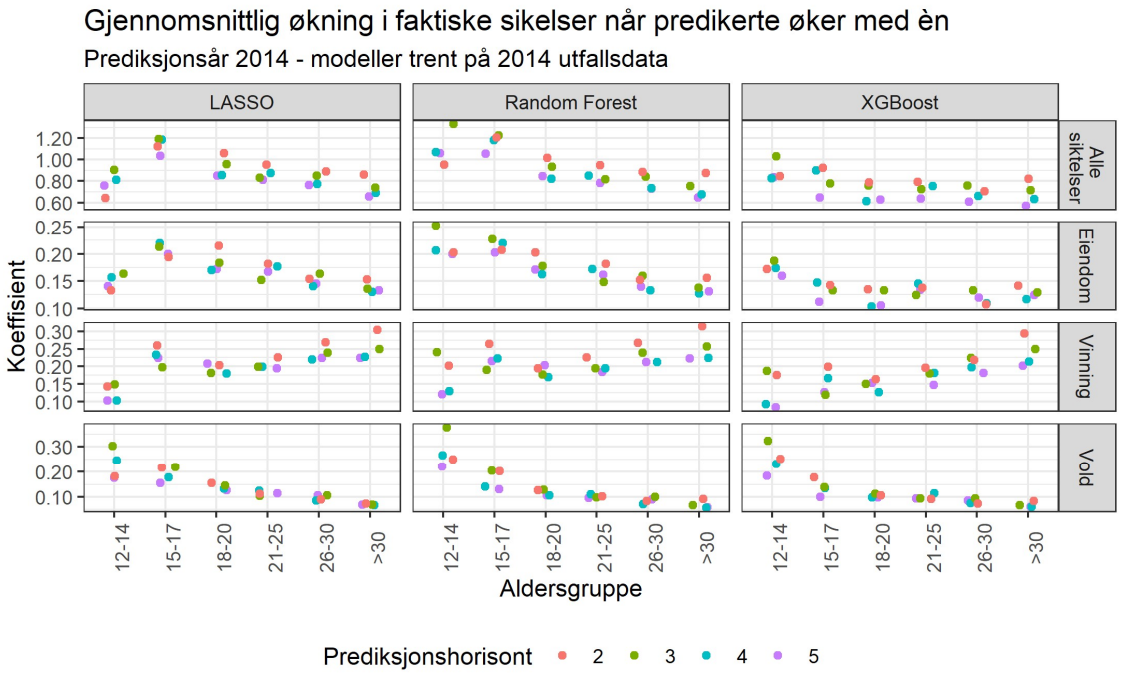
For det første legger strategien vår til grunn at det er *komposisjonsendringer* som driver lovbruddstrender, og det på en nokså subtil måte: Vi bruker prediktorer målt i ett og samme år til å predikere antallet siktelser for hver person i hvert av fire ulike år i fremtiden. Dette gir oss en predikert utvikling over fireårsperioden for hver person i

utvalget, og utviklingen på total eller gruppenivå er bare summen av slike «individ-baner». Får vi flere med «stigende forventede baner» inn i befolkningen (for eksempel 16-åringer med tung barnevernshistorikk og tidligere lovbrudd) vil dette trekke totalprognosene i retning av sterkere vekst i siktelser. Får vi flere med «fallende forventet bane» vil dette trekke prognosene i motsatt retning.

Denne antagelsen om at trender utvikler seg likt for to personer med samme prediktorverdier i ulike kalender år er trolig feil. Selv om vi beskriver en person langs et stort antall observerte registerkjennetegn vil det være ytterligere forskjeller vi ikke har data på, og i tillegg vil personene leve i ulike tider med ulike samfunnsmessige påvirkninger og holdninger, ulik prioritering av saksområder fra politisk og politihold osv. Det kanskje mest åpenbare eksempelet på dette er siktelser for illegale rusmidler, der politiets prioritering av avdekningsvirksomhet i stor grad er bestemmende for siktelsestallene som produseres. På slike felt vil utviklingen i siktelser tildels skyldes endringer i innsatsen, som da daværende Riksadvokat Busch i 2014 og 2015 ga beskjed om endret prioritering og retningslinjer for slike saker[6], hvorpå tallene falt gradvis i mange år. Et annet eksempel kan være trakassering og hatbudskap på sosiale medier, der forekomsten i historiske data var lav fordi teknologiene ikke var rukket å bli populære og utbredt enda. Da vil ikke en modell trent på gamle data være i stand til å predikere korrekte siktelsestall i en senere periode.

En annen feilkilde kan være at modellen fanger opp mer av den fremtidige variasjonen for noen grupper enn andre. Dette undersøker vi i Figur 23, som viser en tydelig variasjon i signalverdien av predikerte siktelser på tvers av aldersgrupper: hvis LASSO modellen sier at en i 15-17 års alderen har 1 siktelse mer enn en annen i samme aldersgruppe er den typiske forskjellen mellom 1 og 1.2. Dersom samme modell gjetter på samme forskjell mellom to personer i aldersgruppen 26-30 er den faktiske forskjellen bare rundt 0.7-0.8. Figuren viser også at det ikke er noen åpenbar sammenheng mellom prediksjonshorisont og signalverdi: for vold gjør modellen med to års horisont det betydelig dårligere enn modellen med 3 års horisont, for eksempel.

[6] Rundskriv 2/2014 påpeker at det ikke skal føres separate siktelser for bruk, besittelse og erhverv i enkle narkotikasaker, mens rundskriv 1/2015 ber om at innsatsen på dette saksfeltet «i større grad rettes mot avdekking av bakmenn»



Figur 23- Signalverdi av predikerte lovbrudd. Figuren viser regresjonskoeffisienter fra lineære regresjonsmodeller, og uttrykker den gjennomsnittlige forskjellen i faktiske siktelser når maskinlæringsmodellen anslår at noen vil ha én siktelse mer enn en annen person.

I tillegg kan det som nevnt tidligere være mer tekniske forhold knyttet til modellkalibrering som gjør at de aggregerte prediksjonene gir misvisende nivå – og disse forholdene kan muligens slå ut ulikt i ulike modell-treninger.

3.6.1.4 Hvilke datakilder er mest prediktive?

Ettersom prognosene ikke viste seg å være tilstrekkelig informative til å brukes i praksis anså vi det ikke som formålstjenlig å gå videre med analyser som lagde prognoser på bakgrunn av mer begrensede sett med datakilder.

3.6.2 Analyser av offer og mørketall

Ovenfor brukte vi maskinlæringsmodeller for å predikere *antall siktelser,* en størrelse som alltid vil være null eller høyere. Her bruker vi maskinlæringsmodeller for å predikere *sannsynligheten for at en hendelse inntreffer*, en størrelse som alltid vil ligge mellom 0 (helt umulig) og 1 (helt sikkert).

Totalt predikerer vi sannsynligheten for tre ulike typer hendelser:

1. Sannsynligheten for å *oppleve* et bestemt type lovbrudd i løpet av et kalenderår. Til dette formålet slår vi sammen data fra to runder med Nasjonal Trygghetsundersøkelse der folk svarer på om de har opplevd ulike typer hendelser som å slås med åpen hånd, og der vi har plassert disse svarene inn i de korrekte lovgruppene hendelsene omfattes av. Konkret så bruker vi SSB sin standard for lovbruddstyper (https://www.ssb.no/klass/klassifikasjoner/146/versjon/575/koder) der utsatthet for noen type lovbrudd i NTU korresponderer med kategoriene 1-5 (Eiendomstyveri, Annet vinningslovbrudd, Eiendomsskade, Vold og mishandling,

Seksuallovbrudd) og dermed ekskluderer Rusmiddellovbrudd, Ordens- og integritetskrenkelse, Trafikkovertredelse og kategorien Annet lovbrudd. For de ulike lovbrudds-kategoriene bruker vi kategori 1 for Vinning, kategori 5 for seksuell vold, og kategori 4A for vold. Dette gir oss utfallet vi er interessert i, som vi deretter predikerer på bakgrunn av de registerdataopplysningene NTU-deltagerne har samtykket i at vi kan koble på.

2. Sannsynligheten for å *rapportere* et bestemt type lovbrudd dersom man opplever det. Også til dette formålet bruker vi data fra de to rundene med NTU, men her kan vi kun bruke et undersett av datamaterialet. Bare de som har opplevd et lovbrudd står overfor valget om å rapportere det eller ikke, så NTU-deltagere som ikke har vært utsatt for lovbrudd utgår fra datamaterialet disse modellene bruker. For analyser av «offer for et lovbrudd (hvilke som helst)» gjør dette at antallet observasjoner faller fra 37 100 til 13 200, men for mer sjeldne offerkategorier faller det betydelig mer (vold: 4 000, seksuell vold: 2 500).
3. Sannsynligheten for å *være registrert* med et bestemt type lovbrudd. Dette er et spørsmål vi kan besvare med registerdata. For å gjøre analysene mer sammenlignbare med de som bruker NTU-data har vi i denne analysen valgt å kun benytte den typen registerinformasjon som NTU-deltagerne samtykket i at NTU-dataene kunne kobles til.

Teoretisk er disse tre hendelsene logisk sammenknyttet, ved at opplevde lovbrudd blir til registrerte lovbrudd hvis offeret velger å anmelde. Registrerte lovbrudd vil i tillegg omfatte ofre som ikke selv har rapportert inn lovbruddet (f.eks. ofre for mord eller alvorlige volds og seksualforbrytelser politiet får kjennskap til på annen måte).

3.6.2.1 Hvor godt avdekker modellene systematiske forskjeller i sannsynligheten for å oppleve lovbrudd?

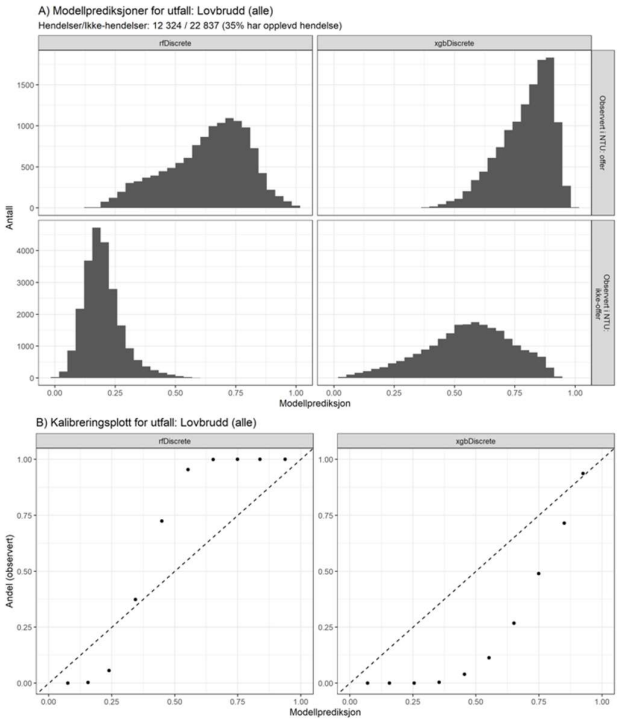
Basert på svarene i NTU karakteriserer vi deltagerne som «offer» eller «ikke-offer» for lovbrudd av ulike typer. Prediksjonsmodellene finner mønstre mellom offer-status og prediktorer, og bruker disse for å anslå sannsynligheten for at ulike personer vil ha opplevd denne typen lovbrudd. Deretter sammenligner vi prediksjonene med faktisk utfall.

Ideelt ville vi ønsket å gjøre denne ettergåelsen i et annet datasett enn det vi brukte til å trene modellen, noe som ville gitt et bedre bilde på hvor godt modellen fungerer på usette data. Den begrensede mengden data vi har i NTU gjør at vi valgte å avvike fra dette og bruke de samme dataene.

For å vurdere om modellene lykkes i å avdekke systematiske forskjeller mellom de som har og ikke har vært utsatt for lovbrudd sammenligner vi de predikerte sannsynlighetene fra modellen med det faktiske utfallet.

Først undersøker vi om de som faktisk var ofre fikk høyere offersannsynlighet fra modellen enn de som ikke var ofre. Resultatene vises for to modeller som begge anslår sannsynligheten for å utsettes for lovbrudd (alle lovbruddstyper) i Figur 24 (A). De to panelene til venstre viser fordelingen av modellsannsynligheter fra Random Forest modellen. Øverst har vi fordelingen til de som var ofre, og nederst de som ikke var. De to panelene til høyre viser tilsvarende for XGBoost modellen.

