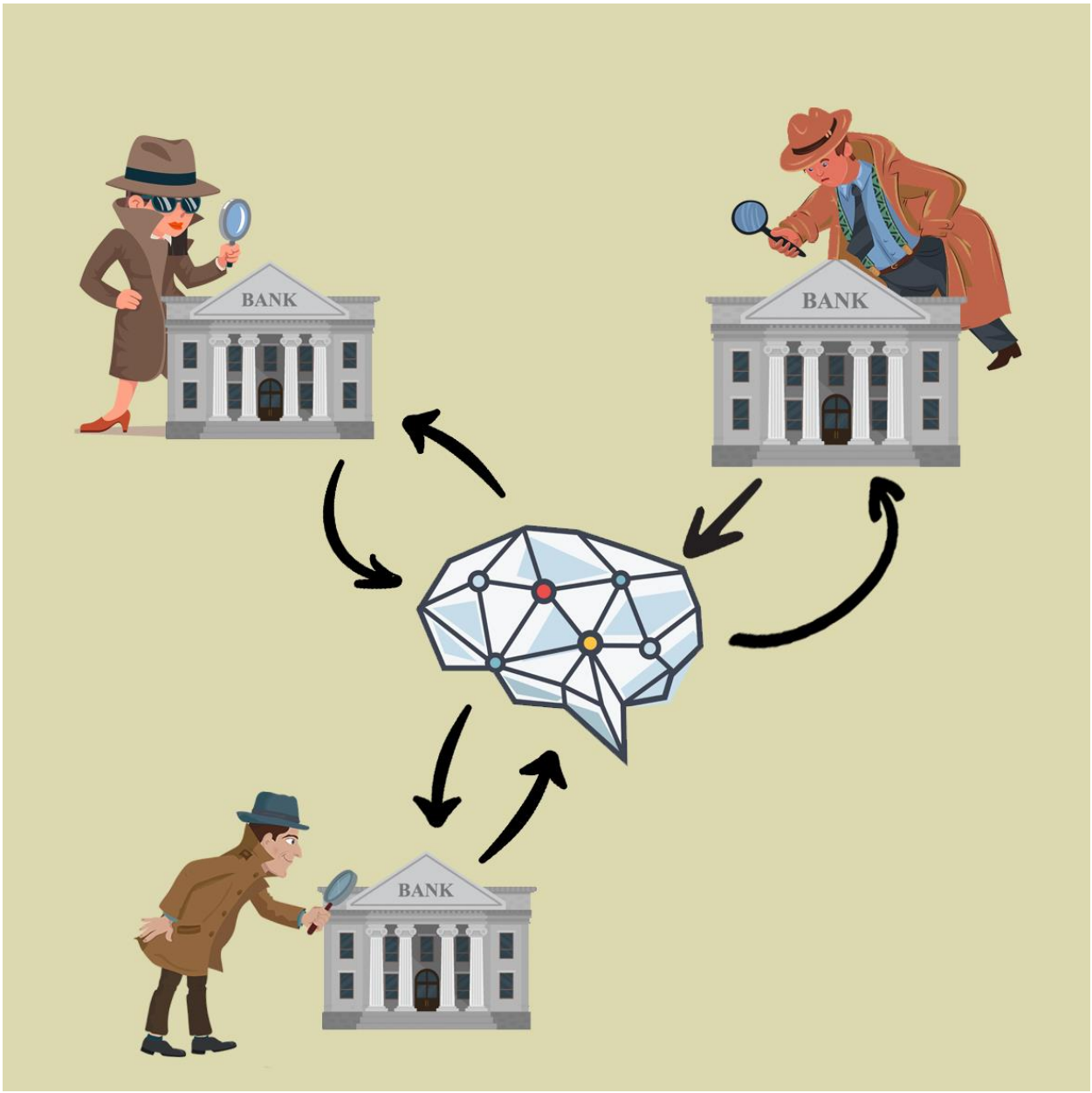

Maskinlæring uten datadeling: Bruk av føderert læring for antihvitvasking

Sluttrapport fra sandkasseprosjektet med Finterai

Oktober, 2022

Datatilsynet

Innhold

Hvordan kan du lære av data du ikke har?

