Rapport

Bokontroll basert på maskinlæring

- En vurdering ved Riksrevisjonen i samarbeid med Statens lånekasse for utdanning

Innhold

Konklusjoner ... 3
Positive funn ... 3
Mulige områder for forbedring ... 3
1. Innledning ... 4
1.1 Bakgrunnen for Riksrevisjonens gjennomgang ... 4
1.2 Avgrensning, mål og problemstillinger ... 5
2. Beskrivelse av området ... 7
2.1 Grunnlag og historikk for Lånekassens bokontroll ... 7
2.2 Generelle krav til saksbehandling ... 8
3. Gjennomgang av maskinlæringsmodellen ... 9
3.1 Datagrunnlag for modellen ... 9
3.2 Om ML-modellen ... 10
3.3 Reproduserbarhet ... 11
3.4 Valg av variabler ... 12
3.5 Overtrening ... 13
3.6 Kvalitetsikring ... 13
4 Hvor god er Lånekassens modell? ... 14
4.1 Målparameter og metodikk ... 14
4.2 Resultat ... 15
Hvor mange skal man kontrollere? ... 16
5. Vurderinger av etiske spørsmål ... 17
5.1 Om forklarbarhet ... 17
5.2 Om likebehandling ... 19
Appendiks ... 21
A1 Begrepsforklaringer ... 21
A2 Equality and Fairness measures in classification models ... 21
A3 Teknisk vedlegg med dokumentasjon av funn ... 23
A3.1 Overtrening ... 23
A3.2 Likebehandling og fairness ... 25
A4 Uttrekk fra Lånekassens dokumentasjon ... 28
A5 Referanser: Dokumentasjon fra Lånekassen ... 30

Konklusjoner

Positive funn

- Lånekassen er i front mht. å ta i bruk maskinlæring i forvaltningen og dette arbeidet har gitt gevinster.
- Datagrunnlaget og utviklingsprosessen for maskinlæringsmodellen virker stort sett solid.
- Det er positivt at Lånekassen har inkludert en tilfeldig utvalgt, representativ kontrollgruppe i arbeidet med modellen.

Mulige områder for forbedring

- For å sikre etterprøvbarhet kunne dokumentasjonen knyttet til utvikling, designvalg og modell vært noe mer utfyllende.
- Lånekassen kunne gjort mer for å sikre at modellen, som er bygd på grunnlag av historiske data, generaliserer godt til ukjente data.
- Lånekassen har valgt å bruke en bestemt type black box[1]-modell. Lånekassen kunne begrunnet dette valget bedre, herunder sannsynliggjøre at denne modellen gir signifikant bedre resultater enn en white box[2]-modell.
- Ettersom Lånekassen bruker en black box-modell kunne de vist tydeligere hvordan man sikrer transparens og forklarbarhet.
- Lånekassen kunne hatt større fokus på å sikre likebehandling og unngå innebygde skjevheter i modellen. Modellen overdriver blant annet betydningen av å være mann.
- Ettersom personopplysninger er inkludert i modellen kan Lånekassen med fordel diskutere denne bruken opp mot personopplysningslovens minimalprinsipp. Bruk av personopplysninger bør således gi signifikant bedre resultater enn en modell der man utelater personopplysninger.
- Lånekassen kunne vært klarere på formålet med kontrollen. Her kunne man vurdert å maksimere kost/nytte heller enn å fokusere på et absolutt antall kontroller. Det kunne potensielt gitt ytterligere gevinstrealisering.
- Forslaget om å gjøre et tilfeldig utvalg som en motvekt mot selvforsterkende mønstre *etter* at ML-modellen har valgt ut personer, istedenfor å bruke en reell kontrollgruppe, kan med fordel revurderes.
- Det kunne vært gjort en mer substansiell vurdering av hvilke variabler som inkluderes i ML-modellen og hvordan dette gjøres. Relativt mange variabler bidrar således lite til ML-modellens prediksjonskraft.

[1] Med «black box» menes modeller der det ikke er mulig å få innsyn i hvordan gitte inndata gir et gitt resultat (se også begrepsforklaringer i appendiks A1).
[2] I motsetning til «black box»-modeller er det i en «white box»-modell mulig å forklare hvordan gitte inndata gir et gitt resultat (se også begrepsforklaringer i appendiks A1).

1. Innledning

1.1 Bakgrunnen for Riksrevisjonens gjennomgang

Algoritmer for maskinlæring (ML) og kunstig intelligens (KI) er i ferd med å bli tatt i bruk i forvaltningens saksbehandling. Det er også liten tvil om at bruken av ML/KI vil øke sterkt fremover.

Dette er åpenbart bra og har mange fordeler. Man kan blant annet tenke seg at produktiviteten øker hvis mennesker blir erstattet med maskiner. Kvaliteten på saksbehandlingen kan potensielt bli bedre mht. for eksempel likebehandling, og det gir helt nye muligheter for deteksjon av avvikende hendelser, handlinger og enheter knyttet til f.eks. tildeling av ytelser. Dette er for øvrig bare noen enkle eksempler og en lang rekke use-cases kan tenkes.

Bruk av ML/KI i forvaltningen innebærer samtidig viktige prinsipielle spørsmål og mer konkrete utfordringer. Et slikt prinsipielt spørsmål er av etisk art og kan oppsummeres på følgende måte[3]:

> *Dersom man kommer dit at maskiner tar beslutninger som gjelder enkeltpersoner, hvordan skal man sikre transparens i saksbehandlingen?*

Dette spørsmålet blir viktig ettersom en del algoritmer for ML/KI er sorte bokser i den forstand at man ikke kan forklare *hvorfor* man får et bestemt utfall. Et hypotetisk eksempel: 100 personer søker om et gitt stønad og det er en maskin som behandler søknadene ved hjelp av en «sort boks algoritme». Si at 50 får innvilget stønad mens 50 får avslag. Da vil det ikke være mulig å si *hvorfor* de konkrete 50 personene fikk innvilget søknaden, mens de andre 50 fikk avslag.

Et eksempel på en annen type spørsmål er hvorvidt bruk av algoritmer og «maskinell saksbehandling» faktisk bidrar til innsparinger og økt effektivitet, som jo er en ofte brukt begrunnelse. Det kan være gode grunner til å etterprøve dette empirisk. En tredje type spørsmål er knyttet til lik behandling av like saker. Veldig forenklet og popularisert: Maskinlæring «brukt feil» kan føre til at noenlunde like saker kan bli behandlet vidt forskjellig på bakgrunn av faktorer som kjønn og etnisk bakgrunn.

Slik spørsmål vil bli svært viktige både for Stortinget og forvaltningen i tiden fremover. De blir derfor også viktige for Riksrevisjonen som Stortingets kontrollorgan.

Riksrevisjonen ønsket derfor å gjennomføre et utviklingsprosjekt der formålet var å skaffe oss erfaring med hvordan vi skal tilnærme oss en "algoritmestyrt forvaltning". Dette prosjektet ble gjennomført i samarbeid med Statens lånekasse for utdanning (Lånekassen), som kanskje er den statlige virksomheten som har kommet lengst mht. å ta i bruk maskinlæring i saksbehandlingen. Det konkrete utgangspunktet for gjennomgangen var maskinlæringsalgoritmene som Lånekassen bruker i sin bokontroll.

Denne rapporten oppsummerer således utviklingsprosjektet.

[3] Et slikt scenario vil være problematisk med tanke på GPDR, så dette er kun en grovkornet illustrasjon for å få fram poenget.

1.2 Avgrensning, mål og problemstillinger

Lånekassen har begynt å bruke maskinlæring i sin såkalte bokontroll, som handler om hvorvidt kunder som sier at de bor hjemmefra faktisk gjør det. Kort sagt, har aktuelle kunder oppgitt riktige opplysninger til Lånekassen mht. bostatus? Riktig bostatus er viktig ettersom borteboere har rett på å få omgjort lån til stipend, mens elever/studenter som bor hjemme hos sine foreldre ikke har rett på en slik omgjøring. Det maskinlæringsalgoritmen gjør er da å produsere en score pr. kunde for hvor sannsynlig det er at de mangler dokumentasjon på riktig bostatus.

Konkret har Riksrevisjonen gjort en gjennomgang av opplegget for bruk av maskinlæring.[4] Formålet med undersøkelsen har vært å teste ut hvilke typer problemstillinger som kan være fornuftige for fremtidige "revisjoner av algoritmer" mht. metodikk, tilnærming, fokus og avgrensning. Dette inkluderer også betraktninger rundt hva som kan være mulige vurderingskriterier (revisjonskriterier) for framtidige revisjoner.

Et mer konkret mål med undersøkelsen har vært å gi en vurdering av den jobben Lånekassa har gjort med å utvikle, dokumentere og kvalitetssikre modellen. Vårt fokus har vært på hvor «god» modellen er i henhold til sentrale hensyn knyttet til offentlig saksbehandling som effektivitet, transparens (særlig forklarbarhet[5] av prosesser og beslutninger) og likebehandling. Prosjektet innbefattet også en vurdering av kost/nytte knyttet til å basere bokontrollen på maskinlæring.

Vi definerte fire problemstillinger for dette prosjektet. Disse er i hovedsak besvart, men det er nok ikke alle kulepunkter under hver problemstilling som har fått like stor oppmerksomhet. Vi tenker dette er greit i denne sammenhengen ettersom vi var temmelig usikre i utgangpunktet når det gjaldt hva som var gode tema å se nærmere på.

Problemstilling 1:

Er de valg og forutsetninger som ble gjort i planleggingen av ML-systemet for bokontroll rimelige og godt dokumentert?

Her har vi i hovedsak sett på arbeidet som ble gjort ifm. planlegging og oppsett av ML-modellen. Det har blant annet handlet om:

- Evaluering av datagrunnlag, inkl. vurderinger rundt personvern
- Om ulike modellkandidater ble vurdert og begrunnelse for valg av modell
- Valg/bortvalg av variabler («features») og verdier for hyperparametre, med tilhørende begrunnelser[6]
- Om det ble gjort tilstrekkelige risikovurderinger mht. valg og oppsett av ML-modell, herunder mht. eventuelle sideeffekter

[4] Ettersom dette ikke er en revisjon, men et utviklingsprosjekt har vi valgt å bruke begreper som «gjennomgang» og «vurdering» og således unngått revisjonsbegrepet.
[5] «Forklarbarhet» handler altså om hvorvidt det er mulig å forklare hvorfor en bestemt person får et bestemt resultat
[6] I ML/KI-terminologi brukes normalt begrepet «features» for det som vanligvis er kjent som variabler. Hyperparametre er i maskinlæring en betegnelse på parameterverdier som settes *før* læringsprosessen igangsettes, og som således blir en del av forutsetningene for hvordan modellen vil oppføre seg.