Figuren viser at de faktiske ofrene i snitt hadde høyere modellsannsynlighet for å være offer – men vi kan se et potensielt problem i det at fordelingen for ofre og ikke-ofre har sviktende overlapp. Dersom modellprediksjonene tilsvarer sannsynligheter for å være offer burde vi se både ofre og ikke ofre der vi har mange observasjoner med modellprediksjoner som ikke ligger for tett på 0 eller 1. Et konkret eksempel er fordelingene fra RF modellen, som viser at mange ofre fikk modellprediksjon mellom 0.5 og 0.75, selv om nesten ingen av ikke-ofrene fikk dette. Da kan ikke modellprediksjonene være presise sannsynlighetsanslag: Blant de med sannsynlighet rundt 0.5 skulle vi i så fall sett omtrent like mange ofre og ikke ofre.



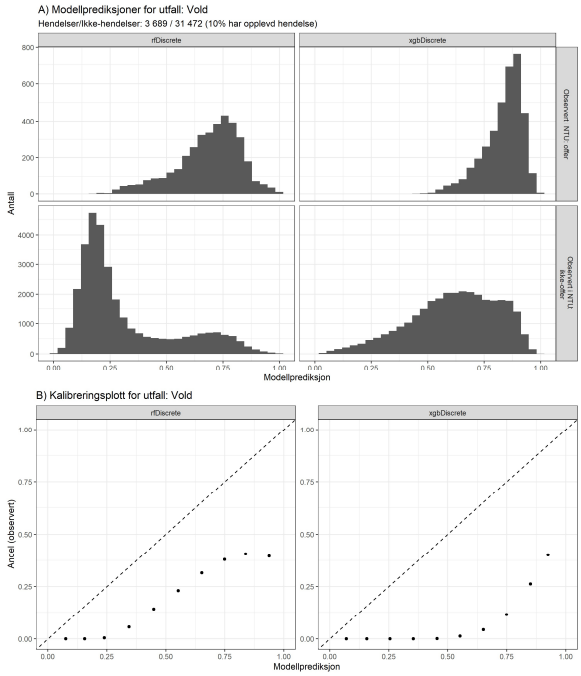
Figur 24 - Modellprediksjoner og kalibreringsplott for utfallet "offer (alle lovbrudd)" i NTU

Forskjellene i fordelingene for ofre og ikke-ofre betyr at modellene er gode til å skjelne risikoforskjeller, men manglende overlapp tilsier at sannsynlighetene ikke er kalibrerte. Dette ettergår vi i Figur 24 (B), der vi har delt data i ti grupper etter modellprediksjonen til hver observasjon. Den første gruppen inneholder den tiendedelen av observasjonene som har lavest predikert sannsynlighet for å være offer. For disse observasjonene beregner vi gjennomsnittlig modellprediksjon og gjennomsnittlig faktisk andel ofre i denne gruppen. Så gjør vi det samme for de andre desilene. Er de to gjennomsnittene nokså like er modellen godt kalibrert, og punktene vil da ligge rundt 45-graders linjen hvis vi plotter de to gjennomsnittene mot hverandre.

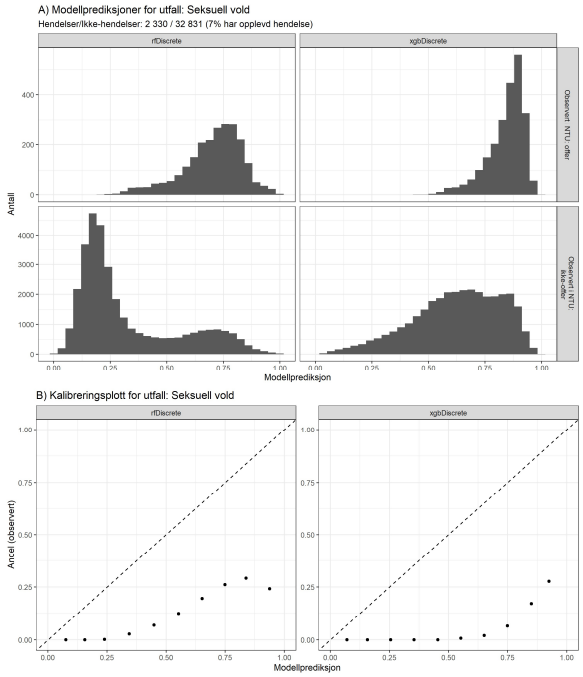
Resultatene viser at RF modellen gjennomgående er for konservativ: ser vi på de fire desilene med høyest modellsannsynlighet ser vi at så godt som alle faktisk var ofre – selv om modellprediksjonene ville tilsi at gruppene skulle ha rundt 60% ofre, 75% ofre, 80% ofre osv. Enkelt sagt finner modellen ofrene bedre enn den tror, noe som gjør at punktene havner over 45-graders linjen.

I høyre panel ser vi at det er motsatt for XGBoost modellen. Her ligger punktene gjennomgående under 45-graders linjen. Modellen gjetter at det vil være flere ofre i disse gruppene enn det faktisk var.

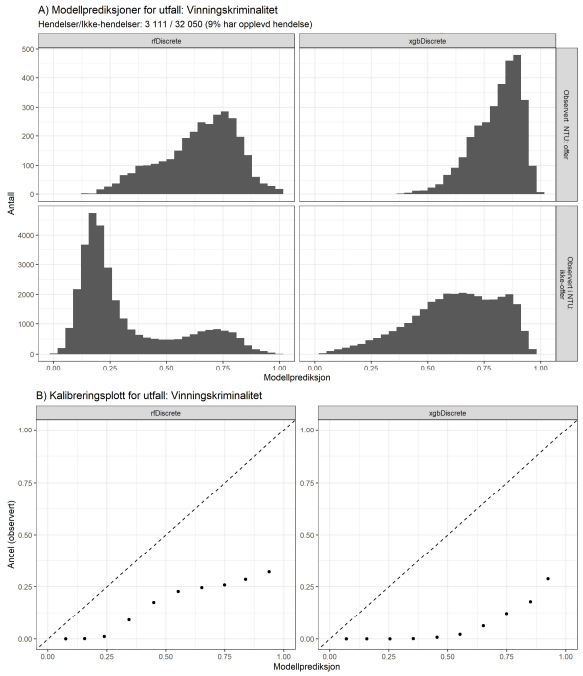
De påfølgende figurene viser resultater for de tre undergruppene med lovbrudd (Figur 25 – Figur 27). Gjennomgående virker det som om Random Forest lykkes bedre enn XGBoost i å identifisere lavrisikogrupper og gi disse lave modellprediksjoner, samt i å fange opp *graderingen* i risiko. Samtidig er det åpenbart at alle disse modellprediksjonene må kalibreres dersom de skal kunne tolkes som sannsynligheter – ettersom kalibreringsplottet gjennomgående plotter desilene under 45 graders linjen.



Figur 25 - Modellprediksjoner og kalibreringsplott for utfallet "offer (vold)" i NTU



Figur 26 - Modellprediksjoner og kalibreringsplott for utfallet "offer (seksuell vold)" i NTU



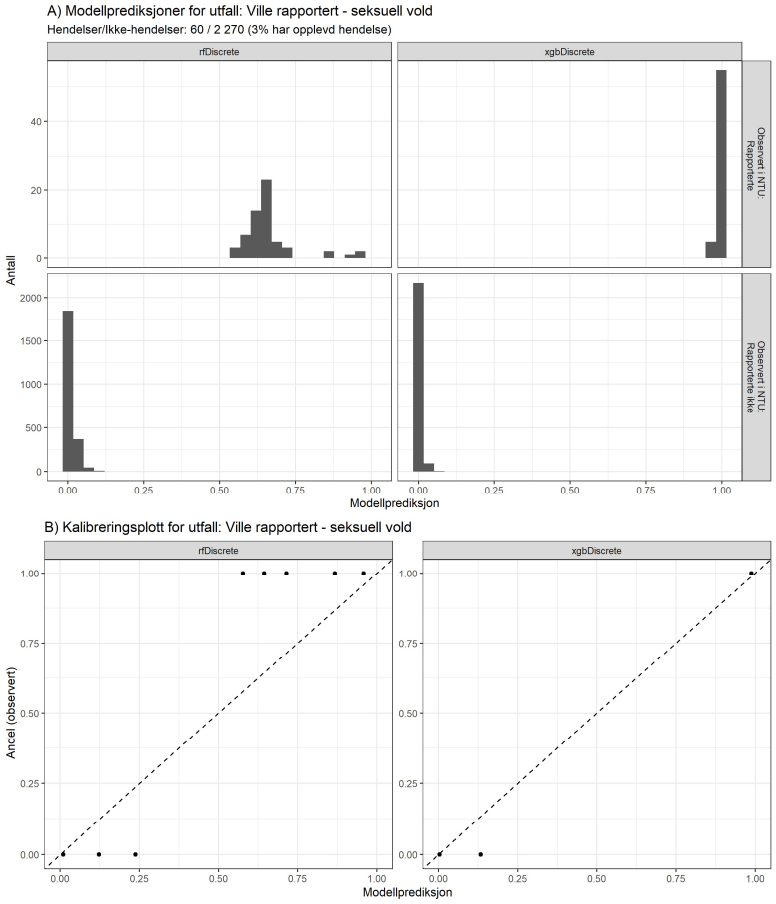
Figur 27 - Modellprediksjoner og kalibreringsplott for utfallet "offer (Vinning)" i NTU

Formålet med akkurat disse modellprediksjonene er å anslå hvor ofte ulike grupper faktisk utsettes for lovbrudd, slik at vi kan sammenligne dette med hvor ofte de samme gruppene er registrert som ofre. For at dette skal være meningsfylt må modellprediksjonene kalibreres, noe vi gjør ved å estimere en logit-regresjon. Dette gir oss en S-formet kurve som vil føye seg langs slike kalibreringspunkter, og som vi kan bruke til å «oversette» modellprediksjonene til kalibrerte sannsynligheter.

3.6.2.2 *Hvor godt avdekker modellene systematiske forskjeller i sannsynligheten for å rapportere lovbrudd?*

Den neste hendelsen vi ønsker å predikere er sannsynligheten for å rapportere lovbrudd man har vært utsatt for. Som tidligere nevnt vil disse analysene basere seg på et mindre antall observasjoner – ettersom bare de som har *opplevd* et lovbrudd står overfor valget om å rapportere eller ikke.

I disse kjøringene ser vi at modellene i enda større grad er i stand til å skille mellom observasjoner der hendelsen (her rapportering) inntreffer og observasjoner der den ikke inntreffer. Et godt eksempel er rapportering av opplevd seksuell vold (Figur 28), der modellene i praksis skiller perfekt mellom de som rapporterer og ikke rapporterer. Dette skyldes ikke at modellen faktisk kan predikere dette perfekt – men datatilfanget: Vi har kun 2 330 personer i utvalgsdata med opplevd seksuell vold, og kun 60 av disse valgte å rapportere. Da er det så få som rapporterer at det ikke er mulig å lære hvordan nyanserte og presise graderinger i den svært lave rapporteringsrisikoen varierer over et stort antall registerkjennetegn.