Kan føderert læring være løsningen, når datadeling er vanskelig? Datatilsynets sandkasse har utforsket utfordringer og fordeler ved føderert læring, en antatt personvernvennlig metode for maskinlæring, som oppstartsvirksomheten Finterai ønsker å bruke i kampen mot hvitvasking og terrorfinansiering.

Sammendrag

Finterai er en norsk oppstartbedrift som vil gå løs på et samfunnsproblem langt større aktører har revet seg i håret over før dem; hvitvasking og terrorfinansiering. Bankene er pålagt å gjøre sitt for å forhindre det, men sliter med å gjøre det på en effektiv måte.

Kjernen i problemet er at hver enkelt bank har «for få» kriminelle transaksjoner til å kunne gi gode nok indikasjoner på hva som faktisk skiller en mistenkelig transaksjon fra mengden. Resultatet er at bankenes elektroniske overvåkningssystemer flagger altfor mange transaksjoner (falske positive), som så utløser et påfølgende tid- og kostnadskrevende, manuelt etterforskningsarbeid. Problemet kan kanskje løses ved å bygge systemer på basis av mer data enn det som foreligger i dag. Utfordringen er at banker ikke kan dele de nødvendige dataene seg imellom, siden transaksjoner inneholder personopplysninger.

Kan føderert læring løse floken?

Finterai vil løse dette datadelingsproblemet ved å anvende en relativt ny metode innen maskinlæring, nemlig «føderert læring». Føderert læring er en desentralisert metode innenfor kunstig intelligens, og blir ansett som mer personvernvennlig enn mange andre former for maskinlæring. Ved å bruke denne metoden kan banker lære av hverandre uten å faktisk dele data om kundene.

I sandkasseprosjektet har vi utforsket tre problemstillinger føderert læring reiser i tilknytning til personvernregelverket som har ledet til de tre konklusjonene under.

Konklusjoner

1. **Behandlingsansvar:** Bankene selv vil alltid ha avgjørende innflytelse på både formålet og midlene til behandlingsaktivitetene diskutert i denne rapporten, og vil derfor være *behandlingsansvarlig*. Finterai vil trolig ikke ha et behandlingsansvar for aktivitetene, med forbehold om at det må gjøres en nærmere vurdering av rettslige grunnlag, samt alle faktiske forhold, før det kan konkluderes. Finterai vil trolig være bankenes *databehandler* for kontroll av sårbarheter i modellene der Finterai skal sørge for at modellene ikke inneholder personopplysninger.

2. **Dataminimering:** Risikoprofilen til en banks kunder påvirker hvilke krav banken må oppfylle etter hvitvaskingsreglene, herunder hvor mye data de må samle inn om kundene. Det kan derfor være krevende å standardisere hvilke datakategorier alle bankene alltid må ha tilgang på for å delta i den fødererte læringen, samtidig som prinsippet om dataminimering overholdes. Vi utelukker likevel ikke at det er mulig å identifisere noen kategorier data, som det alltid kan kreves at bankene har tilgang på. For å oppfylle kravet om dataminimering bør imidlertid systemet rigges slik at bankene kan vente med å innhente personopplysninger til de vet med sikkerhet at de vil få bruk for opplysningene.

3. **Sikkerhetsutfordringer:** Bruk av føderert læring innebærer både styrker og utfordringer når det kommer til *informasjonssikkerhet* og personopplysningssikkerhet. Føderert læring reduserer behovet for deling av data. Samtidig er det en relativt ny metode. Løsningen benytter i utstrakt grad skytjenester som krever sikkerhetskompetanse, men sørger også for at deltakende aktører i stor grad kan benytte egne kapabiliteter og ressurser for å sikre sin del av løsningen. En potensiell angrepsvektor relatert til føderert læring er modellinverteringsangrep, som har som formål å rekonstruere (person)data basert på tilgang til trente modeller. *Risiko* for dette ansees som lav, men også krevende å vurdere.

Hva er sandkassa?

I sandkassa utforsker deltakere sammen med Datatilsynet personvernrelaterte spørsmål, for å bidra til at tjenesten eller produktet deres etterlever regelverket og ivaretar personvernet på en god måte.

Datatilsynet tilbyr veiledning i dialog med deltakerne, og konklusjonene fra prosjektene er ikke vedtak eller forhåndsgodkjenning. Deltakerne står fritt i valget om å følge rådene de får.