Problemstilling 2:

Hvor god er maskinlæringsmodellen?

Hvor «god» en maskinlæringsmodell er, er normalt et spørsmål om prediksjonskraft. I tilfellet med bokontroll kan dette forenklet oppsummeres i «hvor godt treffer modellen mht. å flagge de som faktisk har oppgitt feil bostatus til Lånekassen?» En perfekt modell vil gi et resultat på 100 %. Dvs. alle som faktisk har oppgitt feil bostatus blir fanget opp, og *ingen* av de som har oppgitt riktig bostatus. En perfekt modell er ikke mulig å oppnå, og det blir derfor en avveining av hvor viktig det er å finne flest mulig med feil bostatus, opp mot hvor viktig det er å unngå unødvendige kontroller.

Her har vi sett se på hvor godt modellen fungerer, både som ML-modell isolert sett og sammenlignet med det tidligere systemet man hadde for utvelgelse av personer til bokontroll. Mht. sammenligning med tidligere system for bokontroll har vi også gjort noen enkle betraktninger rundt kost/nytte.

Problemstilling 3:

Har Lånekassen håndtert spørsmål omkring sikkerhet, transparens, likebehandling, "forklarbarhet" og eventuelle etiske problemstillinger på en god måte?

Spørsmål rundt sikkerhet knyttet til bl.a. håndtering av personopplysninger, lik behandling av like saker og transparens i saksbehandlingen er generelt viktige i offentlig sektor. Bruk av ML-modeller medfører imidlertid noen mer særskilte utfordringer.

For eksempel kan transparens og forklarbarhet av beslutninger basert på ML-modeller ofte være utfordrende å få til, og kan kreve spesielle tiltak.

Samspill mellom maskin (modell) og mennesker er også generelt interessant, blant annet med tanke på beslutningsansvaret til involvert personell. Dette er imidlertid mest relevant hvis modellen tar eller støtter beslutninger med vidtgående effekt, og således ikke viktig i forbindelse med Lånekassens bokontroll.

Etiske problemstillinger kan involvere utilsiktet forskjellsbehandling av ulike grupper, eller utilsiktet forsterking av strukturer som finnes i datagrunnlag modellen er trenet på.

Under denne problemstillingen vil vi også vurdere de tiltak som er gjort for å sikre stabil og betryggende drift etter at modellen er satt i produksjon.

Problemstilling 4:

Er kodekvaliteten god?

God kodekvalitet er viktig for å forstå hva som faktisk er gjort mht. oppsett av ML-modellen. Dette er en nødvendig (men ikke tilstrekkelig) forutsetning for reproduserbarhet og bidrar til transparens i den delen av arbeidet som er av teknisk art. Samtidig er god kodekvalitet viktig for modellvedlikehold, overføring av kompetanse og langsiktig stabil og pålitelig bruk av modellen.

2. Beskrivelse av området

2.1 Grunnlag og historikk for Lånekassens bokontroll[7]

Studenter som ikke bor sammen med sine foreldre, dvs. som er borteboere, kan få omgjort lån til stipend ved bestått utdanning. Studenter som bor sammen med sine foreldre, dvs. som er hjemmeboere, har ikke rett til en slik omgjøring. Det er kundens egne opplysninger i søknaden om utdanningsstøtte som avgjør om han eller hun blir registrert som borteboer eller hjemmeboer i Lånekassens saksbehandlingssystem. Her finnes det således en mulighet for at studenter oppgir feil bostatus ved at de oppgir borteboerstatus, mens de i realiteten bor hjemme. På denne måten kan de urettmessig få omgjort lån til stipend.

Grunnlag for Lånekassens bokontroll er lov om utdanningsstøtte[8] og relaterte forskrifter. Særlig forskrift om tildeling av utdanningsstøtte er her relevant.

Fra St.prp. 1 KD 2014 - 2015:

«Departementet foreslår å innføre fast, årleg kontroll av bustatus til eit større utval studentar som har opplyst å vere bortebuarar. Tiltaket blir foreslått for å redusere misbruk av ordninga som gir bortebuande studentar rett til å få omgjort delar av utdanningslånet til stipend ved gjennomført utdanning. Utplukket av dei som skal kontrollerast vil bli gjort ut frå objektive kriterium knytte til alder og avstand mellom lærestad og folkeregistrert adresse, slik at ein så langt som mogleg berre kontrollerer dei som har hatt eit realistisk høve til å skaffe seg stipendet på urettmessig vis. Det er berekna at kontrollen vil gjelde om lag 54 000 studentar. Basert på tidlegare gjennomførte kontrollar av den bustatusinformasjonen som studentane har gitt, legg departementet til grunn at den auka kontrollen vil føre med seg lågare omgjering av lån til stipend. Departementet foreslår å redusere løyvinga med 72 mill. kroner som følgje av forslaget.»

Lånekassen gjennomførte stikkprøvebaserte bokontroller av 1 000 studenter i 2008, 2011 og 2012. I 2015[9] gjennomførte de før første gang en fullskala bokontroll, av 48 000 studenter som hadde oppgitt til Lånekassen at de var borteboere i 2014. Dette var et krav i KDs budsjettproposisjon for 2015, og beregnet innsparing ved at lån urettmessig ikke ble omgjort til stipend var 72 mill. kroner. Lånekassen fikk 8,5 mill. kroner i tilleggsbevilgninger for å gjennomføre fullskalakontrollen. Formålet var altså å bruke 8,5 mill. for å spare 72. mill.

Bokontrollen omfatter de man tenker har en realistisk mulighet til å bo hjemme, men som samtidig oppgir at de er borteboere. Det viktigste vurderingskriteriet her avstand mellom lærested og foreldrenes hjemmeadresse.

Antallet og andelen kunder som i tidligere kontroller har blitt identifisert med feil bostatus fremkommer av tabellen under. Som vi ser varierer andelen identifiserte mellom 3,6 og 4,6 %, og det totale antallet når man kontrollerte alle i 2014 var 2396 personer. Innsparingen for 2014 er

[7] Dette avsnittet er i stor grad basert på, og til dels direkte hentet fra Lånekassens egen rapport «Bokontrollen for 2014 - Planlegging, gjennomføring og resultater av kontrollen» [Ref.3]
[8] https://lovdata.no/dokument/NL/lov/2005-06-03-37
[9] Fullskalakontroller ble også gjennomført i 2016 og 2017

av Lånekassen beregnet til brutto 55, 5 mill. kroner (utgiftene til fullskalakontroll på 8,5 mill. kommer til fratrekk).

Tabell 1 Identifiserte misligholdere i tidligere kontroller

Kalenderår**	2008*	2011*	2012*	2014	2015	2016	2017
Antall kontrollerte kunder	1000	1000	1000	47962	31178	42963	25000 (av 58553 kandidater)
Antall identifiserte misligholdere[10]	45	36	45	2396	1592	2732	2254
Prosent av de kontrollerte kundene	4,5	3,6	4,5	5,0	5,11	6,36	9,02 (3,85% av alle kandidater)

* Stikkprøvekontroller

** Kontroll i av f.eks. 2017 var for hele kalenderåret 2017, og omfatter vårsemesteret for *undervisningsåret* 2016-2017 og høstsemesteret for *undervisningsåret* 2017-2018

Fullskalakontroller er, tross positiv kost/nytte, krevende og kostbart. Lånekassen antok derfor i 2017 at det kunne være rom for å redusere ressursbruken til bokontroll ved å benytte en prediksjonsmodell for uttrekk basert på «sannsynlighet for å ha oppgitt feil bostatus». Da kunne det potensielt være mulig å kontrollere langt færre, samtidig som antallet man avdekket med feil bostatus ikke var vesentlig lavere. Lånekassen gjennomførte derfor i 2018 et «proof-of-concept» for å se på muligheten for å bygge en slik prediksjonsmodell, basert på maskinlæring. Modellen ville da kunne gi hver enkelt «kandidat» en score som ga beregnet sannsynlighet for at hen hadde oppgitt feil bostatus. De med høyest score kunne da bli bedt om å dokumentere at de faktisk var borteboere.

2.2 Generelle krav til saksbehandling

Mht. saksbehandling så skal også ML-assistert saksbehandling være i henhold til krav som stilles i relevante lover og forskrifter, herunder forvaltningsloven. Samtidig er det som kan betegnes som «god forvaltningsskikk» i stor grad ulovfestet, og hva som er god forvaltningsskikk mht. bruk av maskinlæring/kunstig intelligens er ikke alltid åpenbart ettersom det det er maskiner som, i større eller mindre grad, står for saksbehandlingen.

«Bokontroll» er ikke i seg selv et enkeltvedtak. Det må anses som intern saksforberedelse der et enkeltvedtak *kan* være utfallet, dersom man finner at en student ikke kan dokumentere oppgitt borteboerstatus. Vi anser derfor forvaltningsloven som lite relevant for bokontrollen isolert sett, ettersom det ikke bør spille noen stor rolle for Lånekassens kunder *hvordan* man kom fram til listen over kunder som potensielt har oppgitt feil bostatus.

Mht. regler for «god forvaltningsskikk» er det vanskelig å gi en uttømmende liste, men prinsipper som ofte blir listet opp er[11]

- En åpen og forståelig forvaltning
- Likebehandling
- Forsvarlig saksbehandling

[10] Med «misligholdere» betegnes kunder som ikke leverer dokumentasjon om borteboerstatus

[11] Kilder: Etiske retningslinjer for statstjenesten; NOU 2019:5 Ny forvaltningslov kap. 10.3, 10.8 og 11.7

- Nøytralitet og saklighet
- Forholdsmessighet
- Effektivitet

For eksempel, i en maskinlæringskontekst kan åpenbart prinsippet om en åpen og forståelig forvaltning bli utfordret dersom komplekse black box-modeller er brukt. Samtidig kan prinsippet om effektivitet fremme bruk av black box-modeller i en del sammenhenger. Videre kan «likebehandling» defineres og måles på ulike måter, der ulike behov mht. likebehandling kan være vanskelig å oppfylle samtidig (se appendiks A2 og A3.2 for detaljer). Det er derfor viktig å teste, forstå og dokumentere hvordan maskinlæringsmodeller oppfører seg.