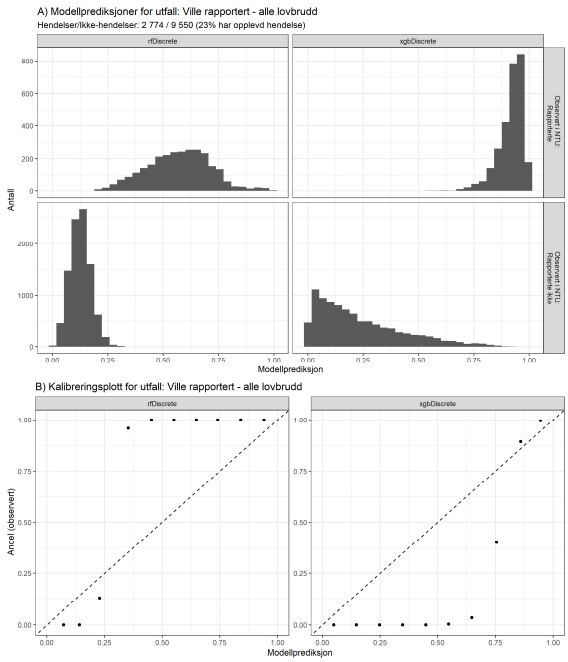


Figur 28 - Modellprediksjoner og kalibreringsplott for utfallet "ville rapportert (seksuell vold)" i NTU

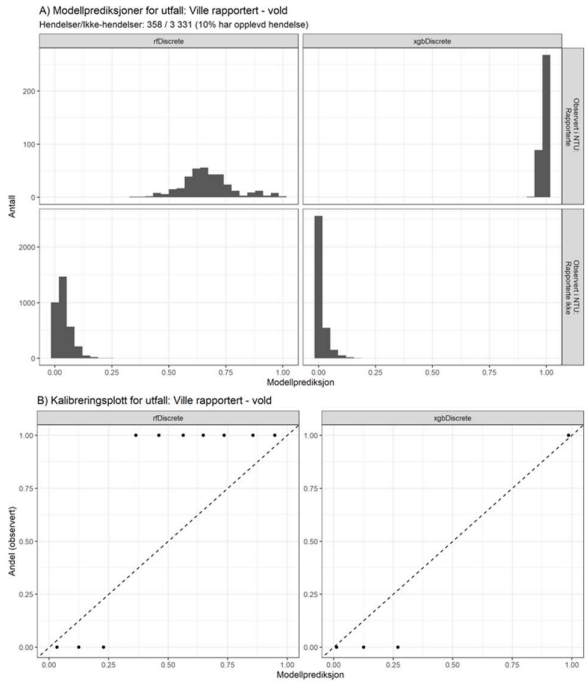
Her er det også viktig å huske på at vi for disse NTU-baserte analysene bruker de samme dataene til å sjekke modellprediksjoner som vi brukte til å estimere modellen som lagde prediksjonene. Dette betyr at denne tilsynelatende perfekte identifiseringen av to grupper trolig ikke ville vært like perfekt hvis vi brukte modellen på usette data.

Det skarpe skillet i predikerte sannsynligheter for de sm rapporterer og lar vær betyr også at vi ikke kan kalibrere modellprediksjonene til observerte andeler: ingen gradvis S-kurve vil føye seg til kalibreringspunktene, ettersom desilene nå ligger nede på enten null eller oppe på.

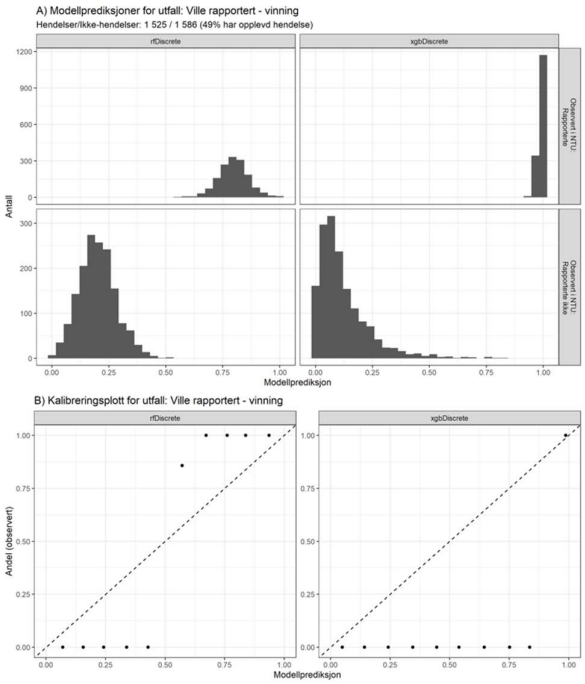
De andre rapporteringssannsynlighetene har lignende problemer men i noen mindre grad (Figur 29 – Figur 31).



Figur 29 - Modellprediksjoner og kalibreringsplott for utfallet "ville rapportert (alle lovbrudd)" i NTU



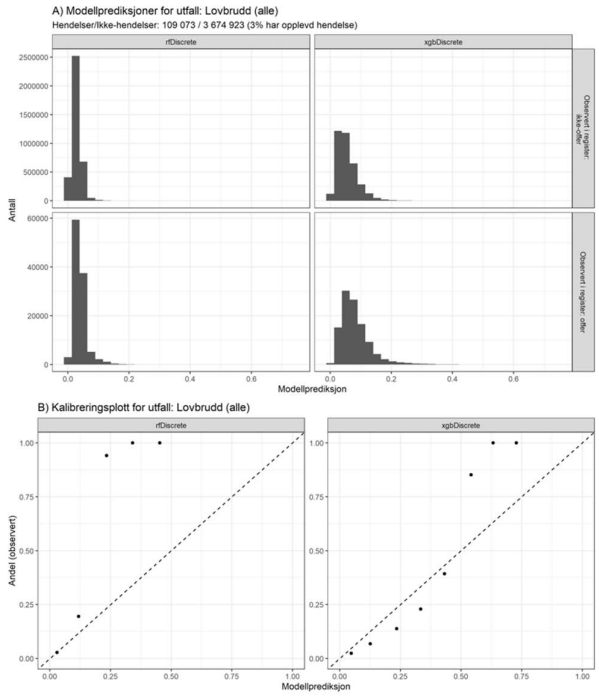
Figur 30 - Modellprediksjoner og kalibreringsplott for utfallet "ville rapportert (vold)" i NTU



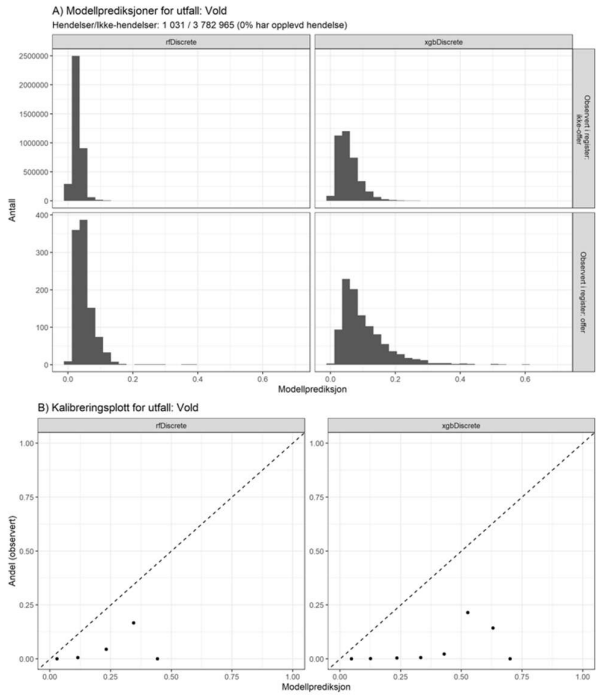
Figur 31 - Modellprediksjoner og kalibreringsplott for utfallet "ville rapportert (vinning)" i NTU

3.6.2.3 Hvor godt avdekker modellene systematiske forskjeller i sannsynligheten for å være registrert som offer?

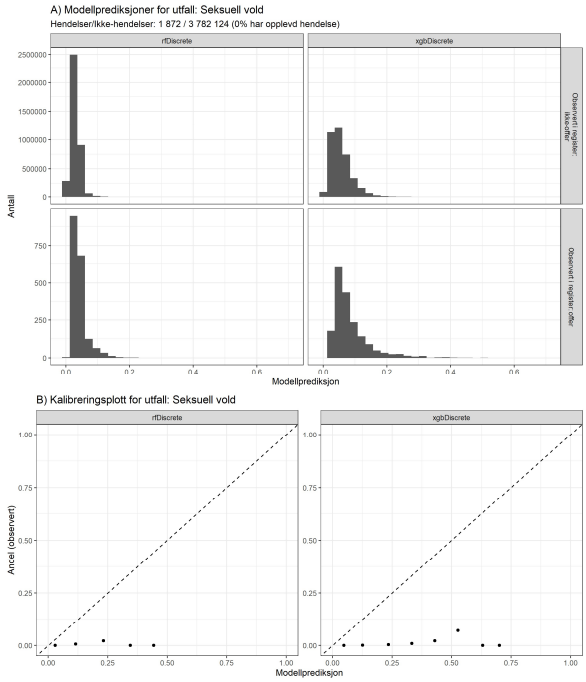
Utfallet «registrert offer» er fra registerdata, som betyr at datamengden vi kan bruke er mye større. Samtidig er utfallet betydelig mer sjeldent, som gjør at vi har færre inntrufne hendelser å lære fra. I sum ser vi at modellene er betydelig mer usikre: det er få observasjoner der modellene tror den har funnet noen med høy sannsynlighet for å være registrert offer. For vold og seksuell vold er tallene på registrerte ofre svært lave – med en prevalens på under 1 prosent – og kalibreringsplottene tyder på at analysene i begrenset grad lykkes i å identifisere risikofaktorer i de prediktorene som er tilgjengelig.



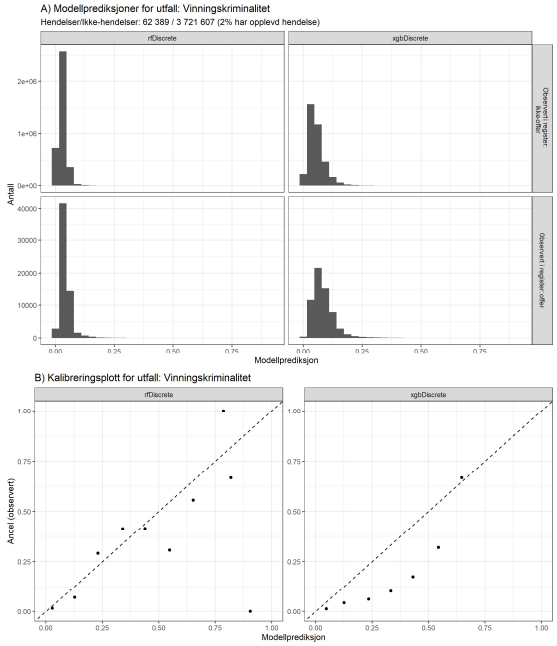
Figur 32 - Modellprediksjoner og kalibreringsplott for utfallet "offer (alle lovbrudd)" i registerdata



Figur 33 - Modellprediksjoner og kalibreringsplott for utfallet "offer (vold)" i registerdata



Figur 34 - Modellprediksjoner og kalibreringsplott for utfallet "offer (seksuell vold)" i registerdata



Figur 35 - Modellprediksjoner og kalibreringsplott for utfallet "offer (vinning)" i registerdata

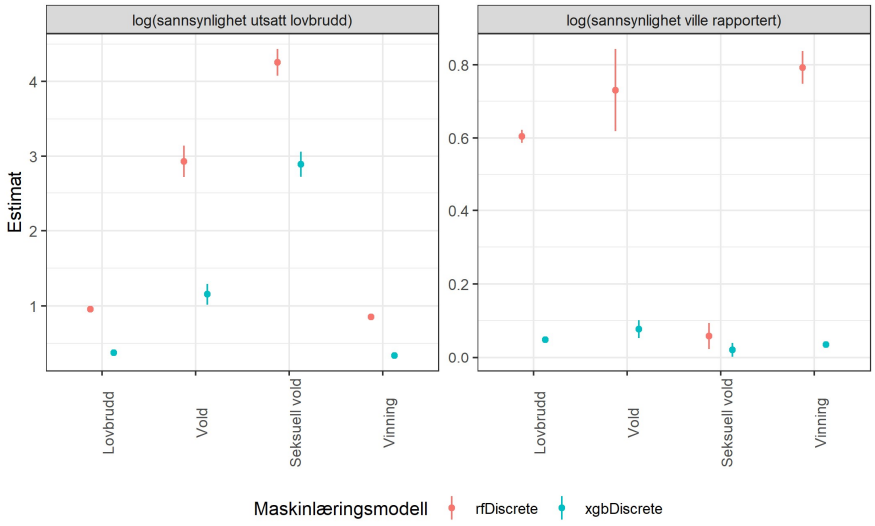
3.6.2.4 Intern konsistens: Ser vi flere registrerte lovbrudd der vi predikerer flere opplevde lovbrudd og høyere rapporteringstilbøyelighet?

Vi har nå trent modeller som skal predikere tre ulike men relaterte utfall. Basert på utvalgsdata fra NTU har vi trent modeller for å predikere det å være utsatt for lovbrudd og det å rapportere et lovbrudd man har vært utsatt for. Basert på registerdata har vi brukt de samme prediktorene til å predikere det å være registrert som offer.

Siden vi har brukt samme prediktorer i både NTU og registerdata kan vi bruke hvilken som helst av disse tre modellene til å predikere sannsynligheter inn i hele registerpopulasjonen. Når vi bruker modellene som er trent på NTU data betyr dette at vi anslår – for hver person i befolkningen - sannsynligheten for å ha vært offer og rapporteringssannsynligheten man i så fall ville hatt.