Sandkassa er en verdifull metode for å utforske problemstillinger der jussen har få praktiske eksempler å vise til, og vi håper konklusjoner og vurderinger i rapporten kan være til hjelp for andre med liknende problemstillinger.

Oktober 2022

Denne pdf-en tilsvarer den første versjonen av rapporten, slik den ble publisert på Datatilsynets sider oktober 2022. Teknologien og jussen er stadig i utvikling, så det *kan* være behov for å justere eller presisere rapportene med tiden.

Dersom denne pdf-en skiller seg fra det som står på Datatilsynets nettsider, kan du ta utgangspunkt i at det er nettsidens tekst som er gjeldende råd.

Om hvitvasking

Finterai er en oppstartvirksomhet, etablert i Oslo i 2021, som skal levere finansiell teknologi til banker og regulatoriske myndigheter. Tjenestene deres adresserer de globale utfordringene ved hvitvasking av penger og terrorfinansiering.

Hva er problemet Finterai forsøker å løse?

Hvitvasking handler overordnet om å sikre et utbytte fra straffbare handlinger. Formålet med hvitvaskingen er derfor å få utbyttet til å framstå som ervervet på lovlig måte ved å tilsløre/skjule pengenes illegale opphav før de integreres i den legale økonomien.

> **Hvitvasking i Finanstilsynets sandkasse**
>
> Antihvitvasking har også vært tema for et prosjekt i Finanstilsynets sandkasse. I deres prosjekt vurderte Finanstilsynet og Quesnay muligheter og begrensninger i hvitvaskingsloven for en teknisk løsning for utveksling av informasjon mellom rapporteringspliktige som kan effektivisere antihvitvaskingsarbeidet og bekjempelsen av terrorfinansiering.
>
> Les sluttrapporten fra prosjektet (finanstilsynet.no).

Terrorfinansiering er å motta, sende og samle inn penger med hensikt eller viten om at pengene skal finansiere en terrorhandling, brukes av en terrorgruppe eller av en person som handler på vegne av en terrorist/terrorgruppe. Terrorfinansiering kan skje både med utbytte fra kriminelle handlinger og/eller med legalt opparbeidede midler.

FN estimerer at hvitvaskede penger utgjør to til fem prosent av verdensøkonomien – omtrent det doble av størrelsen på «oljefondet». Dessverre indikerer internasjonal forskning også, at myndighetene bare klarer å hente tilbake så lite som 0,1 prosent av ulovlig økonomisk gevinst. Dette er altså et stort tap både for menneskene som utsettes for vinningskriminalitet og for samfunnet. EU estimerer at de taper opptil en billion (engelsk: en trillion) euro på hvitvaskingsrelatert skatteunndragelse årlig.

Les mer om hvitvasking på FNs nettsider (engelsk)

Les forskningsartikkelen "Anti-money laundering: The world's least effective policy experiment? Together, we can fix it" (engelsk)

Les mer om EUs estimat av hvor mye som går tapt i hvitvaskingsrelatert skatteunndragelse årlig (engelsk)

Ifølge hvitvaskingsloven er finansforetak pålagt å arbeide for å forhindre hvitvasking av penger og terrorfinansiering. I praksis betyr det for eksempel at de er ansvarlige for å ikke bli misbrukt til å skjule opprinnelsen til ulovlig utbytte. Ansvaret medfører at finansforetakene må forstå kundenes transaksjoner og vurdere risiko for hvitvasking av penger.

Mye bortkastet arbeid

Finterai mener at hovedproblemet med antihvitvasking og kontraterrorfinansiering er den overveldende mengden bortkastet etterforskningsarbeid. Banker er pålagt å bruke elektroniske overvåkningssystemer, men Finterai mener at nøyaktigheten av disse systemene er lav. Det fører til at mange «falske positive» transaksjoner blir etterforsket. Mange falske positive betyr at transaksjoner antas å være mistenkelige når de ikke egentlig er kriminelle.

Etter hvitvaskingsloven er banker pålagt å etterforske alle mistenkelige transaksjoner, og et stort antall falske positive transaksjoner betyr derfor veldig mye etterforskningsarbeid for bankene. I lys av dette tilbyr Finterai derfor et maskinlæringsverktøy for å forbedre de elektroniske overvåkingssystemene.