3. Gjennomgang av maskinlæringsmodellen

3.1 Datagrunnlag for modellen

ML-modellen som Lånekassen bruker til bokontroll ble som nevnt over utviklet i 2018, med en viss grad av videreutvikling senere. Dette utviklingsarbeidet pågikk imidlertid når vårt prosjektet startet opp og vi valgte derfor i samråd med Lånekassen å se på den opprinnelige modellen fra 2018.

Datamaterialet består i hovedsak av

a) Datagrunnlaget, der enheten er perioder og opplysninger om både studieperioder og studentene/elevene er variabler.
b) Selve modellen inkl. kodebase

Datagrunnlaget består av alle studenter/elever som har mottatt utdanningslån som borteboer, men som potensielt kan bo hjemme på grunn av tilstrekkelig geografisk nærhet mellom hjemmet til foreldrene og studiested. I det videre vil disse studentene bli omtalt som *kandidater.* En kandidat vil typisk få studiestøtte i flere *perioder,* der en periode typisk er et semester.

Datagrunnlaget består av *utviklingsdata* (trenings- og testdata), som man bruker til å bygge ML-modellen, og *produksjonsdata*, som er nye, ukjente data man bruker modellen på. Utviklingsdata er data på kandidater fra fullskalakontrollene for 2014-2016, hvor man også har opplysninger på hvorvidt de kunne dokumentere borteboerstatus eller ikke. En slik populasjon med kjent utfall for enhetene er et godt grunnlag å bygge en ML-modell på. Datasettet for 2014-2016 ble videre delt i to, i en 80/20 splitt. 80 % av datagrunnlaget er brukt til å trene modellen, mens de resterende 20 % er brukt som testdata for å validere modellen.[12]

[12] Sagt på en annen måte: Man begynner med et datasett hvor man kjenner utfallet (fra fullskalakontrollene 2014-2016). Så bruker man mesteparten av datasettet (80 %) til å trene modellen. Men så må man vite hvor godt modellen treffer. For å teste dette bruker man de resterende 20 % av materialet.

Datasettet for 2017 er produksjonsdata og består av kandidater *man ikke kjente utfallet for* mht. reell bostatus. Modellen brukes for å predikere sannsynligheten for at disse kandidatene har oppgitt feil bostatus og består av totalt **58553 kandidater** og **82283 perioder**.

Et lite poeng her er at dokumentasjonen fra Lånekassen ikke inneholder noe om eventuelle endringer i variablenes fordelinger og deres prediksjonskraft for de aktuelle årene 2014-2016. Slike vurderinger bør generelt gjøres når man skal trene en modell på historiske data. Et eksempel: Når Lånekassen skal kjøre ML-modellen for bokontroll på 2018-kandidatene, så vil 2017 kunne brukes som treningsdata. Det kan da være lurt å se om betydningen av de variablene som er inkludert har endret seg i perioden 2014-2017. Dette vil være et godt tiltak for å redusere risikoen for over-/undertrening av modellen. Man kan her også tenke seg enkle tester for å vurdere hvor mange (historiske) årsklasser man skal bruke i treningen av neste års modell.

Lånekassen angir at studenter kan ha problemer å fremskaffe dokumentasjon selv om de faktisk er borteboer. F. eks. kan en student bo sammen med andre i et kollektiv og ikke har leiekontrakt selv. Dette er en potensiell usikkerhet knyttet til datamaterialet. Hvis det er slik at enkelte grupper har større utfordringer med å skaffe dokumentasjon enn andre, selv om de er faktisk borteboende, så kan det føre til en skjevhet i data fra tidligere bokontroller. Lånekassen motvirker dette gjennom lempelige dokumentasjonskrav, så risikoen kan anses som lav.

En annen potensiell svakhet mht. datagrunnlag er at dokumentasjon på borteboerstatus kan sendes inn lenge etter at søknaden har blitt avslått, og vedtaket blir da omgjort. Det medfører en usikkerhet i resultatene for bokontrollene for 2014-2016. Det kunne her vært dokumentert hvor mange som har levert dokumentasjon innen fristen på 3 uker og hvor mange som har levert senere, f.eks. innen purrefristen og innen fristen for neste treningsløp for modellen. Utfra dette kunne man estimert usikkerhet i resultatet av kontrollen som har opphav i eventuelle endringer i datagrunnlaget.

Datagrunnlaget inneholder personopplysninger, men ikke opplysninger innen «særlige kategorier»[13]. Det inneholder heller ikke direkte identifiserbare personopplysninger, men Lånekassens prosess for datauttrekk innebærer en midlertidig håndtering av personnummer. Forhold rundt håndtering av personopplysninger er tatt opp med Lånekassen og vi har ingen grunn til å stille spørsmål ved hvordan dette er gjort.

Alle variabler som er brukt i modellen har for øvrig blitt vurdert av juridisk ekspertise i Lånekassen før bruk.

3.2 Om ML-modellen

Når man skal ta i bruk maskinlæring på et nytt område så vil det som regel være flere modeller som kan være aktuelle å bruke. Valg av modell kan ha stor betydning for resultatet og det er ikke alltid åpenbart hvilken modell som er den beste i et gitt tilfelle. Det er således viktig at man her gjør et informert valg. Og, skal man kunne gjøre et «informert valg» mht. modell så krever dette relativt høy kompetanse på maskinlæring.

[13] «Særlige kategorier av personopplysninger» er i GDPR-terminologi det som tidligere normalt var omtalt som «sensitive personopplysninger»

Modellen Lånekassen brukte heter «Catboost», og er en såkalt «gradient boosting»-modell basert på beslutningstrær. Catboost har innbygget håndtering av kategoriske variabler og har gode resultater i standardiserte benchmark-tester, sammenlignet med andre modeller.[14] Catboost er dermed i utgangspunktet en modell som passer godt til Lånekassens datagrunnlag, som består for en stor del av kategoriske variabler.

Mht. «informert valg» spurte vi derfor Lånekassen hvorfor de valgte Catboost og ikke en annen mulig modell. Da var svaret at Catboost var modellen som var anbefalt av de konsulentene som hjalp Lånekassen med å komme i gang med maskinlæring på bokontrollområdet. For en proof of concept-øvelse er dette greit nok, men for ML-modeller som skal brukes i produksjon kunne nok en noe mer grundig vurdering mht. modellvalg vært gjort.

En annen viktig diskusjon rundt modellvalg er det man kan kalle «black box vs. white box». Catboost er black box modell, noe som innebærer at det ikke er mulig å vite sikkert hvorfor man får den outputen man får. Catboost er ikke uten videre «forklarbar». Det er således ikke uten videre mulig å si *hvorfor* en gitt person fikk en høy sannsynlighetsscore og en annen lavere. I tilfellet med bokontroll er det kanskje ikke noe stort problem at man ikke kan vite sikkert hvorfor man får en gitt output. Som sagt, man blir bare bedt om å dokumentere det man tidligere har oppgitt til Lånekassen. I andre tenkte tilfeller, og særlig der output fra en ML-modell blir brukt til å fatte beslutninger, kan det være problematisk å ikke kunne forklare output.

Således, når man skal ta i bruk maskinlæring bør man begrunne hvorfor man bruker en black box-modell. Man bør også sannsynliggjøre at den vil gi bedre prediksjoner enn en white box modell ettersom det ikke gitt at det faktisk er slik.[15] Her er det også mulig at en godt gjennomført white box-analyse vil gi bedre resultat enn en middels gjennomført black box-analyse. Lånekassen har ikke gjort noen slike sammenligninger mellom Catboost og alternative white box-modeller. Vi har imidlertid gjort det som en del av dette prosjektet, se kap. 4 for detaljer.

3.3 Reproduserbarhet

Modellen ble som nevnt over utviklet av konsulenter i samarbeid med Lånekassen, og de som jobber med dette hos Lånekassen synes å ha god forståelse av kode og av utviklingsprosessen. I den dokumentasjonen vi har fått fra Lånekassen finnes det imidlertid lite informasjon om kvalitetssikringstiltak for kodebasen. Kvalitetssikring av kode må sies å «alltid» være viktig, ikke minst når sentrale deler av et prosjekt er gjennomført av eksterne konsulenter.

Den beste måten for å teste om en modell gir de resultatene som eier av modellen påstår, er at noen eksterne reproduserer resultatene. Dette skal i utgangspunktet være relativt greit om man har samme modell, datagrunnlag og kodebase som eier av modellen, og dette er godt dokumentert.

Vi fikk således overlevert både modell, datagrunnlag og kodebase, med tilhørende dokumentasjon fra Lånekassen, med formål å reprodusere resultatene. På tross av at dokumentasjonen var rimelig

[14] Se www.catboost.ai for en oversikt over benchmarkresultater
[15] Se for eksempel https://www.ncbi.nlm.nih.gov/pubmed/30763612

god mht. hva de ulike delene av koden gjorde, så viste det seg likevel vanskelig å få til en nøyaktig reproduksjon, av ulike årsaker.

For det første: I kodebasen fastsettes det enkelte parametere som Catboostmodellen ikke kjenner. I et enklere språk kan dette sammenlignes med å si at «denne innstillingen skal stå på 12», men uten at innstillingen finnes.

For det andre: Kodebasen er i seg selv litt forvirrende, siden flere parameterverdier for trening av modellen er satt i en konfigurasjonsfil, men uten at disse faktisk blir brukt. Igjen veldig forenklet: Det blir som å definere at «denne innstillingen skal stå på 12», men uten å angi hvor innstillingen skal brukes. På tross av at de ikke er brukt er altså de aktuelle parameterverdiene inkludert i dokumentasjonen. Det er derfor noe uklart hvordan ulike deler av koden passer sammen, og hvordan koden passer sammen med dokumentasjonen. Her kunne nok dokumentasjonen vært litt bedre. I alt kunne Lånekassen ha vært klarere i sin dokumentasjon og kommunikasjon ovenfor Riksrevisjonen ved overlevering av kode, modell og datagrunnlag – spesielt med hensyn på versjonering. På grunn av det som sannsynligvis er feil og kvalitetsproblemer i koden oppstod det tvil om det som var overlevert hørte sammen, eller om det var versjoner fra forskjellige tidspunkt.