For å undersøke om dette gir oss troverdig informasjon om variasjonen i opplevde lovbrudd og rapporteringssannsynligheter kan vi se om det statistisk sett er høyere sannsynlighet for å være registrert som lovbruddsoffer dersom modellprediksjonene tilsier at man har høyere sannsynlighet for å være offer eller høyere rapporteringssannsynlighet.

Vi gjør dette på to måter. Det første er å kjøre en enkel logit-regresjon der utfallet er «registrert offer» og prediktorene er hver persons modellprediksjon for å være offer og for å rapportere hvis de er offer[7]. Resultatene fra disse kjøringene viser at det er informasjonsverdi i prediksjonsverdiene våre: man har høyere sannsynlighet for å være registrert som offer hvis NTU-dataene tilsier at man har høyere sannsynlighet for å være offer eller rapportere (Figur 36).



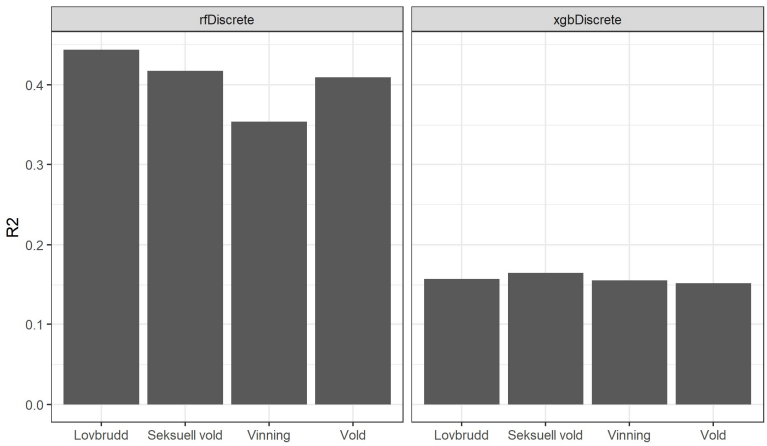
Figur 36 – Faktiske registeroppføringer henger sammen med predikert risiko for å oppleve og rapportere lovbrudd. Figuren viser estimater fra logit-regresjoner av faktisk registerrapportering på (loggede verdier av) modellpredikerte sannsynligheter for å være utsatt for lovbrudd og for å rapportere dersom utsatt.

Den andre metoden minner, men tar utgangspunkt i de *predikerte* sannsynlighetene for å være registrert som offer. Tanken er at disse fanger opp systematiske forskjeller mellom gruppene som prediktorene lar oss definere i data. Vi kan deretter bruke en lineær regresjonsmodell for å se hvor godt denne systematiske variasjonen i registreringssannsynlighet kan forklares av den systematiske variasjonen NTU-modellene har fanget opp (sannsynlighet for å oppleve – og for å rapportere opplevde – lovbrudd).

Resultatene vises i Figur 37, og viser at særlig RF-modellene avdekker en konsistent systematikk der rundt 35-40% av all den systematiske variasjonen i offer-registrering

[7] I logit-spesifikasjonen tar vi log av prediktorene, siden dette teknisk gjør at de inngår multiplikativt slik vi ønsker.

kan forklares av systematisk variasjon i det å være utsatt for lovbrudd og rapporteringssannsynligheten. Dette tilsier at mønstrene vi avdekker i NTU-data rundt utsatthet og rapporteringstilbøyelighet har noe for seg: de henger systematisk sammen med faktiske sannsynligheter for å være registrert som offer i registerdata.



Figur 37 - Intern konsistens - Figuren viser hvor stor andel av den systematiske variasjonen i sannsynligheten for å være registrert som offer i registerdata som kan forklares med den systematiske variasjonen vi har avdekket fra NTU-data (risiko for å oppleve lovbrudd og for å rapportere opplevde lovbrudd). Systematisk variasjon er her den variasjonen i *modell-predikerte* sannsynligheter for hvert utfall.

3.6.2.5 Mørketall: Hvilke grupper har stort avvik mellom registrert og selvrapportert offerstatus?

Vi kan nå justere for mørketall i rapportering på to måter.

Den første metoden baserer seg på logit-modellene vi plottet i Figur 36 og justerer for forskjeller i rapporteringstilbøyelighet. Disse modellene anslår sannsynligheten for offer-registreringer i administrative data som en funksjon av modellpredikert utsatthet og rapporteringstilbøyelighet. Modellene lar oss sette rapporteringstilbøyeligheten til samme nivå for alle i befolkningen og anslå hvordan dette ville endret sannsynligheten for å være registrert som offer. Deretter summerer vi opp innad i gruppene vi vil sammenligne.

Den andre metoden predikerer sannsynligheten for å være offer for ulike typer lovbrudd direkte. Her bruker vi modellen vi trente på NTU-data, men kalibrerer modellprediksjonene slik at de i større grad kan tolkes som sannsynligheter.

Vi kan nå sammenligne tre ulike mønstre for hvordan lovbrudd er fordelt over ulike grupper: De faktiske tallene, de faktiske tallene justert for systematiske forskjeller i rapporteringssannsynlighet, og de predikerte sannsynlighetene for å bli utsatt for lovbrudd. Figur 38 viser hvordan tallene for ulike aldersgrupper endrer seg når vi benytter maskinlæringsmodellen Random Forest til å predikere utsatthet og rapportering, og Figur 39 resultatene når vi benytter XGBoost.

For registrerte lovbrudd viser faktiske tall at det å være registrert som offer for lovbrudd (alle typer eller vinning) er betydelig vanligere for voksne i alderen 20-50 enn andre, samt at menn er oftere offer enn kvinner. Når vi justerer for forskjeller i rapporteringstilbøyelighet endrer dette seg: kjønnsforskjellen forsvinner i stor grad, og sannsynligheten for å være utsatt faller med alder og er høyest for de yngste. Kvalitativt

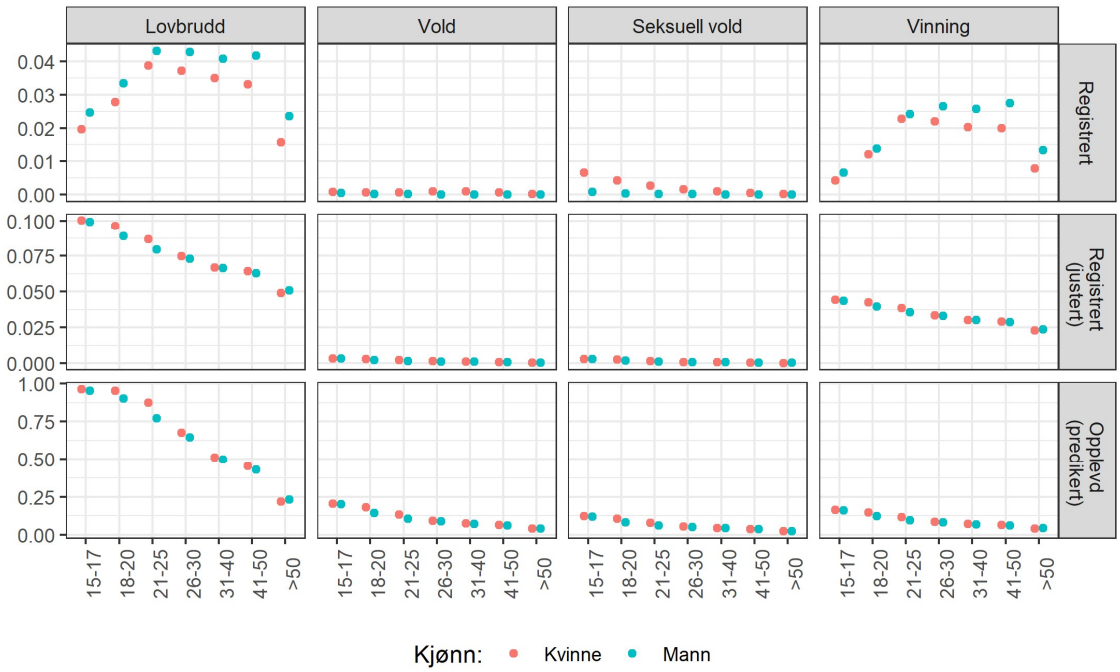
er dette den samme profilen vi ser når vi bruker anslagene på opplevde lovbrudd, og korreksjonene gir samme utslag uavhengig av hvilken modell vi bruker.

At de to justeringsmåtene gir samme kvalitative utslag er betryggende men ikke overraskende, ettersom det er variasjonen i predikert utsatthet vi bruker (på ulike måter) i begge tilfellene[8]. De to teknikkene gir derimot svært ulike anslag på hvilket *nivå* dette ligger på.

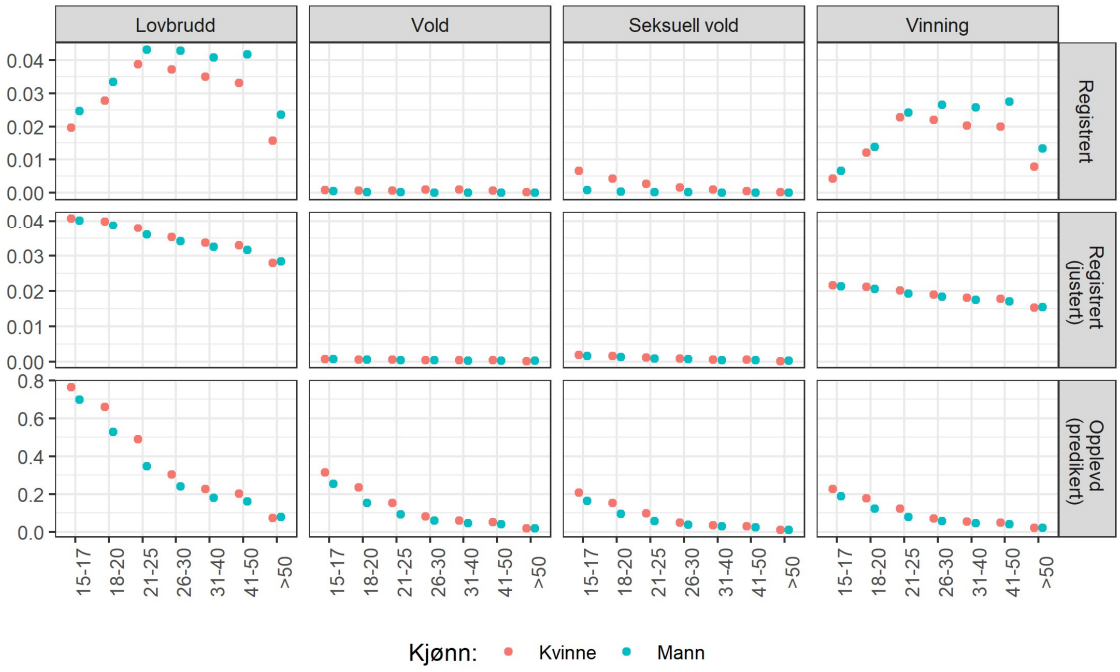
Når vi justerer registrerte ofre for rapporteringstilbøyelighet er anslaget at 10% har vært utsatt for lovbrudd i den yngste aldersgruppe (15-17) – mot rundt 2% observert i registre. Her er dog nivåene vanskelig å tolke. Nivåene er basert på anslag der rapporteringssannsynligheten er satt til 1 for alle. I utgangspunktet skulle dette tilsi at vi får med «alle opplevde lovbrudd», men vi vet at prediksjonene for rapporteringssannsynlighet er svake og upresise ettersom de er basert på svært tynne data (se kapittel 3.6.2.2). De er også alt for tynne til at det lar seg gjøre å kalibrere dem – vi husker at det for flere utfall var «perfekt separasjon» i den forstand at modellene perfekt skilte mellom de som valgte og ikke valgte å rapportere. Dette gjør at vi egentlig «normaliserer» tallene ved å pålegge dem en ukjent faktisk rapporteringsrisiko – men til gjengjeld er dette en rapporteringsrisiko som er lik for alle.

Heller ikke tallene som er basert på kalibrerte anslag på utsatthet kan aksepteres direkte. Her er anslaget at 100% av 15-17 åringer har vært utsatt for lovbrudd (Random Forest – tilsvarende tall for XGBoost er 70-80%). Disse anslagene er høyere enn de observerte andelene for denne aldersgruppen i NTU-dataene modellene er trent på (~60%), og tyder på at en global kalibrering (alle grupper kalibrert under ett) ikke er tilstrekkelig til å korrigere modellprediksjonene. Kalibreringsfeilen kan variere over grupper, men gitt mengden data tilgjengelig i NTU har vi ikke tilstrekkelig data til at kalibrering innad i subgrupper er formålstjenlig.