Utfordringen for maskinlæring til dette formålet, er at hvitvasking og terrorfinansiering utgjør en svært liten del av det totale antallet finansielle transaksjoner i de fleste banker. Det betyr at «kriminalitetssignalet» i dataene er svakt. Da må de fleste maskinlæringsmodeller ha store mengder data for at det skal fungere godt – mer enn det banker har alene.

Effekten av sterkere «kriminalitetssignal»

Hadde bankene hatt mulighet til å dele data med hverandre, ville «kriminalitetssignalet» være sterkt nok til at maskinlæring kan fungere godt. Da ville bankenes elektroniske overvåkningssystemer bli bedre, og øke sannsynligheten for at etterforskningsarbeidet kan innrettes på en mer effektiv måte:

1. Den totale etterforskningsmengden reduseres.
2. Det blir mulig å gjøre praktiske inngrep mot reelle hvitvaskingsforsøk på et tidligere tidspunkt.

Problemet med å dele data er at transaksjoner inneholder personopplysninger. Men kan dette problemet løses med føderert læring?

Hvilke tjenester tilbyr Finterai?

Finterai utvikler maskinlæringsteknologi basert på føderert læring, en tjeneste som hjelper banker med å samarbeide mot finansiell kriminalitet. Modellen trenes til å kunne identifisere mistenkelige transaksjoner blant annet basert på transaksjonshistorikk. Konseptet til Finterai er at bankene skal kunne lære fra hverandres datasett, men de opplever at dette er vanskelig på grunn av juridiske begrensninger på hvilke opplysninger bankene kan utveksle med hverandre.

For at bankene skal kunne dra nytte av hverandres datasett, men uten å utveksle personopplysninger, benytter Finterai seg av føderert læring. Finterai sin ambisjon er å gjøre det enkelt for en bank å utvikle sterke etterforskningssystemer basert på maskinlæring, i felleskap med andre banker.

Om føderert læring

Føderert læring er en metode for maskinlæring. Maskinlæring er et fagområde innenfor kunstig intelligens som innebærer utvikling av algoritmer som «lærer» ved å identifisere mønstre og sammenhenger i store datasett.

I utgangspunktet trenger maskinlæring data for å lære og løse problemer – og mer data gir som regel bedre problemløsning. For utviklere kan det imidlertid være en utfordring å få tilgang til nok data til å utvikle gode algoritmer. Spesielt gjelder dette i tilfeller hvor dataene består av personopplysninger, som det er strenge regler for om og hvordan man kan behandle.

Dersom en organisasjon ser at de trenger mer data, kan de samarbeide med andre organisasjoner. Dette gjøres som regel ved at organisasjonene laster opp sine data til en felles sentral server eller maskin, som alle de samarbeidende organisasjonene kan bruke til trening av maskinlæringsmodeller. Hvis man ikke har et behandlingsgrunnlag – det vi si et rettslig grunnlag for å behandle personopplysninger – for å dele opplysninger med andre, er dette imidlertid ikke mulig. Det er derfor et stort potensiale for kunstig intelligens som kan anvende store mengder data i sin læring, men som samtidig ikke krever deling av personopplysninger. Og det er nettopp dette som er hovedmålet ved føderert læring: å få til «big data» maskinlæring uten datadeling.

Historikk

Føderert læring ble utviklet av Google i 2016. Google brukte metoden for å trene en maskinlæringsmodell på data lokalisert på mobiltelefoner, men uten å laste opp dataen til et sentralisert nettverk. Formålet var å bygge maskinlæringsmodeller som ble oppdatert basert på data som lå på brukernes mobiltelefoner. Teknologien ble blant annet brukt i tastaturapplikasjonen Gboard for å predikere hvilke ord som tastes inn. I etterkant har teknologien blitt delt og brukt i andre sammenhenger.

Les Googles eget blogginnlegg om føderert læring (engelsk)

Les artikkelen "Federated Learning for Mobile Keyboard Prediction" på Google Research (engelsk)

De siste årene har ulike aktører forsket på føderert læring, som har generert flere type alternative oppsett for metoden. Men føderert læring er fortsatt et nytt verktøy og det er foreløpig begrenset kommersiell eller offentlig bruk som involverer store mengder data.

Hvordan foregår føderert læring?