Det er også enkelte andre forhold som ikke var dokumentert og som derfor bidro med vanskeligheter mht. reproduksjon. For eksempel er enkelte defaultverdier i Catboost avhengig av om modellen kjøres på CPU- eller GPU-hardware[16], uten at det var dokumentert hva Lånekassen hadde brukt. I den Catboost-versjonen som Lånekassen brukte ble i tillegg variabler inkludert basert på rekkefølge i koden og ikke basert på variabelnavn. Hvis rekkefølgen til variablene da ikke var eksakt den samme for treningsdata og produksjonsdata så ville modellen gi meningsløse prediksjoner. Rekkefølgen på variablene var heller ikke godt dokumentert.

Etter en del prøving og feiling klarte vi å reprodusere resultatet for 99,7 % av *periodene* og med litt bedre resultat mht. prediksjon enn det Lånekassen fikk. Dette må anses som bra, selv om vi ikke klarte å få til identiske resultater. Vi brukte dog en god del tid på dette, og det kan nok tenkes at dokumentasjonen av modellutviklingen kunne vært noe mer utfyllende.

3.4 Valg av variabler

Med tanke på resultatet av en ML-modell så er naturlig nok input, i form av hvilke variabler som inkluderes, viktig. Lånekassens modell inneholder 35 ulike variabler, herunder informasjon om tidligere kontroll, studieretning, bosted og finansielle opplysninger, men også mindre intuitive variabler som fødselsmåned, og variabler med overlappende informasjon som postnummer og kommunenummer.

Betydning av de fleste variablene er forklart i dokumentasjonen, men variabellisten beskrevet i dokumentasjonen er ikke identisk med variablene som faktisk er brukt[17]. Det kunne således vært bedre dokumentert hvilke variabler som faktisk inngår i modellen.

[16] CPU er vanlige mikroprosessorer, mens GPU er prosessorer du finner i grafikkort.
[17] F.eks. kunde_antall_status_a/b/d/h/i er ikke i dokumentert i variabel-listen «Flight 1» i avsnitt 5.1 i Lånekassens Behovs- og Løsningsbeskrivelse [Ref. 1]

Begrunnelser særlig for ikke-intuitive variabler er et poeng ettersom det er inkludert en rekke variabler som nesten ikke bidrar med prediksjonskraft i det hele tatt, verken alene eller i kombinasjon med andre variabler. Dette ser vi av figurer for hhv. «feature importance» og «feature interaction», se appendiks A4[18].

Når man bygger en modell er det generelt viktig å gjøre en vurdering av hvilke variabler som skal inkluderes, om variabler skal transformeres eller rekodes på noen måte mv. For eksempel kan inkludering av mange variabler som bidrar med lite informasjon føre til unødvendig lang kjøretid for beregninger, høyere risiko for en overtrent eller skjev modell, og lavere grad av forklarbarhet.

3.5 Overtrening

Det som er kjent som «overtrening» er et vanlig problem når man jobber med maskinlæring. Enkelt forklart handler dette om at modellen lærer seg detaljerte mønstre som finnes i data den er trent på, men som kanskje ikke fines i nye, ukjente data. Dette kan medføre at modellen ikke er allmenn nok til å beskrive generelle sammenhenger i hele populasjonen. [19]

En modell som således ser veldig bra ut mht. utviklingsdata kan gi dårlige resultater på nye data. Derfor deler man ofte utviklingsdata i trenings- og testdata, lar modellen bare lære fra treningsdata og vurderer treffsikkerhet på testdata. Forskjell mellom modellens oppførsel på trenings- vs. testdata gir da en indikasjon på hvor godt modellen generaliserer til nye data.

Lånekassen har brukt data for årene 2014-2016 for å predikere hvem som er «sannsynlige misligholdere» i 2017.

Et spørsmål da er hvor godt modellen basert på tall for 2014-2016, og uten spesielle tiltak mot overtrening passer for 2017. Her er konklusjonen at resultatene ikke er fullt så gode på produksjonsdata som forventet utfra resultatene på test-data. Det ser således ut til at modellen er noe overtrent (se tabell 2 i avsnitt 4, og appendiks A3.1 for detaljer).

3.6 Kvalitetsikring

For å unngå og trene fremtidige modeller bare på data den allerede kjenner, plukker Lånekassen tilfeldige kandidater for kontroll, i tillegg til kandidater valgt av ML-modellen. P.t. er det 10000 tilfeldige kandidater i kontrollgruppen og 15000 kandidater i ML-gruppen, totalt 25000 som blir kontrollert. Dvs. i underkant av 50 % av det totale antallet kandidater i 2017-datasettet. I kontrollen av 2017-data ble kontrollgruppen plukket først. Dette er fornuftig ettersom kontrollgruppen ikke lenger ville være et tilfeldig utvalg dersom man først skulle plukke 15 000 utfra ML-modellens sannsynlighetsscorer.

[18] Disse figurene indikerer at variabler kan droppes, men uten at det er noe bevis. En slik indikasjon burde imidlertid medført at saken ble undersøkt videre

[19] Undertrening finnes også, og er en situasjon der modellen du har laget er for generell. Da har du ikke klart å fange opp mønstre av interesse som faktisk finnes. Dette er vanligvis et mindre problem og vil derfor ikke bli omtalt.

Ifølge dokumentasjonen fra Lånekassen står det imidlertid at ML-gruppen skal bli plukket først i senere kontroller, og kontrollgruppen plukkes fra resten, som da er kandidater som ML-modellen gir lav sannsynlighet.[20] Det fører til at denne «kontrollgruppen» ikke lenger er representativ for hele populasjonen. ML-modellens utplukk vil i så tilfelle vise kunstig gode resultater sammenlignet med «kontrollgruppen», nettopp fordi alle som er i «kontrollgruppen» vil ha lav sannsynlighet for å ha oppgitt feil bostatus. Lånekassen kunne derfor vært klarere i sin bruk av ordet «kontrollgruppe». Det kunne også vært klargjort at gruppen i senere kontroller således ikke vil være en reell kontrollgruppe, men fungere som et motvekt mot selvforsterkende mønstre i datagrunnlaget. En slik motvekt er ikke ufornuftig, men man vil ikke lenger ha en nøytral "baseline" som man kan sammenligne resultatene fra ML-modellen med. Vi mener således at en nøytral baseline er viktig.

En mulig løsning her er at både ML-modell og tilfeldig utvalg plukker fra hele populasjonen. Da vil man få noe overlapp, men man sikrer både at ML-modellen plukker ut de med høyest sannsynlighet, og man vil beholde en reell tilfeldig valgt kontrollgruppe. Det vil også bety at man vil kontrollere et noe lavere antall.

4 Hvor god er Lånekassens modell?

I Lånekassens tilfelle var utgangpunktet for å ta i bruk en ML-modell en fullskalakontroll av *alle.* Da må man anta at man også avdekker alle som har oppgitt feil bostatus, og formålet med å bruke en ML-modell kan således ikke være å avdekke flere. Man kan gjennom bruk av ML imidlertid potensielt avdekke nesten like mange ved å kontrollere vesentlig færre. Lånekassen har selv sagt [Ref. 4] at de anslagsvis kunne avdekke 7 av 10 misligholdere ved å kontrollere 75 % færre. En rimelig tolkning er således at Lånekassen ønsket å bruke minst mulig ressurser på kontroll, og samtidig finne flest mulig kunder uten dokumentasjon på bostatus.

Målet er således å finne en «best mulig» modell. Så finnes det noen alternativer for hvordan dette skal måles, ettersom det kommer an på hva man vil maksimere.[21]

4.1 Målparameter og metodikk

Om man ønsker å effektivisere kontrollen i betydningen «finne samme antall antatte misligholdere med færre kontroller» så vil *presisjon*[22] være måleparameteren man bør være opptatt av. Dog, hvis formålet er høyest mulig kost/nytte, så vil det sannsynligvis være bedre å skjele til *sensitivitet*[22], ettersom kostnaden ved å kontrollere én ekstra er veldig liten, sammenlignet med innsparingen man får ved å finne en ekstra misligholder.

[20] Se Lånekassens Brukerveiledning [Ref.2], kapittel 5, s. 26
[21] De vanligste måleparameterne er presisjon (andel treff av alle kontrollerte), og sensitivitet (andel av alle «misligholdere» som er funnet i kontroll). Lånekassen bruker AUC som kombinerer sensitivitet med falske positive rate (andel unødvendig kontrollerte av alle som har angitt riktig bostatus).
[22] Se appendiks A2 for definisjoner av relevante måleparameter. Sensitivitet også kalles «recall» eller «TPR» (true positive rate)

Lånekassen plukket ut de 15 000 kundene fra 2017-data med høyest predikert sannsynlighet for å ha oppgitt feil bostatus. I tillegg plukket de ut 10 000 kunder tilfeldig til kontroll.[23] Grenseverdien for modellens sannsynlighets-score er dermed definert som 15000-største sannsynlighet[24]. Kunder med større sannsynlighet er klassifisert som «antatte misligholdere», kunder med lavere sannsynlighet er predikert som «antatt riktig bostatus».

Resultatene for de 10 000 tilfeldige kan brukes som målestokk for ML-modellen, fordi de ligner hele populasjonen. Resultatene for de 15000 kundene som ble plukket av modellen kan imidlertid ikke brukes for å evaluere modellen, fordi kunde-populasjonen her inneholder bare de som er predikert «antatte misligholdere». Vi har ingen informasjon om faktisk bostatus for kunder som er predikert som har angitt riktig bostatus, ettersom disse ikke ble kontrollert. I det følgende er derfor resultater på produksjonsdata alltid evaluert opp mot de 10000 tilfeldig valgte kundene.

4.2 Resultat

Tabellen under viser sensitivitet og presisjon av Lånekassens modell, både for tilfeldig utvalg fra produksjonsdata og for utviklingsdata:

Tabell 2 Modellens treffsikkerhet på utviklings- vs produksjonsdata

Måleparameter	Lånekassens ML-modell Produksjonsdata (2017)	Lånekassens ML-modell Utviklingsdata (2014 - 2016)
Sensitivitet	0,63	0,93
Presisjon	0,10	0,15

Som nevnt i avsnittet om overtrening gir modellen bedre resultat på utviklingsdata enn på produksjonsdata, og en sammenligning av hvordan modellen oppfører seg på hhv. trenings- og testdata viser at Lånekassens modell kan anses som overtrent (jf. appendiks A3.1).

Som en ekstra test gjorde vi noen enkle tiltak[25] for å redusere overtreningen. Tabellen nedenfor sammenligner *forskjell* i sensitivitet og presisjon på utviklings- og produksjonsdata for vår alternative modell med Lånekassens modell. Vi viser her verdiene for produksjonsdata *minus* verdiene for utviklingsdata. Negative verdier betyr altså at måleparameterverdien er lavere på produksjonsdata enn utviklingsdata.