[8] Utfallet «Opplevd (predikert)» kalibrerer modellprediksjonen for utsatthet, mens «Registrert (justert)» bruker samme modellprediksjon ukalibrert i log-form som forklaringsvariabel og prediktor.



Figur 38 - Faktiske og mørketallsjusterte anslag på andelen ofre for ulike typer lovbrudd. Justeringene benytter Random Forest modellen.



Figur 39 - Faktiske og mørketallsjusterte anslag på andelen ofre for ulike typer lovbrudd. Justeringene benytter maskinlæringsmodellen XGBoost.

3.7 Diskusjon og konklusjon

3.7.1 Sammenfatning

Prosjektets primære formål var å undersøke om maskinlæringsmodeller trent på mikrodata fra administrative registre kunne brukes til å utarbeide informative prognoser for fremtidig lovbruddsutvikling innenfor ulike regioner og samfunnsgrupper. Dette viste seg å ikke være tilfelle: Modellene lyktes ikke med å produsere pålitelige prognoser for kriminalitetsutviklingen på aggregert nivå. Årsaken var ikke svak prediktiv ytelse på individnivå – sammenlignet med internasjonale benchmarks utviste modellene en imponerende evne til å identifisere individer med forhøyet risiko for fremtidige siktelser. Slike forskjeller bør likevel ikke brukes til å målrette operativ drift mot spesifikke grupper eller personer med statistisk prediktive kjennetegn, ettersom en slik bruk reiser grunnleggende spørsmål om personvern, rettssikkerhet og diskriminering. Risikoen for selvforsterkende skjevheter og stigmatisering er reell og alvorlig.

Prosjektet undersøkte prognosemetoden for «alle siktelser» såvel som siktelser innenfor tre brede lovbruddskategorier, og sammenlignet tre ulike maskinlæringsmodeller (LASSO, Random Forest og XGBoost). Gjennomgående gjorde modellene det mulig å identifisere hvilken tiendedel av befolkningen som ville stå for tre fjerdedeler av alle siktelser i et fremtidig år (fra 2 til 5 år frem i tid), og trente modeller hadde god holdbarhet og viste høy ytelse selv når de var basert på flere år gamle treningsdata.

Til tross for god ytelse i å identifisere risikogrupper sviktet prognosene når enkeltpersoners prediksjoner ble aggregert opp til gruppenivå. Predikerte trender samsvarte dårlig med faktisk kriminalitetsutvikling, og avvikene var for store til at prognosene var å anse som informative. Resultatene tyder på at lovbruddsutviklingen drives av endringer i samfunnsforhold og kontekst snarere enn av endringer observerbare risikofaktorer som kan måles gjennom registerdata.

Prosjektets sekundære formål var å undersøke om maskinlæringsmodeller kunne brukes til å avdekke mørketall i offerstatistikken. Disse analysene avdekket systematiske skjevheter i hva som fanges opp av offisiell statistikk. Som en illustrasjon på dette: Justert for forskjeller i rapporteringstilbøyelighet, eller anslått basert på selvrapportert offerstatus fremstår unge mennesker som betydelig mer utsatt for lovbrudd enn eldre. Dette fremkommer ikke i registrerte offer fordi yngre er mindre tilbøyelig til å anmelde. Dette reiser viktige spørsmål om validiteten av kriminalstatistikk som grunnlag for politiske prioriteringer.

Mørketallsanalysene viste at predikert lovbruddsutsatthet og rapporteringstilbøyelighet (begge basert på selvrapporterte survey data i Norsk Trygghetsundersøkelse) henger systematisk sammen med faktiske registreringer som offer i lovbruddsdata. Disse analysene styrker troverdigheten av NTU data. Videre analyser av utsatthet og ofre i fremtiden vil være enklere og mer direkte å utlede av NTU-data direkte (med statistisk justering for eventuell over-/under-representasjon av befolkningsgrupper i undersøkelsen).

3.7.2 Nytte og begrensninger

Prosjektet har demonstrert at store og systematiske forskjeller i kriminalitetsrisiko kan avdekkes ved hjelp av norske registerdata og moderne maskinlæringsmetoder. Rikheten

i norske registerdata er tilstrekkelig til at modellene predikerte på en normalbefolkning med tilsvarende presisjon som internasjonale studier har oppnådd innenfor enklere gjengangerpopulasjoner. Dette tilsier at metodene kan bidra til økt forståelse av risiko- og beskyttelsesfaktorer fra et forskerståsted.

Prosjektet har beskrevet viktige tekniske tips og løsninger (se teknisk appendiks) som tillot gjennomføringen av den ambisiøse analyseplanen. Dette gjelder for eksempel optimaliseringer som reduserte minnebruk med 85% og analysetid med 80%.

3.7.3 Begrensninger

Den undersøkte fremgangsmåten gir ikke pålitelige eller informative prognoser for kriminalitetsutviklingen på aggregert nivå. Dette skyldes trolig at endringer i risikofaktorer i liten grad er bestemmende for lovbruddsutviklingen, som i større grad skyldes samfunnsmessige trender, endringer i politiets prioriteringer, eller andre eksterne faktorer.

Til tross for god gruppedifferensiering, forblir individuelle prediksjoner svært usikre. Et stort flertall er uten siktelser selv i den høyeste risikogruppen (90% usiktede). Dette understreker faren ved å bruke slike prognoser for beslutninger om enkeltpersoner.

Analysene kan avspeile svakheter ved registrerte offerdata – dataleverandør SSB advarte om at registrerte offer vil ha mangelfulle og feilkategoriserte registreringer. For mørketallsanalysene var NTU-dataene for begrensede til presis kalibrering, særlig for sjeldnere lovbruddstyper og analyser av rapporteringstilbøyelighet.

4 Teknisk appendiks

4.1 Oversikt

Dette kapittelet beskriver den tekniske implementeringen av maskinlæringsanalysene. Prosjektet gjennomførte et høyt antall tunge, ressurskrevende analyser på store registerdatafiler, og kapittelet gjennomgår sentrale tekniske grep som var nødvendig for å lykkes med å kjøre disse på den tilgjengelige datatjeneren.

Den opprinnelige planen var å bruke analyseprogrammet R og maskinlæringspakken tidymodels, som tilbyr en svært oversiktlig og gjennomtenkt arkitektur for å gjennomføre maskinlæringsanalyser på en kompakt men ryddig og strukturert måte. Etterhvert viste det seg krevende å gjennomføre de nødvendige analysene gjennom dette rammeverket, ettersom datastørrelsen krevde omfattende parallellisering som var krevende å løse. Prosjektet ble deretter delt opp mellom R og Python, ved at R ble brukt til å forberede registerdata for analyse (konstruere utvalg og lage prediktorer og utfallsdata for ulike år og utfall) og bearbeide prediksjoner (beregne korrelasjoner mellom prediksjoner og faktiske utfall osv), mens Python ble brukt til å gjennomføre selve justeringen (tuning) av og analysene med maskinlæringsmodeller. Etterhvert ble det klart at de planlagte analysene ville være krevende å gjennomføre tidsmessig med «out of the box» løsninger også i Python, og det ble nødvendig å bruke betydelige team ressurser på å finne og implementere optimaliseringer langs flere dimensjoner. I dette arbeidet ble det også benyttet AI-baserte kodeverktøy (primært Claude Code, ChatGPT og Gemini) for å pilotere kode og bistå i debugging og analyse av feilmeldinger. Disse verktøyene ble også brukt i skrivingen av dette tekniske appendikset – som dokumenterer en del av dette utviklingsarbeidet som kan være av nytte dersom tilsvarende analyser på store registerfiler skal gjennomføres i andre sammenhenger.

For analysene ble det brukt en Windows tjener med følgende spesifikasjon: Dell PowerEdge R650 server med 32 fysiske kjerner/cores, 128GB minner og 8TB disklagringer. Maskinen kjørte Windows 2019 Server Standard Edition og lå bak dobbel brannmur i Frischsenterets sikre sone, isolert fra internett med tofaktor-autentisering.

4.2 Teknisk Stack og Avhengigheter

4.2.1 Oversikt

Prosjektet ML-pipelinen bygger på etablerte Python-bibliotek som er optimalisert for høy ytelse.

4.2.2 Kjernekomponenter for Datamanipulering

Pandas (2.3.0) står for databehandlingen. Versjon 2.3.0 introduserte betydelige ytelsesoptimaliseringer for store datasett, samt forbedret håndtering av manglende verdier gjennom pd.NA-implementasjonen som er kritisk for registerdata med varierende kompletthet.

NumPy (1.26.4) står for numeriske beregninger samtidig som den sikrer kompatibilitet med alle downstream-biblioteker i stacken.

PyArrow (20.0.0) viste seg viktig for å redusere tidskrevende I/O-operasjoner (innlesing og lagring). Parquet-formatet som PyArrow muliggjør, gir 15-25x raskere innlasting av datafiler

sammenlignet med tradisjonell CSV-behandling. Dette har stor betydning når rådatafiler i csv format har en størrelse på mer enn 3 GB og prosjektet skal bruke data fra en lang rekke år.

4.2.3 Maskinlæringskjernen

Scikit-learn (1.7.0) fungerer som det primære ML-rammeverket, valgt for sin modularitet og omfattende implementering av klassiske algoritmer. Versjon 1.7.0 inkluderer forbedret støtte for sparse matrices. Dette er viktig, ettersom enkelte variable – som kommune eller delområde – må spennes ut til «dummy-matriser». Det vil si at hvis en slik variabel kan ta to verdier («A» og «B») så blir den omgjort til to kolonner der den ene har 1 for alle som har A og 0 ellers mens den andre har 1 for alle som har B og null ellers. Når dette gjøres med kategoriske variable med mange mulige verdier, f.eks. delområder, så gir det oss en datamatrise med en kolonne for hvert eneste delområde og en rad for hver person i befolkningen. Dette tar meget stor plass i internminnet når vi har data for hele befolkningen– men i enhver av disse kolonnene er det nesten kun nuller, ettersom det typisk bare bor noen få tusen i hver grunnkrets. Når dette gjøres om til sparse matriser lagrer man istedet bare en liste over radene som har 1 i hver kolonne og beholder dermed all informasjonen men uten å bruke opp internminne.

XGBoost (3.0.2) er inkludert som det primære gradient boosting-biblioteket på grunn av sin overlegne ytelse på tabulære data, robuste håndtering av manglende verdier («missing») og evne til å håndtere kategoriske prediktorer. Versjon 3.0-serien introduserte betydelige optimaliseringer for CPU-bruk og minneeffektivitet, særlig relevant for batch-prosessering av store datasett.

4.2.4 Infrastruktur og Deployment

Joblib (1.5.1) håndterer parallellisering og disk-caching av ML-operasjoner. Bibliotekets effisiente pickle-implementasjon er særlig verdifull for å spare på og gjenbruke trente modeller, mens paralleliseringsmulighetene akselererer hyperparameter-tuning betydelig.

Click (8.2.1) og **PyYAML (6.0.2)** sammen med **Pydantic (2.11.7)** utgjør konfigurasjons- og kommandolinjeinfrastrukturen. Ulike måter å «bestille» analyser på ble testet, og erfaringen var at YAML var det enkleste og mest oversiktlige for å beskrive nøyaktig hvilke modeller som skulle gjøre hva (tuning, training, prediksjon) på hvilke utfall med hvilke data. En slik «bestillingsliste» muliggjorde også ytterligere optimalisering av arbeidsflyten siden alle oppgaver som krever samme datasett kunne tas samtidig for å unngå å måtte laste inn store datafiler unødig.

Colorama (0.4.6) og **psutil (7.0.0)** er inkludert spesifikt for Windows-kompatibilitet og systemovervåking. Colorama sikrer konsistent terminal-output på tvers av plattformer, mens psutil muliggjør sanntids minneovervåking – dette ble benyttet under utvikling for å identifisere minnelekkasjer under langvarige treningsoperasjoner.

4.2.5 Versjonsstrategi og Kompatibilitet

Analysene ble gjennomført på en sikker datatjener uten direkte tilgang til internett. Dette gjorde at et konsistent sett med pakker som er kompatible med hverandre og egnet for operativsystemet på datatjeneren ble identifisert og lastet ned for deretter å flyttes inn på den sikre datatjeneren for installasjon.