Forskjellige modeller for føderert læring

Den vanligste arkitekturen for federert læring er såkalt «horisontal føderert læring». Det mindre brukte alternativet er «vertikal føderert læring» som er mer vanlig hvis to aktører deler et datasett. Andre arkitekturer inkluderer «federated transfer learning», «cross-silo federated learning, og «cross-device federated learning». Vertikal føderert læring er situasjonen der aktører har forskjellige kolonner/kategorier med data – i denne situasjonen trengs ikke datastandardisering.

Trening av kunstig intelligens ved bruk av føderert læring kan skje på flere forskjellige måter. Under har vi beskrevet de trinnvise prosessene i en vanlig modell for føderert læring (basert på modellen Google utviklet i 2016):

1. Deltaker mottar en maskinlæringsalgoritme.
2. Deltaker bruker det lokale datasettet til å trene maskinlæringsalgoritmen.
3. Deltakere krypterer sin lokale «læringspakke» som de sender til en ekstern sentral server. Læringspakkene inneholder ikke personopplysninger.
4. Serveren utfører en sikker aggregering av pakkene.
5. Aggregeringen av læringspakkene blir brukt til å oppdatere maskinlæringsmodellene som er lagret sentralt, med læring fra deltakerne. Maskinlæringsmodellen som er lagret sentralt er den samme som opprinnelig ble sendt ut til deltakerne for lokal trening.
6. Steg 1 til 5 repeteres inntil maskinlæringsmodellen er ferdig opplært.
7. Deltaker mottar den ferdig opplærte maskinlæringsmodellen og får nå bedre lokale prediksjoner.

At det bare er modellparameterne som utveksles, betyr at lokale data – som ofte består av personopplysninger – i teorien ikke trenger å overføres mellom deltakere eller mellom deltakere og den sentrale serveren. Den innebygde begrensningen av deling av lokale data, gjør at føderert læring er ansett som en mer personvernvennlig tilnærming til kunstig intelligens.

Om prosjektet

Et sentralt spørsmål ved bruk av føderert læring – og i dette sandkasseprosjektet – er: inneholder maskinlæringsmodellene som utveksles mellom deltakere personopplysninger fra lokale data? Svaret har avgjørende betydning regulatorisk, ettersom personvernregelverket bare gjelder hvis man behandler personopplysninger.

Når man bruker føderert læring-metode på personopplysninger, er det i prinsippet bare læringen, eller «modellparametrene», som skal utveksles mellom deltakerne. Det er likevel en hypotetisk mulighet å avlede personopplysninger hvis modellen har sårbarheter. Selv om personopplysningene ikke blir sendt eller lagret eksternt, blir vektene (modellparametrene) utvekslet. Modellparametre er vektene som representerer modellens læring. Og om modellen har lært personopplysninger, kan vektene hypotetisk sett avsløre denne informasjonen til ondsinnede deltakere, som aktivt angriper modellen.

Men hvis lokale data ikke forlater bankenes lokale datasett, hva er det da som utveksles? Svaret er modellparametre og hyperparametre.

Modellparametre og hyperparametre

Hyperparametere setter rammen for hvordan maskinlæringen skal utføres. Det vil si at den definerer hva læringen skal baseres på, samt at den bestemmer hvordan datapunkter skal henge sammen. På den andre siden inneholder modellparametrene de konkrete vektene (innholdet) som modellen skal lære av.

Til å lære opp modellparametrene brukes en «tilbakeforplantningsalgoritme» (kjent på engelsk som «backpropagation algorithm»), som identifiserer hvordan vekter bør endres for at maskinlæringens prediksjoner skal bli mer presise. Prediksjonene i læringen er det som til slutt skal resultere i identifisering av *risiko* for hvitvasking. Prosessene som skal gjøre dette mulig består av flere trinn, men hvor anvendelsen av føderert læring er sentral.

Hvorvidt en maskinlæringsprosess basert på føderert læring tillater re-identifisering av data som modellen er trent på, har sammenheng med utformingen av den konkrete modellen og treningsprosessen. Utfordringer må derfor vurderes med utgangspunkt i det konkrete valget av løsningsarkitektur og maskinlæringsmodell. Dette adresseres i kapittelet om sikkerhetsutfordringer.

Hvordan skal Finterai bruke føderert læring?

For å muliggjøre sin ambisjon, vil Finterai bruke føderert læring på en litt annen måte enn Google sin versjon av teknologien. Den største forskjellen kommer helt i begynnelsen, med et steg før modellen distribueres til deltakerne. Her skal en av deltakerne selv bestemme hva slags modell som skal trenes, hvor deltakeren selv definerer hyperparameterne til modellen. Altså er det Finterais kunder og ikke Finterai selv som definerer hvilke maskinlæringsmodeller som skal trenes føderert.