Med mindre overtrening generaliserer modellen bedre til nye data[26].

[23] De 10 000 tilfeldige ble plukket utfra hele populasjonen på 58 553 personer. Det var altså de resterende 48 554 som var grunnlag for å plukke 15 000 basert på sannsynlighetsscore
[24] På kunde nivå. Merk at modellen gir sannsynlighet på periode nivå, etter det er maksimal sannsynlighet per kunde evaluert. I tillegg er 15000-største tatt av alle kunder uten de som er i tilfeldig utvalg, fordi tilfeldige ble plukket først
[25] Bare 19 variabler ble brukt, bruk av regulariseringsparameter l2_leaf_reg=32 (som i dokumentasjon og i gridsearch, men ikke i Lånekassens endelige modell), max_ctr_complexity=2, maksimal 1000 iterasjoner. Merk at disse parameterverdier ikke er optimert.
[26] Denne testen burde strengt tatt gjøres på trenings- vs. testdata, elle trening/test- vs. separate valideringsdata i utviklingen av modellen.

Tabell 3 Forskjell treffsikkerhet på utviklings- og produksjonsdata, mindre overtrent modell vs. Lånekassens modell

Måleparameter	Alternativ ML-modell v. RR Produksjon minus Utvikling	Lånekassens ML-modell Produksjon minus Utvikling
Sensitivitet	-0,12	-0.30
Presisjon	-0,01	-0,05

Med dette som utgangspunkt kommer vi til et reelt interessant spørsmål: Er Lånekassens black-box modellen signifikant bedre enn en forklarbar white-box modell? Vi har testet dette ved å lage en alternativ modell basert på en (white box) beslutningstremodell[27]. Vi brukte ikke mye tid på å optimalisere denne modellen, så vi antar at man kan få bedre resultater med en bedre optimert white box modell.

Som måleparameter brukte vi sensitivitet. Bare 16 inputvariabler[28] ble brukt og ingen variabler som inneholdt «direkte» personopplysninger[29] ble inkludert.

Tabell 4 Treffsikkerhet av en white-box modell vs. Lånekassens black-box modell

Måleparameter	Alternativ white box-modell Utviklingsdata vs. Produksjonsdata	Lånekassens ML-modell Utviklingsdata vs. Produksjonsdata
Sensitivitet	0,83 vs. 0,67	0,93 vs. 0,63
Presisjon	0,06 vs. 0,08	0,15 vs. 0,10

Om vi sammenligner denne white box modellen med Lånekassens (black box) modell, så ser vi at white box modellen gjør det litt dårlige enn Lånekassens modell på trenings-/testdata, men forskjellen i sensitivitet er ikke veldig stor. På produksjonsdata viser white box modellen bedre sensitivitet, fordi den enklere modellen generaliserer bedre til de ukjente data. Alt i alt er det således vanskelig å si at black box modellen til Lånekassen er signifikant bedre enn en forklarbar white box modell.

Man skal være forsiktig med å generalisere, men en konklusjon kan være: Black box kan åpenbart være bedre enn white box mht. prediksjonskraft. Dette forutsetter imidlertid at man har tid, ressurser og kompetanse til å optimalisere modellen. Har man ikke det, så kan det godt tenkes at en enklere white box modell fungerer like bra. White box har også en åpenbar fordel i at det er mulig å forklare hva som skjer inne i modellen. Mht. potensielle problemer knyttet til bruk av black box modeller i offentlig sektor så tenker vi at man bør kunne vise rimelig tydelig at black box er signifikant bedre enn white box, dersom man skal bruke slike modeller.

Hvor mange skal man kontrollere?

[27] *R-pakken `caret`* ble brukt her
[28] Mot 35 i Lånekassens opprinnelige modell
[29] Alle variabler kan anses som personopplysninger avhengig av kontekst. Med «direkte personopplysninger» mener vi demografiske variabler som alltid må oppfattes som personopplysninger. (Her: Variablene kjønn, alder, fødselsmåned, postnummer, kommunenummer og foreldres kommunenummer er ikke brukt. Formue og inntekt er delt inn i hhv. 3 og 5 kategorier.)

Som sagt over så kontrollerte Lånekassen totalt 25 000 personer i 2017, hvor 15000 var plukket ut av ML-modellen, og 10 000 ble plukket ut tilfeldig. Noe forenklet kan man si at Lånekassen klarte å avdekke nesten like mange «antatte misligholdere», samtidig som de kontrollerte 25 000 i 2017, i stedet for nærmere 50 000 i 2016. Dette er selvsagt bra.

Samtidig, det er ikke sikkert at 25 000 personer er det optimale antallet om maksimal kost/nytte er formålet. Her er det et poeng at det koster svært lite å kontrollere én person ekstra, sammenlignet med den potensielle gevinsten, dersom denne personen faktisk ikke kan dokumentere borteboerstatus. Således, hvis alle kandidatene i datamaterialet får tildelt en sannsynlighetsscore, hva er fornuftig cut-off mht. antallet man ønsker å kontrollere? Det er mulig å gjøre noen slike beregninger, og la resultatene fra ML-modellen i større grad styre hvor cut-off settes.

Man kan også diskutere om det er lurt å kontrollere 10 000 tilfeldig utvalgte kandidater. Veldig mange av disse vil nødvendigvis ikke være misligholdere. Her må vi imidlertid huske på at det er to gode grunner til å velge 10 000 tilfeldige. For det første, disse 10 000 gir oss en baseline for å kunne vurdere hvor god modellen er. Det er viktig. For det andre, disse 10 000 fra 2017 vil inngå som treningsdata for 2018. Det gjør at modellen for 2018 ikke bare trener på modellresultater fra året før. Man unngår dermed faren for å miste nye mønstre og faren for å overdrive de mønstre ML-modellen allerede har lært seg.

5. Vurderinger av etiske spørsmål

Som nevnt over er modellen Lånekassen bruker (Catboost) en black box modell. Det er derfor ikke helt godt å forklare hvorfor modellen predikerer som den gjør. I Lånekassens tilfelle er ikke dette noe stort problem ettersom resultatene kun blir brukt til å plukke ut personer som blir bedt om å dokumentere det de tidligere har oppgitt til Lånekassen.

Samtidig vil «alltid» forklarbarhet og likebehandling være viktig. Dersom f.eks. noen skulle spørre om hvorfor de ble plukket ut til kontroll, så er det ikke urimelig å forvente at Lånekassen kan svare. På samme vis, dersom modellen skulle diskriminere enkelte grupper, f.eks. ved å overdrive betydningen av et gitt kjennetegn så kan det være problematisk. En slik diskriminering vil også kunne ha betydning for hvor god modellen er jf. kapittel 4. Og uansett, dersom det er slik at modellen faktisk er diskriminerende så er det greit å vite.

5.1 Om forklarbarhet

Forklarbarhet handler enkelt sagt om at det bør være mulig å forklare hvorfor modellen gir det resultatet den gjør. Selv om dette kan være vanskelig for black box modeller, så finnes det teknikker og analyser man kan gjøre for å få innsikt i «hva som skjer inne i den sorte boksen».

Et sted å begynne er å se på hvor viktig hver variabel er totalt sett mht. å «forklare» resultatet fra modellen. Dokumentasjonen fra Lånekassen inneholder noe informasjon om dette i form av et «feature importance plot», jf. figur 6 in appendiks A4). Denne figuren viser imidlertid bare hvor mye

resultatet totalt sett endrer seg med utgangspunkt i en variabel[30]. Man har heller ikke informasjon om i hvilken retning resultatet endrer seg, verken totalt sett eller på personnivå.

En første, enkel tilnærming kunne være å vise variablenes globale betydning for resultatet. Man kan også på overordnet nivå se på fordeling av riktige eller falske prediksjoner avhengig av ulike variabler. For Lånekassens modell kan man på denne måten for eksempel se at sannsynlighet for å bli plukket for kontroll er høyere for menn enn for kvinner, høyere for lærlinger og høyere for studenter på lavere grad.

Samtidig, informasjon om variablenes betydning totalt sett kan ikke si noe hvor viktig hver enkelt variabel er for sannsynlighetsscoren på personnivå. Derfor, selv om en konkret person er mann og lærling, så kan ikke en betraktning om variablenes generelle betydning fortelle oss om han ble valgt ut *fordi* han er mann og lærling. Eller om det var på grunn av noe helt annet. Det er dette som er den sorte boksen. Vi vet rett og slett ikke, og det finnes ingen enkel måte som kan gi oss et sikkert svar.

Det finnes dog flere ulike metoder som kan gi oss informasjon om variablers lokale betydning, dvs. betydning for enkeltpersoner. To slike er hhv. LIME[31] og SHAP values[32]. I dokumentasjonen vi fikk fra Lånekassen er det ikke inkludert noe materiale som omhandler bruk av slike metoder for å øke «forklarbarheten» til ML-modellen. Det ser således ikke ut til å være gjort, og er nok noe som kunne vært gjort.

Vi har derfor selv gjort en analyse av forklarbarhet ved bruk av SHAP values. For det første kan SHAP values gi oss mer detaljert informasjon om variablenes generelle betydning ettersom vi blant annet kan få plot der alle enhetene (personene) er plottet inn pr. variabel. Vi får også ut informasjon om *retning*. Dvs. om en variabel bidrar positivt eller negativt totalt sett, f.eks. om økning i utdanningsnivå bidrar til større eller mindre sannsynlighet for å ha oppgitt feil bostatus. For det andre, og viktigere, SHAP values gir oss informasjon om hvilken betydning de enkelte variablene *sannsynligvis* har[33] for enkeltpersoner. Catboost har i nyere versjoner innebygd funksjonalitet for SHAP values, og man kan få ut resultater som illustrert i figuren 1 under.[34]

Vi ser at denne personen har fått en «output value» på -0,66, som da er relatert til sannsynlighet for å ha oppgitt feil bostatus. Denne er høyere enn «base value» på -3,308, som er relatert til gjennomsnittlig beregnet sannsynlighet. Denne personen har altså en god del *høyere* predikert sannsynlighet[35] for å ha oppgitt feil bostatus enn gjennomsnittet. Vi får også oppgitt hvilke variabler som drar i hvilken retning. Denne personen har tydeligvis vært kontrollert tidligere og alt var da «ok», mens bl.a. variablene kjønn og postnummer drar opp sannsynligheten.