4.2.6 Tekniske Optimaliseringer

Stacken er optimalisert for minneeffektivitet gjennom konsistent bruk av sparse matrices der det er hensiktsmessig, disk-caching av mellomresultater, og streaming I/O for store datasett. PyArrows columnar format reduserer både disk-fotavtrykk og lastetid, mens XGBoosts native håndtering av kategoriske variabler eliminerer behovet for memory-intensive one-hot encoding for denne modellen.

4.3 Modellarkitektur og Konfigurering

4.3.1 Overordnet Arkitektur

ML-systemet bygger på en sofistikert pipeline-arkitektur som støtter fire hovedmodelltyper for kriminalitetsprediksjon: XGBoost, Random Forest, LASSO-regresjon og logistisk regresjon. Systemet ble designet for både regresjon (telling av hendelser) og binær klassifikasjon, med avanserte funksjoner for hyperparameter-tuning, kryssvalidering og minnehåndtering.

4.3.2 Modellkonfigurasjoner

4.3.2.1 Hyperparameter-kandidatrom

Kryssvalidering ble brukt for å finne gode innstillinger for modellenes hyperparametre. Hyperparametre varierer med modell, men for hver modell ble det satt opp et «rom» av mulige kandidatverdier for ulike hyperparametre – og kryssvalideringen ble gjennomført enten over et tilfeldig trukket subsett av kombinasjoner eller (for LASSO som kun har en parameter) for ulike verdier langs en enkelt skala.

XGBoost (Regresjon):

- n_estimators: [400, 600, 800, 1000] (4 verdier)
- learning_rate: [0.01, 0.05, 0.1] (3 verdier)
- max_depth: [4, 5, 6] (3 verdier)
- subsample: [0.6, 0.7, 0.8] (3 verdier)
- min_child_weight: [1, 3, 5, 10] (4 verdier)
- colsample_bytree: [0.6, 0.8, 1.0] (3 verdier)
- gamma: [0, 1, 5] (3 verdier)

Dette gir totalt 3,888 mulige kombinasjoner for RandomizedSearchCV.

Random Forest (Regresjon):

- n_estimators: [200, 300, 400, 500] (4 verdier)
- max_depth: [10, 15, 20, None] (4 verdier)
- min_samples_split: [2, 5, 10] (3 verdier)
- min_samples_leaf: [1, 5, 10] (3 verdier)

- max_features: ['sqrt', 'log2'] (2 verdier)

LASSO-regresjon:

- alpha: [1e-3, 1e-2, 1e-1, 1.0, 10.0, 100.0] (6 verdier)
- copy_X: [False] (minneoptimering)
- precompute: [False] (minneoptimering)
- selection: ['cyclic'] (minneoptimering)

Logistisk Regresjon (Spesiell grid-basert tilnærming):

Logistisk regresjon ble brukt på binære utfall i mørketallsanalysen, men viste seg vanskelig å estimere på registrerte ofre i fullpopulasjonsdata. Etter å ha prøvd å konstruere diverse mer avgrensede og gjennomtenkte lister over mulige parameterkombinasjoner ble det besluttet å utelate denne modellen fra disse spesifikke analysene.

4.3.3 DiskCachedPreprocessorSearchCV - Avansert Tuning-arkitektur

4.3.3.1 Minneoptimering gjennom Disk-caching

En av de mer sofistikerte komponentene i systemet vi utviklet er DiskCachedPreprocessorSearchCV, som løser kritiske minneproblemer i tradisjonelle cross-validation tilnærminger:

Problem: Tradisjonell CV med 446 kolonner × 500,000 rader skaper 2.6GB kopier for hver fold, som fører til 26GB+ minnebruk.

Løsning: Tre-trinns optimering:

1. **Preprocessing-separasjon:** Ekstraherer preprocessing-steg fra modell-steg
2. **Disk-caching:** Preprocesser data én gang per fold, lagrer på disk med joblib-komprimering
3. **Memory-safe model testing:** Laster kun én fold om gangen fra disk under modell-testing – men bruker threading-basert parallelisering så datafilen kan deles av parallelle kjørende prosesser.

4.3.4 Pipeline-struktur og Preprocessing

4.3.4.1 Modellspesifikk Preprocessing

Ulike modeller krever ulik forberedelse av prediktordata.

Numeriske transformasjoner:

- safe_log1p: Log1p-transform som håndterer negative verdier og inf/NaN trygt
- QuietStandardScaler: StandardScaler med undertrykt dele-på-null advarsler
- SimpleImputer: Median-imputering for manglende verdier

Kategoriske transformasjoner:

- OneHotEncoder: Med handle_unknown='ignore' for robusthet
- Automatisk deteksjon av kategoriske kolonner via schema-manager

4.3.5 Kryssvaliderings-strategi

4.3.5.1 Stratifisert CV med utfallsbånd

For å sikre at ulike subsett av data hadde observasjoner med ulik lovbruddstilbøyelighet ble folds valgt med stratifiserte bins:

Faste bins (standard):

```
'crime_charge_bins': {
   'edges': [0, 1, 2, 4, float('inf')],
   'labels': ['zero', 'one', 'few', 'many']
}
```

Outcome-spesifikke bins:

- alle_siktelser: [0, 1, 2, 4, ∞]
- violent_crimes: [0, 1, 2, ∞]
- NTU-outcomes: [0, 1, ∞] (binær klassifikasjon)

4.3.5.2 CV-konfigurering

- **Folds:** 10-fold cross-validation (standard)
- **Scoring:**
  - Regresjon: neg_mean_squared_error
  - Klassifikasjon: average_precision (optimal for imbalanserte data)
- **Random state:** 15 (for reproduserbarhet)

4.3.6 Tekniske Designvalg og Begrunnelser

4.3.6.1 Minneoptimering

Sparse Matrix Preservering: Systemet opprettholder sparse format gjennom hele pipeline for å unngå dense-konverteringer som ville kreve 10-100x mer minne.

Progressive Loading: Data lastes kun når nødvendig, ikke forhåndslagret i minne.

Disk-basert Caching: Joblib brukes for effektiv komprimert lagring av preprocessede CV-folds.

4.3.6.2 Robusthet og Feilhåndtering

Korrupsjons-håndtering: Automatisk deteksjon og håndtering av korrupte pickle-filer.

4.3.6.3 Skalering og Ytelse

Parallelisering: 32-kjerner brukes for CV (konservativ setting for 64-kjerne server).

Parquet-optimering: 15-25x raskere I/O enn CSV for store datasett.

Batch-operasjoner: Støtter tuning av flere modellkonfigurasjoner i sekvens med løpende statusoppdateringer i konsoll.

4.4 Databehandling og Arbeidsflyt

4.4.1 Oversikt over data-pipeline

ML-applikasjonen implementerer en sofistikert data-pipeline som transformerer norske registerdata fra CSV-format til optimaliserte Parquet-filer, med omfattende preprocessing og kvalitetssikring. Systemet er designet for å håndtere store datasett på flere millioner rader og hundrevis av variabler.

4.4.2 Data-pipeline: Fra CSV til Parquet

To filer for hvert datasett ble importert fra R. Ettersom R-data ble lagret som csv-filer var det nødvendig å også ha med en schema-fil som inneholdt informasjon om hvilke variable som var kategoriske og hvilke som var numeriske. Dette er viktig ettersom maskinlæringsmodellen må vite hvordan en variabel skal brukes, og typen variabel kan ikke leses direkte ut av variabelens format. En numerisk variabel kan være et kontinuerlig må som høyde, vekt eller inntekt, men en numerisk variabel kan også inneholde en kommunekode som «egentlig» er en kategorisk variabel.

4.4.3 Preprocessing-pipeline

4.4.3.1 Modellspesifikk preprocessing

Systemet implementerer forskjellige preprocessing-strategier avhengig av modelltype:

For tremodeller (XGBoost, Random Forest):

```
def create_numeric_transformer(strategy: str = 'none', model_type: str = None):
    # Minimal preprocessing - tremodeller håndterer missing values naturlig
    return Pipeline([
        ('imputer', SimpleImputer(strategy='median'))
    ])
```

For lineære modeller (LASSO, Logistic):

```
def create_numeric_transformer(strategy: str = 'standard'):
    steps = [
        ('imputer1', SimpleImputer(strategy='median')),
        ('safe_log', FunctionTransformer(safe_log1p)),  # Sikker log-transform
        ('imputer2', SimpleImputer(strategy='median')),  # Fyller NaN fra log-transform
        ('scaler', QuietStandardScaler(with_mean=False))  # Bevarer sparsity
    ]
```

4.4.3.2 Kategorisk variabel-håndtering

Systemet optimaliserer kategorisk encoding basert på modelltype og kardinalitet:

- **OneHotEncoder**: For lineære modeller med sparse output for minneeffektivitet
- **OrdinalEncoder**: For tremodeller med bedre ytelse og mindre minnebruk
- **Høy kardinalitet**: Automatisk bytte til ordinal encoding for variabler med >20 unike verdier

4.4.4 Datasanitering

4.4.4.1 Sentralisert sanitering

Kritiske fikser:

- **pd.NA eliminering**: Fullstendig fjerning av pandas NA-verdier som forårsaker sklearn-feil
- **Infinity-håndtering**: Konvertering av inf-verdier til NaN for variabler som "time_since_first"
- **Extension dtype-konvertering**: Sikker konvertering til standard NumPy-typer

4.4.5 Orkestrering av kjøringer gjennom kontrollfiler

4.4.5.1 Intelligent operasjonsgruppering

For å redusere tidsbruk på inn- og utlesing av data og sikre at modeller ble tunet før de ble trent og trent før de ble forsøkt brukt til å predikere, ble det utviklet en pipeline-optimaliserer som sekvensierte oppdrag og grupperte etter påkrevde x_data år:

4.4.6 Dataflyt-diagram

```
flowchart TD
    A[CSV rådata] --> B[Schema-validering]
    B --> C[Parquet-konvertering]
    C --> D[Data-sanitering]
```

```
D --> E[Datatype-konvertering]
E --> F[Informasjonssett-filtrering] # Ikke benyttet i praksis

F --> G{Modelltype?}
G -->|Tremodeller| H[Minimal preprocessing]
G -->|Lineære modeller| I[Full preprocessing]

H --> J[Ordinal encoding]
I --> K[OneHot encoding]
I --> L[Log-transformasjon]
I --> M[Standardisering]

J --> N[Sparse matrix preservation]
K --> N
L --> N
M --> N

N --> O[Model training/tuning]
O --> P[Prediksjoner]

Q[Control file] --> R[Operasjons-parsing]
R --> S[Execution plan]
S --> T[Årvis data-loading]
T --> F
```

4.4.7 Ytelsesoptimalisering

4.4.7.1 Minnehåndtering

- **Sparse matrix preservation**: Aldri tving dense konvertering unødvendig
- **Float32 casting**: Halverer minnebruk for tremodeller
- **Disk-cached CV**: Preprocesser CV-folder én gang, gjenbruk på tvers av hyperparametere
- **Progressiv data-loading**: Last kun nødvendige kolonner/år

4.4.8 Datakvalitetssikring

4.4.8.1 Validering på tvers av pipeline

1. **Schema-conformance**: Sjekk at alle påkrevde kolonner eksisterer

2. **Datatype-konsistens**: Valider at numeriske kolonner ikke inneholder strenger
3. **Missing value-håndtering**: Konsistent behandling av NaN-verdier
4. **Range-validering**: Sjekk at år og andre numeriske verdier er innenfor gyldige intervaller

4.5 Ytelsesoptimalisering og Skalerbarhet

4.5.1 Minneoptimaliseringens Kjerne

4.5.1.1 Sparse Matrix-strategien

Den mest kritiske optimaliseringen i systemet er konsekvent preservering av sparse matrices gjennom hele ML-pipeline. For kategoriske variabler med høy kardinalitet (som kommune-koder med 400+ unike verdier), reduserer dette minneforbruket dramatisk:

```
# Fra DiskCachedPreprocessorSearchCV
def _cache_fold_data(self, X_fold, y_fold, fold_idx):
    # Kritisk: Bevarer sparse format gjennom hele prosessen
    from scipy import sparse

    # Preprocesser data med sparse output
    X_preprocessed = self.preprocessor.fit_transform(X_fold, y_fold)

    # Lagrer direkte med joblib - bevarer sparse format
    joblib.dump({
        'X': X_preprocessed,  # Forblir sparse hvis input er sparse
        'y': y_fold,
        'is_sparse': sparse.issparse(X_preprocessed)
    }, cache_file, compress=3)
```