Dette fører så til en ny systemforskjell. Finterais fødererte læring er seriell heller enn parallell. Det betyr at en maskinlæringsmodell først trenes hos en deltaker, før den sendes til neste deltaker. Løsningen står i kontrast til Googles tilnærming, som sender maskinlæringsmodeller ut til deltakerne parallelt, som så oppdaterer den sentrale modellen kontinuerlig. Her er det også en annen viktig forskjell: Google får bare en modelloppdatering (gradienter) tilbake fra sine deltakere, mens Finterai får tilbake hele maskinlæringsmodellen.

Modelloppdateringer er mindre enn maskinlæringsmodellene i sin helhet, og dette sparer dermed nettverkstrafikk. Likevel er det både tekniske, sikkerhetsmessige og forretningsmessige hensyn som gjør at Finterai velger å overføre hele maskinlæringsmodeller. Finterai implementerer også «secure aggregation» på en annen måte enn Google. Forskjellen er blant annet delvis en funksjon av at virksomhetene har forskjellig «use case».

Finterai skal utføre eksplisitte tester av sikkerhetstrusler, skjevhetsproblemer og datalekkasjetrusler som kan oppstå under den fødererte læringen. Dette er en sterkere grad av personopplysningsbeskyttelse og systembeskyttelse enn det Googles opprinnelige modell legger opp til. Det er verdt å merke seg at slike problemer vil oppstå i enhver situasjon der maskinlæringsmodeller deles eller tilgjengeliggjøres – det er altså ikke trusler som er unike i kontekst av føderert læring.

Forenklet, trinnvis fremstilling av Finterais fødererte læring:

1. En deltakende bank sender forespørsel til Finterai om å bygge en maskinlæringsmodell. Deltakeren oversender sine egne hyperparametere og andre treningsinstrukser til Finterai.
2. Finterai bygger en modell basert på mottatte instrukser.
3. Finterai sender denne modellen med hyperparametere til første deltakende bank for trening på deres lokale datasett.
4. Første deltakende bank mottar modellen og hyperparametre som beskriver treningen. Denne treningen gjøres lokalt hos deltakeren på standardiserte transaksjonsdata og andre data (KYC- og tredjepartsdata).
5. Finterai får modellen og hyperparametere i retur når treningen er gjennomført lokalt hos den deltakende bank. Modellen lagres deretter i Finterais database.
6. Finterai kvalitetssikrer modellen, og kontrollerer for blant annet datalekkasjer og skjevheter.
7. Finterai sender den oppdaterte modellen og relevante hyperparametre til neste deltakende bank.
8. Deltakeren mottar modellen og hyperparameterne. Modellen trenes lokalt hos deltakeren på samme type data som i steg 4.
9. Steg 5 til 9 repeteres inntil modellen er ferdig utlært – altså at den har konvergert.
10. Finterai lagrer den ferdig trente modellen på en server. Alle deltakerne i den fødererte læringen har tilgang til modellene. Disse kan lastes ned fra serveren, og umiddelbart brukes med bankenes lokale datasett for å identifisere mistenkelige transaksjoner.

I denne modellen skal all lagring av data i forbindelse med disse prosessene (inkludert transaksjonsovervåking) skje hos bankene. Finterai skal ikke ha tilgang til bankenes lokale data med transaksjonsopplysninger for å utvikle eller drifte tjenesten.

Drøftinger mellom Datatilsynet og Finterai

Datatilsynet og Finterai har hatt fem arbeidsmøter hvor vi har diskutert teknologien Finterai planlegger å benytte, og utfordringer knyttet til personvernregelverket. Finterai var ved første arbeidsmøte i konseptstadiet av løsningen sin. Derfor har mange av diskusjonene handlet om hvordan Finterai kunne utforme løsningen sin på en måte som best ivaretar personvernet. Datatilsynet har ikke prøvd å påvirke Finterais metode, men diskusjonene har bidratt til å belyse konsekvensene av veivalgene de tar når de utformer sin løsning.