[30] Det fins ulike metoder for å beregne «feature importance», og i fravær av dokumentasjon er det vanskelig å si hvilken metode ble brukt. Vanlig for beslutningstrær baserte modeller som Catboost er også å måle hvor ofte en variabel er brukt for å dele en tre.
[31] Se: https://github.com/marcotcr/lime
[32] Se: https://github.com/slundberg/shap
[33] Alle slike eksisterende metoder, herunder både SHAP values og LIME, er i seg selv modeller. De gir således ikke 100 % pålitelige og endelige svar som forklarer en black box-modell. De gir oss således mer informasjon, men vi kan fortsatt ikke vite sikkert.
[34] Denne funksjonaliteten var imidlertid ikke tilgjengelig når Lånekassen laget den opprinnelige modellen i 2017.
[35] Sannsynlighet er med sigmoid(-0,66)=0,34 ca en faktor 10 høyere enn basis verdi sigmoid(-3,308)=0,035

Figur 1 "Shap values" for et "true positive" eksempel (klassifiseringen her er at «dokumentasjon mangler», og dette viste seg å være riktig)

5.2 Om likebehandling

Likebehandling i en ML-kontekst betyr enkelt forklart at en modell behandler ulike grupper eller personer på samme måte. Ulike former for gruppetilhørighet kan ofte være variabler som har forklaringskraft og som derfor bør inkluderes i modellen. F.eks. dersom det faktisk er slik at menn reelt sett oppgir feil opplysninger oftere enn kvinner, så er det grunn til å ta med denne variabelen i modellen. Kjønn kan slik sett være en god forklaringsvariabel.

Samtidig er det viktig å passe på at modellen ikke forskjellsbehandler grupper eller personer på en urimelig måte. Dette er også viktig ettersom produksjonsdata for ett år normalt blir treningsdata for året etter. Det vi snakker om her er et viktig etisk spørsmål. Det er ofte nødvendig å inkludere variabler knyttet til gruppetilhørighet i ML-modeller. Dette gjelder også variabler som må karakteriseres som personopplysninger (f.eks. kjønn). Samtidig kan det være en fare for at en ML-modell overdriver (eller underdriver) effekter av slike variabler. Noe som kan gi oss modeller som er diskriminerende.

Punkt 1 vil således normalt være å teste modellen for indikasjoner på urimelig forskjellsbehandling. Igjen, for å ta kjønn som eksempel: Som sagt over øker sjansen for å bli plukket ut til kontroll om du er mann, rett og slett fordi de er noen flere menn enn kvinner som reelt sett oppgir feil bostatus.

For klassifikasjonsmodeller betyr likebehandling at sannsynlighet for riktig eller feil klassifisering er (omtrent) den samme uavhengig av forhold nevnt ovenfor. En indikator for dette er at «false positive rate» (FPR) skal være den samme for ulike kategorier innen samme gruppe. FPR er altså andelen som blir plukket ut som sannsynlige «misligholdere», men som reelt sett ikke er det. FPR angir altså andelen personer som modellen feilklassifiserer som «misligholdere». Det skal altså ikke være flere menn som er feilklassifisert enn kvinner.

Det er ingenting i dokumentasjonen vi har fått som tilsier at Lånekassen har vurdert likebehandling i utviklingen av ML-modellen. Det kunne nok med fordel vært gjort, særlig ettersom det er mulig å belyse ved enkle analyser. Vi har derfor gjort noen slike analyser, som viser at modellen overdriver relativt kraftig betydningen av nettopp kjønn. Det ser derfor ut til at menn i større grad blir plukket ut som «sannsynlige misligholdere» enn det det er grunnlag for i datamaterialet. Den reelle andelen som har oppgitt feil bostatus[36] er 4 % kvinner og 7 % menn. Det er derfor gode grunner til at menn vil ha noe høyere sannsynlighet for å bli plukket ut til kontroll enn kvinner. Imidlertid er FRP-verdien[37] 2,27 ganger så stor for menn sammenlignet med kvinner. Det betyr at menn som faktisk kan

[36] Basert på treningsdata

[37] "False positive rate", se appendiks A2

dokumentere borteboerstatus likevel er plukket ut mer enn dobbelt så ofte til kontroll enn kvinner som kan dokumentere borteboerstatus. En viss kjønnsforskjell gir altså store utslag mht. prediksjonsresultat.

Appendiks

A1 Begrepsforklaringer

- «Black box» modell: Med «black box», eller «sort boks», menes modeller der man bare kan se inndata og resultat og ikke har innsyn i *hvordan* inndata gir et gitt resultat. Klassiske eksempler er «ensemble models» som Catboost hører til, og nevrale nettverk.
- «White box» modell: Dette er da modeller hvor man har innsyn i hvordan inndata gir et gitt resultat, normalt i form av tolkbare koeffisienter. Klassiske eksempler er beslutningstrær og logistisk regresjon.
- «Misligholdere»: Kunder som ikke leverer dokumentasjon om borteboer status

A2 Equality and Fairness measures in classification models

Dette avsnittet er på engelsk fordi det er et uttrekk fra en utkast av en white paper om «audit of machine learning algorithms»

The performance of classification models is usually evaluated based on the confusion matrix and derived metrics. In a binary classification with two possible outcomes, one class is defined as the positive outcome, e.g. in an image classification model testing for the occurrence of skin cancer.

The confusion matrix then shows how many of actual positive/negative results the model has predicted to be positive/negative:

Tabell 5 Confusion matrix[38]

	Actual – Positive	Actual – Negative
Predicted – Positive	**True Positive (TP)** PPV = $\frac{TP}{TP+FP}$ TPR = $\frac{TP}{TP+FN}$	**False Positive (FP)** FDR = $\frac{FP}{TP+FP}$ FPR = $\frac{FP}{FP+TN}$
Predicted – Negative	**False Negative (FN)** FOR = $\frac{FN}{TN+FN}$ FNR = $\frac{FN}{TP+FN}$	**True Negative (TN)** NPV = $\frac{TN}{TN+FN}$ TNR = $\frac{TN}{TN+FP}$

For example, the true positives (TP) in the "bokontroll" case study are students where the model correctly predicts that they are not living apart from their parents (or they cannot document it), and the true positive rate (TPR) is the fraction of such correctly positive prediction in all actual positive cases. The TPR thus describes the fraction of customers that indeed do not provide valid documentation found by the model, while the false positive rate (FPR) describes the fraction of customers that do provide valid documentation incorrectly predicted by the model as without documentation.

[38] Copied from: Verma, Sahil and Rubin, Julia: «Fairness Definitions Explained», 2018 ACM/IEEE International Workshop on Software Fairness, http://fairware.cs.umass.edu/papers/Verma.pdf

Where TPR and FPR evaluate the model prediction for a given actual outcome, another important aspect is the evaluation of the actual outcome given a certain model prediction: The positive predicted value (PPV), better known as precision, describes the fraction of positive predictions that are correct, while the negative predictive value (NPV) describes the fraction of correct negative predictions.

In the bokontroll case, these two different view can be summarized to:

- How likely is a student to be controlled (i.e. predicted positive) when he/she is actually living apart from the parents or not

VS

- How likely is it that a student predicted to be living with his/her parents (or not) is actually living with his/her parents (or not)

A common approach to fairness is to demand different groups of people to be treated in the same way according to either of these views, with the first view leading to measures categorized as disparate (mis)treatment, procedural fairness or equality of opportunity, and the second classified as disparate impact, distributive justice or minimal inequality of outcome.

A third approach to group based fairness is to demand that the fraction of predicted positives (predicted prevalence) is independent of any group affiliation, irrespective of possible differences in the actual fraction of positives (prevalence) in these groups.

In order to calculate measures for (in)equality in treatment of different groups of people by a model, and assess the extent of a possible disparity, the following metrics are helpful:

- Prevalence : fraction actual positive
- Predicted prevalence : fraction predicted positive
- Precision, aka positive predicted value (PPV) : fraction TP in all predicted positive
- False positive rate (FPR) : fraction FP in all real negative
- True positive rate (TPR), aka recall aka sensitivity : fraction of TP in all real positive
- Negative predictive value (NPV) : fraction TN in all predicted negative

Based on these, the following group fairness metrics can be calculated:

- **Statistical parity** (aka demographic parity): The predicted prevalence is the same between groups - Probability for pos/neg prediction is equal
- **Equalized odds** (aka disparate mistreatment): Same TPR and same FPR - The probability for a pos prediction given a pos/neg truth is equal
- **Sufficiency** (aka predictive rate parity): Same PPV and same NPV - The probability of a real pos/neg given a predicted pos/neg is equal

It is important for the auditor to understand that in the common case of different prevalence in different groups, no imperfect model can satisfy any two of these three measures at the same time. It is therefore important to take the prevalence (often called base rates) into account when assessing

the seriousness of violations of these fairness principles, as well as the magnitude of the difference and (obviously) the practical implications in the specific ML application.

The group fairness measures discussed here have the advantages that they can easily be calculated and related to discrimination laws. Other fairness concepts the auditor should be aware of include:

- Fairness through unawareness: The naïve idea that an algorithm cannot be discriminating wrt. a certain personal attribute if that attribute is not given to the model can be regarded as too simplistic for ML applications used in public services, as it neglects correlations
- Individual fairness: Focusing on individual cases, this approach demands similar cases to be treated in a similar way by the model. It is much more challenging to calculate a metric for individual fairness compared to group fairness, as the notion of "similarity" needs to be defined in appropriate distance measures in both the feature space (model input) and the prediction space (model output).
- Counterfactual fairness: This approach tries to determine the influence of a personal attribute on the prediction by changing that attribute plus all correlated attributes. It can thus help to analyze the reasons and mechanisms behind a possible bias rather than just observe and quantify it. It is, however, unclear how to implement this approach, as one needs to make sure all relevant variables and correlations are correctly taken into account, and define a causal graph that relates them.