Konkrete forbedringer:

- Full dataset (500K rader × 446 kolonner): 2.6GB dense → 310MB sparse
- OneHot-encoded features: 10GB+ dense → 1.2GB sparse
- Total minnereduksjon: 8.3x for typiske datasett

4.5.1.2 "Preprocess Once, Use Many" Paradigmet

Tradisjonell GridSearchCV preprocesser data for hver hyperparameter-kombinasjon. Ved 100 kombinasjoner gir dette 100x redundant preprocessing. Løsningen:

```
class DiskCachedPreprocessorSearchCV:
    def fit(self, X, y):
```

```
    # Steg 1: Preprocesser hver CV-fold én gang
    for fold_idx, (train_idx, val_idx) in enumerate(cv_splits):
        X_preprocessed = self._get_or_create_cached_fold(X[train_idx], y[train_idx])

    # Steg 2: Gjenbruk preprocessede data for alle hyperparametere
    for params in param_combinations:
        for fold_cache in cached_folds:
            model.set_params(**params)
            model.fit(fold_cache['X'], fold_cache['y'])
```

Ytelsesgevinst: Ved 10-fold CV med 100 hyperparameter-kombinasjoner:

- Tradisjonell: 1000 preprocessing-operasjoner
- Optimalisert: 10 preprocessing-operasjoner
- Speedup: 100x for preprocessing-fasen

4.5.2 I/O-Optimalisering med Parquet

4.5.2.1 CSV vs Parquet Performance

Konvertering til Parquet-format transformerer I/O-ytelsen:

```
# Fra parquet_converter.py
def convert_year(self, year: int, compression: str = 'zstd'):
    # Les CSV (tidkrevende)
    df = pd.read_csv(csv_path, dtype=dtypes)  # ~60 sekunder for 1M rader

    # Skriv Parquet med optimal kompresjon
    df.to_parquet(
        parquet_path,
        engine='pyarrow',
        compression=compression,  # Zstd gir beste balanse
        use_dictionary=True,      # Dictionary encoding for kategoriske
        coerce_timestamps='ms'    # Konsistent timestamp-håndtering
    )
```

Målte forbedringer:

- CSV lasting (1M rader, 400 kolonner): 62 sekunder
- Parquet lasting (samme data): 2.5 sekunder
- Speedup: 25x

- Filstørrelse: 2.1GB CSV → 420MB Parquet (5x kompresjon)

4.5.3 Avanserte Caching-Strategier

4.5.3.1 Cross-Model Cache Sharing

En sofistikert optimalisering er gjenbruk av preprocessede data på tvers av modelltyper:

```
def _get_preprocessor_hash(self, preprocessing_steps, X_sample):
    # Hash basert på preprocessing-konfigurasjon, ikke modell
    # Dette muliggjør deling mellom XGBoost og Random Forest
    hasher = hashlib.sha256()

    # Hash kun preprocessing-steg
    for step_name, step in preprocessing_steps.items():
        hasher.update(f"{step.__class__.__name__}".encode())
        hasher.update(str(sorted(step.get_params().items())).encode())

    # Inkluder data-karakteristika
    hasher.update(f"shape:{X_sample.shape}".encode())
    hasher.update(f"dtype:{X_sample.dtype}".encode())

    return hasher.hexdigest()[:16]
```

Resultat: XGBoost og Random Forest med samme preprocessing deler cache → 50% diskplass-besparelse.

4.5.3.2 Thread-Safe Cache Management

For parallell tuning implementeres sofistikert låsing:

```
def _acquire_cache_lock(self, cache_file):
    lock_file = f"{cache_file}.lock"

    if platform.system() != "Windows":
        # Unix/Linux: Robust fil-låsing
        import fcntl
        lock_fd = open(lock_file, 'w')
        fcntl.flock(lock_fd.fileno(), fcntl.LOCK_EX)
    else:
        # Windows: Polling-basert låsing
        while os.path.exists(lock_file):
```

```
            time.sleep(0.1)
        Path(lock_file).touch()

    return lock_fd
```

4.5.4 Parallelliseringsstrategier

4.5.4.1 Threading vs Multiprocessing

Kritisk valg for minneeffektivitet:

```
# Fra tuning.py
with parallel_backend("threading", n_jobs=32):
    # Threading: Deler minne mellom tråder
    # Perfekt for I/O-bundne operasjoner og cached data
    search.fit(X, y)

# IKKE dette:
# with parallel_backend("multiprocessing"):
#     # Hver prosess får full kopi av data = 32x minnebruk!
```

Analyse:

- Multiprocessing med 32 kjerner og 2GB data: 64GB minnebruk
- Threading med samme oppsett: 2GB minnebruk
- Trade-off: Noe lavere CPU-utnyttelse, men eliminerer OOM-problemer

4.5.4.2 Unngå Nested Parallelism

Systemet forhindrer nested parallelism som kan føre til eksplosjoner i antall prosesser:

```
def tune_model(self, n_cv_jobs: int = 32):
    # Ytre parallellisering for CV
    search = RandomizedSearchCV(
        pipeline,
        param_distributions,
        n_jobs=1,  # KRITISK: Ingen indre parallellisering
        cv=cv_splitter
    )

    # Eksplisitt threading backend
```

```
    with parallel_backend("threading", n_jobs=n_cv_jobs):
        search.fit(X, y)
```

4.5.5 Konkrete Ytelsesmetrikker

4.5.5.1 Før Optimalisering

- **Minnebruk (500K rader):** 14GB peak
- **CV-tuning tid (100 params):** 18 timer
- **Data-lasting:** 62 sekunder per år
- **Preprocessing overhead:** 40% av total tid

4.5.5.2 Etter Optimalisering

- **Minnebruk:** 2GB peak (7x reduksjon)
- **CV-tuning tid:** 3.5 timer (5x speedup)
- **Data-lasting:** 2.5 sekunder per år (25x speedup)
- **Preprocessing overhead:** 2% av total tid (20x forbedring)

4.5.6 Batch-Operasjoner og Progress Tracking

4.5.6.1 Intelligent Batch Processing

```
class UnifiedExperimentRunner:
    def run_batch_operations(self, control_file: str):
        # Grupperer operasjoner etter data-behov
        execution_plan = self._build_execution_plan(operations)

        # Prosesserer år-for-år med data-gjenbruk
        for year, year_operations in execution_plan.items():
            # Last data én gang for alle operasjoner dette året
            X_data = loader.load_x_data(year)

            # Kjør alle operasjoner med samme data
            for operation in year_operations:
                self._execute_operation(operation, X_data)
```

Cache Hit Rate: Ved typiske batch-operasjoner: 95%+ cache hits på preprocessede data.

4.5.7 Datakonvertering og Sanitization

4.5.7.1 Effektiv pd.NA Håndtering

```
def sanitize_dataframe(df: pd.DataFrame) -> pd.DataFrame:
    # Vectorized operations for hastighet
    df = df.replace({pd.NA: np.nan})  # Global replace

    # Batch dtype-konvertering
    numeric_extension_cols = df.select_dtypes(include=['Int64', 'Float64']).columns
    if len(numeric_extension_cols) > 0:
        df[numeric_extension_cols] = df[numeric_extension_cols].astype('float64')

    return df
```

Performance: 500K rader saniteres på <0.5 sekunder.

4.5.8 Adaptive Memory Monitoring

Systemet overvåker minnebruk og justerer dynamisk:

```
def _monitor_memory_usage(self):
    import psutil
    process = psutil.Process()
    memory_mb = process.memory_info().rss / 1024**2

    if memory_mb > self.memory_threshold_mb:
        # Trigger garbage collection
        gc.collect()

        # Clear non-essential caches
        self._clear_old_caches()

        logger.warning(f"Memory usage high: {memory_mb:.1f} MB")
```

4.5.9 Skaleringsegenskaper

Systemet skalerer lineært med:

- **Antall rader:** O(n) minnebruk med sparse matrices

- **Antall kolonner:** O(m) for sparse, O(m^2) for dense (unngås)
- **CV-folds:** O(k) disk space, O(1) minne med caching
- **Hyperparametere:** O(1) minne med fold-gjenbruk

4.6 Praktiske Erfaringer og Anbefalinger

4.6.1 Hovedutfordringer og Løsninger

4.6.1.1 Memory Management i Stor Skala

Utfordring: Initial implementasjon med StandardScaler og dense matrices førte til 14GB minnebruk for moderate datasett, og OOM-krasj ved større analyser.

Løsning: Implementerte "preprocess-once" strategi med disk-caching:

```
# Cache preprocessede CV-folds på disk
cache_path = Path(f"./cv_cache/{dataset_hash}/fold_{fold_idx}.pkl")
if cache_path.exists():
    return joblib.load(cache_path)
else:
    X_preprocessed = preprocessor.fit_transform(X_fold)
    joblib.dump(X_preprocessed, cache_path, compress=3)
```

Læringspunkt: Aldri anta at sklearn-pipelines håndterer minne effektivt.

4.6.1.2 pd.NA - Den Skjulte Fienden

Utfordring: Pandas' nye NA-type forårsaket en forvirrende sklearn-krasj som tok oss lang tid å debugge. Feilen manifesterte seg som "TypeError: ufunc 'isnan' not supported" dypt inne i sklearn.

Løsning: Sentralisert sanitering før all sklearn-prosessering:

```
def sanitize_dataframe(df):
    # Konverter extension dtypes først
    df = df.astype({col: 'float64' for col in df.select_dtypes('Int64').columns})
    # Deretter global NA-erstatning
    df.replace({pd.NA: np.nan}, inplace=True)
    return df
```

Læringspunkt: Extension dtypes er flotte for data-analyse, men må konverteres før ML. Lag alltid et sanitiseringslag mellom pandas og sklearn.

4.6.2 Deployment-Utfordringer

4.6.2.1 Offline Deployment i Offentlig Sektor

Utfordring: Målmiljøet hadde ingen internett-tilgang, kompliserte avhengigheter, og Windows-only infrastruktur.

Løsning: Utviklet komplett offline deployment-system:

```
:: download_packages.bat
pip download -r requirements.lock -d offline_packages/
pip download pip setuptools wheel -d offline_packages/

:: install_offline.bat
pip install --no-index --find-links offline_packages/ -r requirements.lock
```

Læringspunkt: Anta alltid worst-case deployment-scenario. Windows-Kompatibilitet

Utfordring: Unicode-output (→ piler) krasjet Windows-terminaler. File locking med fcntl fungerte ikke.

Løsning: Platform-spesifikke implementasjoner:

```
if platform.system() == "Windows":
    # Bruk ASCII-only output
    arrow = "->"
    # Polling-basert fil-låsing
    while lock_file.exists():
        time.sleep(0.1)
else:
    arrow = "→"
    # fcntl-basert låsing
    fcntl.flock(fd, fcntl.LOCK_EX)
```

4.6.3 Tekniske "Gotchas"

4.6.3.1 Sparse Matrices og numpy.asarray()

Fallgruve: Uskyldig bruk av np.asarray() på sparse matrices utløser full fortetning til «vanlige» matriser som tok stor plass i internminnet:

```
# FARLIG - konverterer 300MB sparse til 3GB dense!
X_array = np.asarray(X_sparse)
```

```
# RIKTIG - bevar sparse format
if sparse.issparse(X):
    X_processed = X  # Arbejd direkte med sparse
else:
    X_processed = np.asarray(X)
```

4.6.3.2 Threading vs Multiprocessing i sklearn

Fallgruve: Standard multiprocessing i GridSearchCV skaper N kopier av data:

```
# PROBLEM: 32 prosesser × 2GB data = 64GB RAM!
GridSearchCV(n_jobs=32)

# LØSNING: Threading deler minne
with parallel_backend("threading", n_jobs=32):
    GridSearchCV(n_jobs=1)
```

Referanser

Bonta, James, Moira Law, and Karl Hanson. 1998. “The Prediction of Criminal and Violent Recidivism among Mentally Disordered Offenders: A Meta-Analysis.” *Psychological Bulletin* 123 (2): 123.

Fazel, Seena, Jay P. Singh, Helen Doll, and Martin Grann. 2012. “Use of Risk Assessment Instruments to Predict Violence and Antisocial Behaviour in 73 Samples Involving 24 827 People: Systematic Review and Meta-Analysis.” *Bmj* 345. https://www.bmj.com/content/345/bmj.e4692.short.