Én konkret lærdom fra arbeidsmøtene er at utviklere kan utforme føderert læring-metoden på mange forskjellige måter. De forskjellige utformingene vil påvirke personvernet i løsningen og i varierende grad ivareta viktige personvernhensyn og åpne opp for sårbarheter. Veivalg som ville medført å samle og sentralisere bankenes transaksjonsopplysninger på en sentral server, vil potensielt kunne skape en stor angrepsflate og utløse store krav til tekniske og organisatoriske tiltak.

På tidspunktet denne sluttrapporten skrives, har Finterai valgt å gå for en mer desentralisert løsning – hvilket minimerer systemets angrepsflate, ettersom forskjellige datalagringssystemer sjeldent kan angripes med samme sårbarhet. Det vil også ha konsekvenser for sikkerhetstrusler, som omtalt i kapittelet om sikkerhetsutfordringer. Informasjon om kravene og konsekvensene fra de ulike systemarkitekturene har vært svært viktige for Finterai, da denne informasjonen har hjulpet dem med å ta gode valg i en tidlig fase preget av mye usikkerhet.

Finanstilsynets involvering i prosjektet

Dette prosjektet berører forholdet mellom hvitvaskingsreglene og personvernreglene, som begge ivaretar viktige samfunnshensyn. Det er Finanstilsynet som fører *tilsyn* med at rapporteringspliktige etterlever hvitvaskingsreglene, men Finanstilsynet har ikke hatt noen formell rolle i Finterai-prosjektet i Datatilsynets sandkasse.

Hensynene disse to regelverkene skal ivareta inneholder til en viss grad motstridende prinsipper, med enkelte uavklarte grenser mellom kundetiltak og dataminimeringprinsippet. Datatilsynet har gjennom dette prosjektet erfart at det kan være krevende for oss som tilsynsmyndighet å gi tydelige anbefalinger og veiledning om godt personvern i anti-hvitvaskingsarbeidet, uten involvering også fra Finanstilsynet. Det har derfor vært naturlig å konsultere Finanstilsynet om relevante forhold knyttet til tolkningen og praktiseringen av hvitvaskingsreglene underveis i prosjektet.

Finanstilsynet har også deltatt som observatør på ett av arbeidsmøtene i sandkasse-prosjektet. Det er imidlertid viktig å presisere at denne rapporten gir uttrykk for Datatilsynets vurderinger og synspunkter. Finanstilsynet har vurdert om gjengivelser i hvitvaskingsloven er uriktige, men har ikke tatt stilling til faktumbeskrivelser og Finterais vurderinger av regelverket. Finanstilsynet har ikke vært involvert i skrivingen av rapporten.

Finterai og Finanstilsynet har parallelt med sandkasseprosjektet også hatt dialog om en rekke spørsmål knyttet til tolkning av konkrete bestemmelser i hvitvaskingsloven. Denne dialogen har primært dreid seg om spørsmål, som har hatt til hensikt å avklare om hvitvaskingsloven legger begrensninger på hvilke typer opplysninger som kan deles mellom rapporteringspliktige. Disse spørsmålene er besvart av Finanstilsynet i brev sendt direkte til Finterai.

Mål for sandkasseprosessen

Basert på Finterais bruk av kunstig intelligens, har Datatilsynet og Finterai sammen identifisert tre problemstillinger for hvordan føderert læring utfordrer personvernregelverket og de registrertes personvern i dette prosjektet. Disse problemstillingene kan også være relevante for andre utviklere og virksomheter som vil bruke føderert læring:

1. **Hvilke roller og ansvarsforhold påtar Finterai seg i anvendelsen av føderert læring-modellen?** Hvilken behandling av personopplysninger skjer ved bruk av føderert læring, herunder standardisering som ledd i forberedelsene? Hvilket ansvar utløser dette for Finterai?

2. **Hvordan kan Finterai legge til rette for dataminimering tilknyttet behandling av transaksjonsopplysninger og tredjepartsopplysninger?** Hva er konsekvensene for personvernet når Finterai setter rammer for hvilke datakategorier alle deltagerne må ha tilgang på?

3. **Kan Finterai ivareta personopplysningssikkerheten ved utveksling av maskinlæringsalgoritmer?** Hvordan kan Finterai ta stilling til eventuelle brudd og sårbarheter i løsningen? Hvordan finner man frem til et akseptabelt risikonivå?