A3 Teknisk vedlegg med dokumentasjon av funn

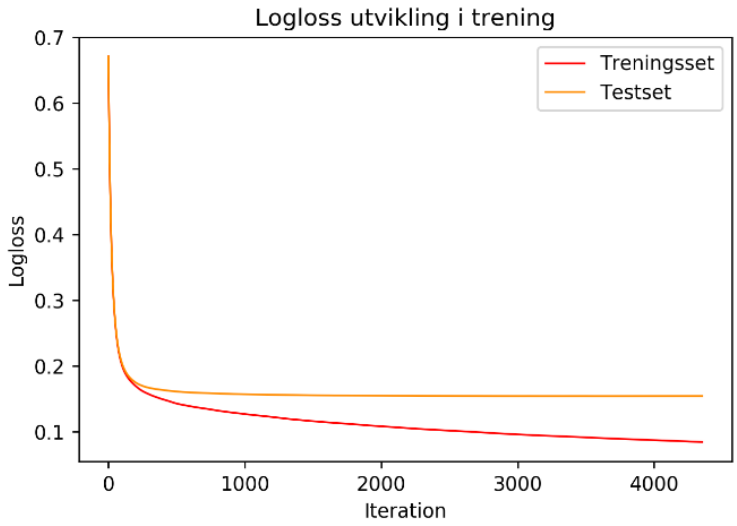
A3.1 Overtrening

Catboosts innebygde «overfitting detector» er brukt på Lånekassens modell for å stoppe iterasjonene hvis visse overtreningskriterier er oppfylt, men ingen test for overtrening, som eksempelvis sammenligning av treffsikkerhet på trenings- og testdata, er dokumentert. En test med kryssvalidering for å verifisere resultatene er gjort, men denne testen bruker 800 iterasjoner som maksimalt antall, som er mye mindre enn maksimalt antall i trening av modellen[39] (og overtrening forsterkes med flere iterasjoner).

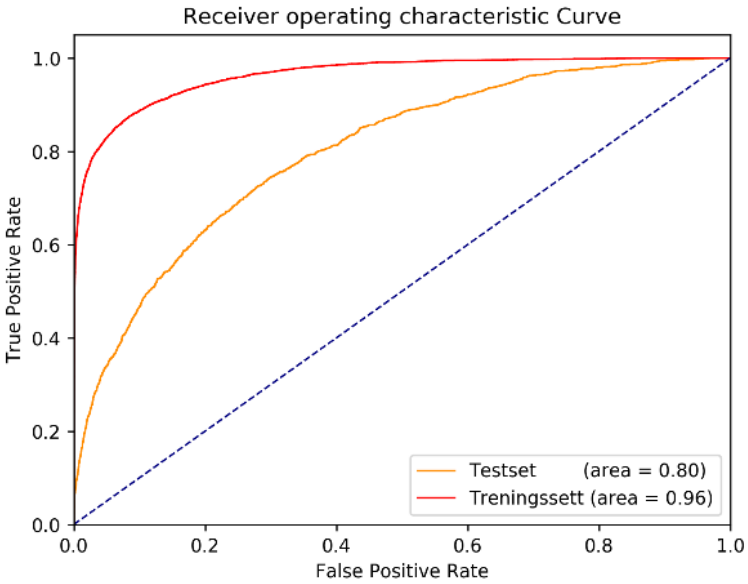
En modellparameter som kan justeres for å mindre overtrening, såkalt L2 regularisering, er testet i gridsearch men evaluering her skjer bare mht. AUC på testdata. I tillegg er parameteret ikke brukt i den endelige modellen: dette er en av parametrene som får en verdi i konfigurasjonsfilen, men som ikke er gitt til modellen for trening senere i koden. Så det er uklart om denne parameter er tenkt brukt eller ikke.

I beste rekjøring av modellen, dvs. trening av en Catboost-modell med samme parametre, på samme datasett og med samme kode[40], får vi en utvikling av logloss funksjon og AUC kurven for trenings- og testdata som vises i figur 2:

[39] Hvor mange iterasjoner som ble brukt i modellen er ikke dokumentert, bare det maksimale antallet på 5000. I vår reproduksjon stopper Catboost sin «overfitting detector» treningen etter ca. 4000 iterasjoner
[40] Modifisert bare sant at koden kjører, og skriver ut litt mer informasjon



Figur 2 Logloss utvikling i trening av vår beste reproduksjon av Lånekassens modell, på treningsdata og testdata



Figur 3 ROC kurve og AUC av vår beste reproduksjon av Lånekassens modell, på treningsdata og testdata

Logloss utvikling i figur 2 visualiserer data som Catboost produserer automatisk i form av egne filer.[41] Vi fikk ikke disse filene fra Lånekassen, men vi går ut fra at Lånekassen har lignende filer og kan se hvordan forskjellen mellom trenings- og testdata øker med flere iterasjoner.

ROC kurven i figur 3 har samme areal (0.80) i vår rekjøring som i Lånekassens dokumentasjon. Vi viser i tillegg samme kurve for treningsdata, som har et areal på 0.96. Forskjell mellom trenings- og test-performance er dermed ganske stor, og viser at modellen ikke kan forventes å generaliserer godt til nye data. Som rapporten viser er dette også realiteten.

[41] Filene får automatisk følgende navn: «training files/learn_error.tsv» og «training files/test_error.tsv»

A3.2 Likebehandling og fairness

Se appendiks A2 (på engelsk) for utførlige definisjoner av begreper. De fleste måleparametre (metrics) ble beregnet med Pythonpakken *aequitas*.

Vi vil også presisere at betraktningene her er ment som et eksempel på hva kunne vært gjort, men at det finnes mange måter å vurdere dette på.

Likebehandling

Likebehandling i ML-kontekst betyr at en modell behandler ulike grupper eller personer på samme måte. Gruppebasert likebehandling er enklest å teste, og skulle være tilstrekkelig for å bekrefte at modellen oppfyller generelle krav til allmenn likebehandling.

For klassifikasjonsmodeller betyr likebehandling at sannsynlighet for riktig eller feil klassifisering er (omtrent) den samme uavhengig av forhold nevnt over. For eksempel, en kunde som ikke har gitt falske opplysninger skal ikke ha høyere sannsynlighet å bli plukket ut til kontroll pga. kundens kjønn eller sosiale faktorer. Sagt på en annen måte: False positive rate burde være samme for alle grupperinger, og relative rater sammenlignet med en referansegruppe burde således være 1.

Den største gruppen er systematisk brukt som referansegruppe. Alle *metrics* for ulikhet (*disparity*) er pr. definisjon 1 for referansegruppen. Standard aksepten for ulikhet har vi satt til 20 %, dvs. ulikhetsverdier mellom 0.8 og 1.25[42] er tolket som «fair» (grønn i tabell), større ulikhet som «unfair» (rød).

I tabeller nedenfor er verdier under «group» absolutte, mens verdier under «disparity» er relative til referansegruppe og avvik fra 1 viser ulikhet til referansegruppe. Metrics for ulikhet baserer seg på «confusion matrix»: se appendiks A2 for definisjoner.

Vi viser således her to eksempler, hhv. for kjønn og alder.[43]

Tabell 6 "Disparity" metrics mht. kjønn

group				disparity				
attribute	**value**	**size**	**prevalence**	**PPREV**	**precision**	**FPR**	**TPR**	**NPV**
kunde_kjonn	M	4136	0.07	2.23	1.14	2.27	1.35	0.99
kunde_kjonn	K	5847	0.04	1	1	1	1	1

[42] $(1-\tau) \leq$ disparity $\leq 1/(1-\tau)$ er en standard, med $\tau=0.2$

[43] Vi har også sett på forskjell mht. økonomiske forhold (inntekt og formue), sosiale forhold (gini-difference) og migrasjonsbakgrunn (med post- og kommunenummer som mulige proxy-variabler).

Tabell 7 "Disparity" metrikker mht. alder

group				disparity				
attribute	**value**	**size**	**prevalence**	**PPREV**	**precision**	**FPR**	**TPR**	**NPV**
kunde_alder	<=20	1284	0.06	1.06	1.34	1.04	1.1	1
kunde_alder	21-22	3155	0.04	1	1	1	1	1
kunde_alder	23-24	2964	0.05	1.02	1.13	1.01	1.04	1
kunde_alder	>24	2580	0.06	1.31	1.15	1.31	1.13	0.99

Den største forskjellen vi finner mellom grupper er relatert til kundens kjønn: Menn har 0.07/0.04 = 1,75 ganger høyere sannsynlighet enn kvinner for å ikke kunne dokumentere bostatus, men er predikert til å ha 2,23 ganger høyere sannsynlighet for dette.

De som faktisk er «misligholdere» har en 35 % høyere sannsynlighet for å bli funnet av modellen hvis de er menn, sammenlignet med kvinnelige «misligholdere». Mest bekymringsfullt er det at menn som faktisk IKKE er «misligholdere» har en 127 % høyere sannsynlighet å bli kontrollert enn kvinnelige ikke-«misligholdere».

Fairness

Som forklart i appendiks A2, fokuserer vi på følgende standardkriterier for *fairness*:

- Statistical parity (aka demographic parity): Sannsynlighet å bli plukket ut til kontroll er den samme uavhengig av gruppekategori: samme *PPREV* (predicted prevalence) i tabell
- Equalized odds (aka disparate mistreatment): Sannsynlighet for å bli plukket ut til kontroll med god grunn er uavhengig av gruppekategori, dvs. sannsynligheten er den samme for de som har og ikke har dokumentasjon: samme *TPR* (true positive rate) og samme *FPR* (false positive rate)
- Sufficiency (aka predictive rate parity): Hvor mange som er predikert riktig skal være uavhengig av gruppekategori: samme *PPV* (positive predicted value, aka presisjon) og samme *NPV* (negativ predictive value)

Tabell 8 Fairness kriterier mht. kjønn og alder

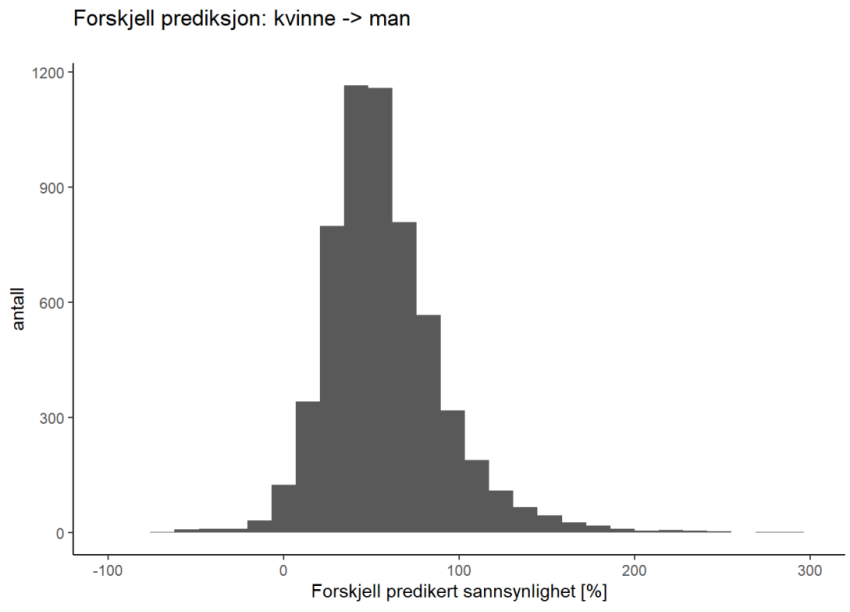
group				fairness		
attribute	**value**	**size**	**prevalence**	**Statistical Parity**	**Equalized Odds**	**Sufficiency**
kunde_kjonn	M	4136	0.07	FALSE	FALSE	TRUE
kunde_kjonn	K	5847	0.04	(-)	(-)	(-)

Vi viser bare metrics for kjønn siden den største ulikheten finnes her.