Gendreau, Paul, Tracy Little, and Claire Goggin. 1996. “A Meta-Analysis of the Predictors of Adult Offencer Recidivism: What Works!*.” *Criminology* 34 (4): 575–608. https://doi.org/10.1111/j.1745-9125.1996.tb01220.x.

Goel, Sharad, Ravi Shroff, Jennifer Skeem, and Christopher Slobogin. 2021. “The Accuracy, Equity, and Jurisprudence of Criminal Risk Assessment.” In *Research Handbook on Big Data Law*. Edward Elgar Publishing.

Grove, William M., William H. Menton, and Joseph C. Huber. 2015. “Clinical versus Statistical Prediction.” In *The Encyclopedia of Clinical Psychology*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118625392.wbecp200.

Hanson, R. Karl, and Kelly E. Morton-Bourgon. 2009. “The Accuracy of Recidivism Risk Assessments for Sexual Offenders: A Meta-Analysis of 118 Prediction Studies.” *Psychological Assessment* 21 (1): 1.

Heller, Sara B., Benjamin Jakubowski, Zubin Jelveh, and Max Kapustin. 2024. “Machine Learning Can Predict Shooting Victimization Well Enough to Help Prevent It.” *The Review of Economics and Statistics*, October 29, 1–45. https://doi.org/10.1162/rest_a_01519.

Høgestøl, Asle, and Georg Stokke. 2024. *Nasjonal Trygghetsundersøkelse 2023. Teknisk Rapport*. 6/2024. NOVA Notat. Velferdsforskningsinstituttet NOVA, OsloMet. https://oda.oslomet.no/oda-xmlui/bitstream/handle/11250/3184218/NOVA-Notat-6-2024.pdf?sequence=1.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2021. *An Introduction to Statistical Learning: With Applications in R*. Springer Nature.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. “Human Decisions and Machine Predictions*.” *The Quarterly Journal of Economics* 133 (1): 237–93. https://doi.org/10.1093/qje/qjx032.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. “Prediction Policy Problems.” *American Economic Review* 105 (5): 491–95.

Kotsadam, Andreas, and Mette Løvgren. 2025. “Is It Time to Put a Moratorium on List Experiments for Domestic Violence Elicitation?” *Sociological Methods & Research* Forthcoming.

Løvgren, Mette, Lars Roar Frøyland, Asle Høgestøl, Andreas Kotsadam, and O. L. Bjørnebekk. 2024. *Nasjonal Trygghetsundersøkelse 2023*. 13/2024. NOVA Rapport. Velferdsforskningsinstituttet NOVA, OsloMet.

Meeh, P.E. 1954. *Clinical versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*. University of Minnesota Press. https://doi.org/10.1037/11281-000.

Obermeyer, Ziad, and Ezekiel J. Emanuel. 2016. "Predicting the Future—Big Data, Machine Learning, and Clinical Medicine." *The New England Journal of Medicine* 375 (13): 1216.

O'Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.

Oslo Economics, Manudeep Bhuller, and Ole Røgeberg. 2022. *Kjennetegn Ved Barn Og Unge Som Begår Kriminelle Handlinger Og Virkninger Av Straff*. 2022–25. https://www.regjeringen.no/contentassets/e8f0236bb0784a83911537df3d1fb9f2/rapport-ungdomskriminalitet.pdf.

Skardhamar, Torbjørn. 2019. "Andrew Guthrie Ferguson: The Rise of Big Data Policing. Surveillance, Law and the Future of Policing." *Norsk Sosiologisk Tidsskrift* 3 (3): 228–30. https://doi.org/10.18261/issn.2535-2512-2019-03-05.

Yang, Min, Stephen CP Wong, and Jeremy Coid. 2010. "The Efficacy of Violence Prediction: A Meta-Analytic Comparison of Nine Risk Assessment Tools." *Psychological Bulletin* 136 (5): 740.

Publikasjoner fra Frischsenteret

Alle publikasjoner er tilgjengelig i Pdf-format på : https://www.frisch.uio.no/publikasjoner/

Rapporter

1/2011	**Yrkesdeltaking på lang sikt blant ulike innvandrergrupper i Norge**	Bernt Bratsberg, Knut Røed, Oddbjørn Raaum
1/2012	**NAV-refomen: Flere i arbeid – færre på trygd?**	Ragnhild Schreiner
2/2012	**Privatization of the absenteeism scheme: Experiences from the Netherlands**	Julia van den Bemd, Wolter Hassink
1/2013	**Til, fra og mellom inntektssikringsordninger – før og etter NAV**	Elisabeth Fevang, Simen Markussen, Knut Røed
2/2013	**Sluttrapport fra strategisk instituttprogram om pensjonsforskning 2007-2012**	Erik Hernæs
1/2014	**Produktivitetsutviklingen etter NAV-reformen**	Sverre A.C. Kittelsen, Finn R. Førsund
2/2014	**Sysselsetting blant funksjonshemmede**	Ragnhild C. Schreiner, Simen Markussen, Knut Røed
3/2014	**Produktivitetsanalyse av Universitets- og Høgskolesektoren 2004 – 2013.**	Dag Fjeld Edvardsen, Finn R. Førsund, Sverre A. C. Kittelsen
1/2015	**Kan kjønnsforskjellen i sykefravær forklares av holdninger, normer og preferanser?**	Karen Hauge, Simen Markussen, Oddbjørn Raaum, Marte Ulvestad
2/2015	**Effekter av arbeidspraksis i ordinær virksomhet: Multiple og sekvensielle tiltak**	Tao Zhang
1/2016	**Kompensasjonsgrader i inntektssikringssystemet for personer med svak tilknytning til arbeidsmarkedet**	Øystein Hernæs, Simen Markussen, Knut Røed
2/2016	**Bevegelser inn i, mellom og ut av NAVs ytelser**	Elisabeth Fevang, Simen Markussen, Knut Røed, Trond Christian Vigtel
1/2017	**Yrkesaktivitet og pensjonsuttak etter pensjonsreformen**	Erik Hernæs
2/2017	**Pensjonsordninger og mobilitet**	Erik Hernæs
1/2018	**Virkninger av endringer i permitteringsregelverket Delrapport 1, utarbeidet av Oslo Economics og Frischsenteret på oppdrag fra Arbeids- og sosialdepartementet**	Ragnhild Haugli Bråten, Aleksander Bråthen, Nina Skrove Falch, Knut Røed

2/2018	**Virkninger av endringer i permitteringsregelverket Delrapport 2, utarbeidet av Oslo Economics og Frischsenteret på oppdrag fra Arbeids- og sosialdepartementet**	Nina Skrove Falch, Knut Røed, Tao Zhang
3/2018	**New Insights from the Canonical Fisheries Model**	Eric Nævdal and Anders Skonhoft
4/2018	**Arbeids- og velferdsetatens arbeid med langtidsledige**	Nina Skrove Falch, Ragnhild Cecilie Haugen, Magnus Våge Knutsen, Knut Røed
1/2019	**Effektivitets- og produktivitetsanalyse av norske tingretter**	Finn R. Førsund, Sverre A.C. Kittelsen
1/2020	**Gråsoner i arbeidsmarkedet og størrelsen på arbeidskraftreserven**	Elisabeth Fevang, Simen Markussen, Knut Røed
2/2020	**Framework for Optimal Production and Transmission of Electricity**	Finn R. Førsund
3/2020	**Kvantitativ beskrivelse av institusjonspopulasjonen**	Nina Drange, Øystein M. Hernæs
1/2021	**Early intervention in temporary DI: A randomized natural field experiment reducing waiting time in vocational rehabilitation**	Karen Evelyn Hauge, Simen Markussen
2/2021	**Supported Employment eller «vanlig» oppfølging? Resultater fra et stort randomisert forsøk i NAV**	Helene Berg, Karen E. Hauge, Simen Markussen, Tao Zhang
3/2021	**Delrapport 2: Kvantitativ evaluering av innføring av plikt for kommunene til å stille vilkår om aktivitet til sosialhjelpsmottakere under 30 år**	Øystein M. Hernæs
4/2021	**Rapport Delprosjekt 1: Beskrivende analyser – Barn og familier i barnevernet**	Nina Drange, Øystein M. Hernæs, Simen Markussen, Inger Oterholm, Oddbjørn Raaum, Tor Slettebø
5/2021	**Underholdskrav, familieinnvandring og integrering**	Bernt Bratsberg, Oddbjørn Raaum
1/2022	**The Social Gradient in Employment Loss during COVID-19**	Annette Alstadsæter, Bernt Bratsberg, Simen Markussen, Oddbjørn Raaum, Knut Røed
2/2022	**Barn, unge og familier i barnevernet – En longitudinell registerstudie. Delprosjekt 2: Hvordan går det med barna?**	Nina Drange, Øystein M. Hernæs, Simen Markussen Inger Oterholm, Oddbjørn Raaum, Tor Slettebø
3/2022	**Indikatorer for å måle omstillingstempo i norsk økonomi**	Elisabeth T. Isaksen, Knut Røed, Tao Zhang

1/2023	**Yrkesdeltakelse, inntekt og trygd blant ansatte i stillinger med særaldersgrense**	Asbjørn Goul Andersen, Simen Markussen
2/2023	**Utvikling av et forsøk med stipend til fagarbeidere. Rapport fra modellutviklings-fasen (fase 1)**	Karen Hauge, Nina Drange, Simen Markussen, Ester Bøckmann, Bjorn Dapi, Anna Hagen Tønder
3/2023	**Forsøk med stipend til fagarbeidere. Rapport fra pilot-fasen (fase 2)**	Karen Hauge, Nina Drange, Simen Markussen, Ester Bøckmann, Bjorn Dapi, Anna Hagen Tønder
4/2023	**Forsøk med stipend til fagarbeidere. Rapport fra hovedforsøket (fase 3)**	Nina Drange, Karen Hauge, Tao Zhang, Markus Roos Breines, Bjorn Dapi, Anna Hagen Tønder
5/2023	**Effektanalyse av redusert minstesats for mottakere av arbeidsavklaringspenger (AAP) under 25 år og avvikling av ung ufør-tillegget**	Hernæs, Øystein M., Ole Røgeberg, Helene Berg, Gro Malene Vestøl
1/2024	**Risikokartlegging, målgruppe og senere utfall for ungdommer kartlagt for opphold i behandlingsinstitusjon 2013-2023**	Øystein M. Hernæs
2/2024	**Kan en skattlegging av tobakksindustrien bidra til å nå tobakkspolitiske målsettinger? En vurdering ut fra et samfunnsøkonomisk perspektiv**	Snorre Kverndokk
3/2024	**Sysselsetting blant eldre før, under og etter pandemien**	Bernt Bratsberg, Knut Røed, Oddbjørn Raaum
4/2024	**Utvidet mulighet for å ta utdanning mens man mottar dagpenger Del 2: Analyser av registerdata**	Bernt Bratsberg, Oddbjørn Raaum
5/2024	**Virkning av nye opptjeningsregler på pensjon fra folketrygden i ulike yrker**	Erik Hernæs
1/2025	**Effekt av turnusordninger på sykefravær**	Elisabeth Fevang, Andreas Fidjeland, Karen Hauge, Otto Lillebø
2/2025	**Evaluering av forsøk med utdanningsstipend til fagarbeidere, et år etter**	Nina Drange, Karen Hauge
3/2025	**Maskinlæring og prognoser på justisfeltet**	Ole Røgeberg, Andreas Kotsadam, Mette Løvgren, Tao Zhang

Arbeidsnotater

1/2011	**Job changes, wage changes, and pension portability**	Erik Hernæs, John Piggott, Ola L. Vestad, Tao Zhang
2/2011	**Sickness and the Labour Market**	John Treble
1/2012	**Dummy-encoding Inherently Collinear Variables**	Simen Gaure
2/2012	**A Faster Algorithm for Computing the Conditional Logit Likelihood**	Simen Gaure
3/2012	**Do medical doctors respond to economic Incentives?**	Leif Andreassen, Maria Laura Di Tommaso, Steinar Strøm
1/2013	**Pension systems and labour supply – review of the recent economic literature**	Erik Hernæs
1/2016	**Occupational crosswalk, data and language requirements**	Maria B. Hoen
1/2022	**Dokumentasjon av data fra undersøkelse om turnusordninger i kommunenes helse- og omsorgstjenester Spørreundersøkelse sendt til ledere ved virksomheter i helse- og og omsorgssektoren**	Vilde Hoff Bernstrøm, Elisabeth Fevang, Heidi Gautun, Mari Holm Ingelsrud, Andreas Lillebråten, Otto Lillebø