Merk at det er matematisk umulig for en modell å være *fair* på alle disse 3 definisjoner samtidig hvis basisfordeling (*prevalence*) ikke er den samme for alle gruppekategorier.

Mht. kjønn, hvor basisfordelingen er veldig forskjellig, er det derfor ikke overraskende at 2 av 3 fairness kriterier ikke er oppnådd. Lånekassen burde dog diskutere hvilke kriterier som er viktigst i konteksten av bokontroll. I tillegg er størrelse av ulikbehandling (se FPR i tabell 6) bekymringsfull.

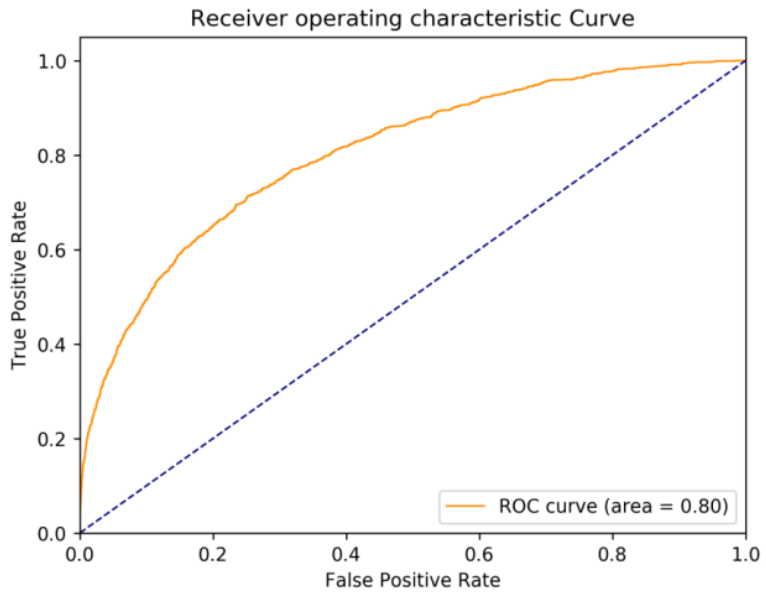
En test i tillegg er å se på endring i modellens prediksjon med endring av kjønn: Modellens prediksjon hvis kjønn endres fra kvinne til mann, og alle andre variabler holdes uendret, vises i figur 4. På periodenivå følger en predikert sannsynlighet som er gjennomsnittlig 59 % høyere for menn enn for kvinner med ellers samme variabler.



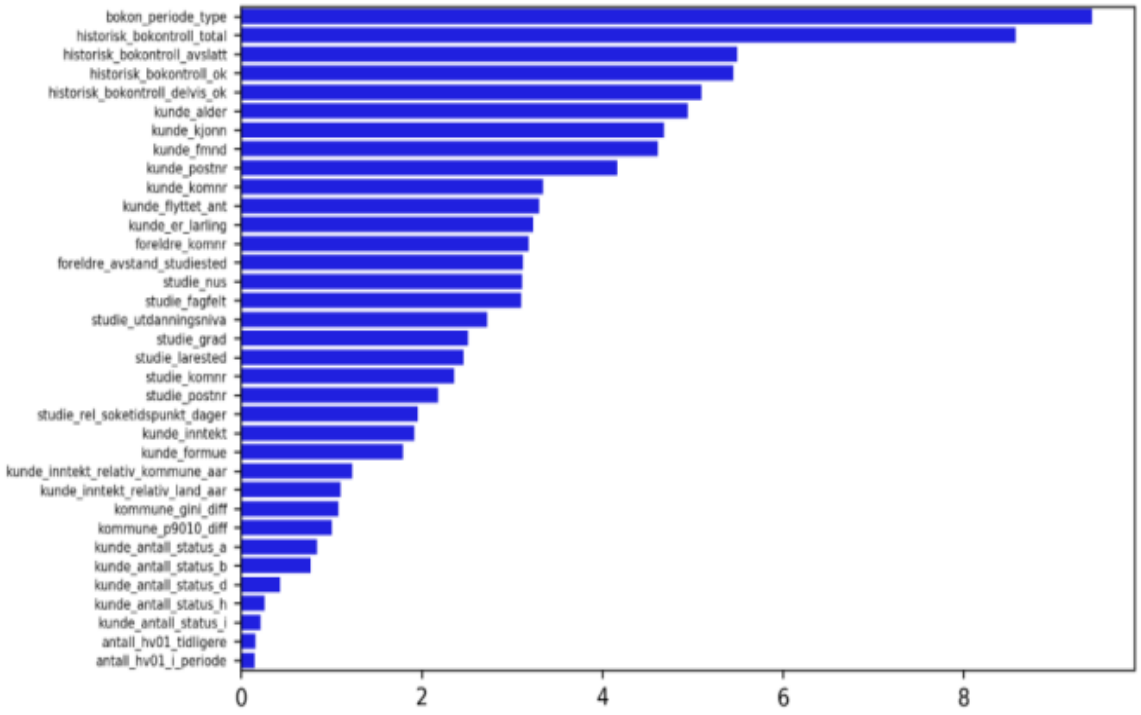
Figur 4 Prosentvis forskjell i predikert sannsynlighet hvis kjønn er endret fra kvinne til man, og alle andre variabler blir som de er

A4 Uttrekk fra Lånekassens dokumentasjon

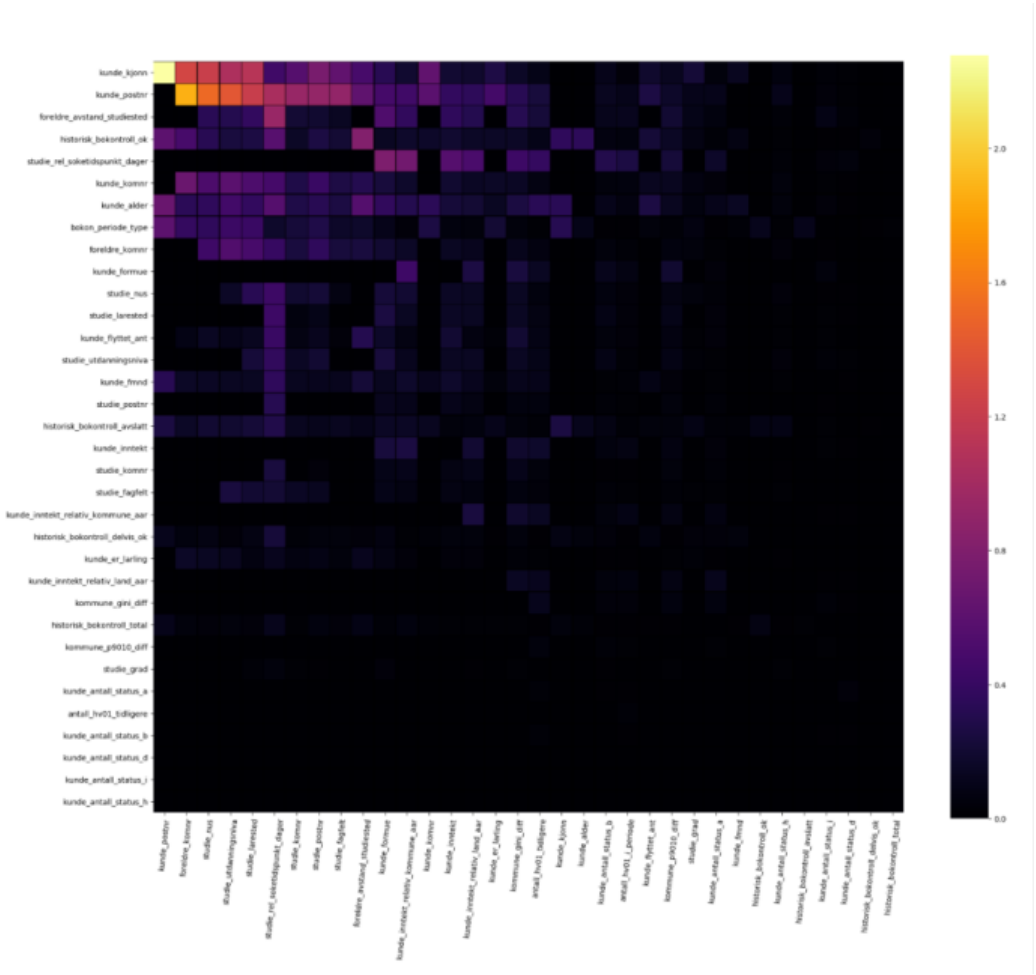
Om modellens treffsikkerhet, jf. S. 35/36 i Lånekassens Behovs- og løsningsbeskrivelse [Ref.1]:



Figur 5 "Performance": ROC kurve



Figur 6 "feature importance"



Figur 7 "feature interaction"

A5 Referanser: Dokumentasjon fra Lånekassen

[1] «30377 Bokontroll 2018 - Behovs- og Løsningsbeskrivelse», 17.12.2017, Versjon: 0.60, Ansvarlig produkteier: Robin Sande

[2] «Brukerveiledning - Kjøring av Bokontroll vha maskinlæring (Catboost)», Forfatter: Snorre Visnes

[3] «Bokontrollen for 2014 - Planlegging, gjennomføring og resultater av kontrollen», Versjon: 1.0 Forfatter: Fagavd v/ Bjørn Rossevatn, Arkivref: 201200898

[4] «Machine Learning and Lånekassen», 19.03.2019, Oslo Big Data Day, Johan Fu, https://drive.google.com/file/d/1R-pIMo5cWQvklcWksBZBnGacWTLS8_uu/view