Behandlingsgrunnlag for bruk av personopplysninger

For at det skal være lovlig å behandle personopplysninger, må det alltid være et rettslig grunnlag for behandlingen.

Personvernforordningen (artikkel 6 nr. 1 bokstav a til f) inneholder en uttømmende liste med seks rettslige grunnlag for en lovlig behandling av personopplysninger.

I dette sandkasseprosjektet har vi ikke tatt stilling til om banker vil ha et rettslig grunnlag for å behandle personopplysninger i KI-verktøyene som Finterai tilbyr. Dette gjelder både for bruk av KI-verktøyene som ledd i bankenes antihvitvaskingsarbeid, og eventuell bruk av personopplysninger til trening av algoritmene. Vi har heller ikke tatt stilling til om Finterai har rettslig grunnlag for behandling av personopplysninger, dersom det skulle bli aktuelt.

Drøftelsene i dette sandkasseprosjektet forutsetter at de behandlingsansvarlige, enten det er Finterai selv eller bankene, finner rettslig grunnlag for behandling av personopplysninger ved bruk og videreutvikling av tjenesten. I motsatt fall vil ikke tjenesten lovlig kunne tas i bruk.

Når det gjelder bankenes adgang til å behandle personopplysninger for formål knyttet til å avdekke mistenkelige transaksjoner, antar vi at rekkevidden av bankenes forpliktelser og eventuelle handlingsrom etter hvitvaskingsreglene vil være det naturlige utgangspunktet ved vurderingen av rettslig grunnlag (artikkel 6 nr. 1 bokstav c). I noen av drøftelsene har vi derfor vist særskilt til dette regelverket, uten at vi med det har tatt stilling til om bankenes bruk av tjenesten kan hjemles i hvitvaskingsreglene.

Dersom behandlingen ikke kan hjemles i hvitvaskingsreglene, må bankene selv identifisere et annet rettslig grunnlag for behandlingen. Legitime interesser (artikkel 6 nr. 1 bokstav f) vil trolig være det mest aktuelle alternativet, uten at vi har tatt stilling til dette her.

Hvilke personopplysninger behandles?

SWIFT-meldinger

SWIFT-meldinger består av transaksjonsdata. SWIFT er et internasjonalt betalingsnettverk for overføring av penger mellom banker som ikke er i samme land. De kan inneholder personopplysninger hvis et individ er avsender eller mottaker i transaksjonen.

Kjenn-din-kunde-data (forkortet KYC av engelsk «know your customer»)

Finansforetakene er pålagt å innhente opplysninger om kundene sine (herunder hvem de er), formålet med kundeforholdet og dets tilsiktede art, hvilke tjenester og produkter de benytter hos den rapporteringspliktige og kilden til midler mv. Dette omtales som «kjenn- din-kunde-prinsippet», og brukes til å klassifisere kunder innenfor ulike risikogrupper og til å overvåke at transaksjoner foretas i samsvar med innhentede opplysninger.
Ikke alle kategoriene KYC-data inneholder nødvendigvis personopplysninger, men noen kategorier gjør det. KYC-data innhentes både fra kunden selv, og fra offentlig tilgjengelige nettsider og eksterne leverandører som tilbyr denne typen data som en betalt tjeneste. Data som innhentes fra offentlig tilgjengelige nettsider og eksterne leverandører omtales som tredjepartsdata.

Tredjepartsdata

Tredjepartsdata brukes for å tilføye ny informasjon eller verifisere informasjon gitt av kunden selv, f.eks. opplysninger om noen er en politisk eksponert person (PEP), om de står oppført på sanksjonslister, om det er negative medieoppslag eller opplysninger fra andre offentlige kilder knyttet til kriminalitet, søksmål og lignende. Tredjepartsdata kan inkludere datasett som er "sydd sammen" fra ulike kilder. Ofte samles tredjepartsdata gjennom ulike plattformer og nettsteder, som deretter aggregeres av en dataleverandør.
Tredjepartsdata inneholder ikke alltid personopplysninger.

Vurderingene i denne rapporten gjelder kun for behandling av data som er å anse som personopplysninger.

Roller: Hvilket ansvar har Finterai?

Et sentralt spørsmål i sandkasseprosjektet var hvordan ansvarsfordelingen mellom Finterai og deres kunder skulle være.

Bakgrunnen for spørsmålet var todelt. For det første krever føderert læring at alle involverte har de samme datapunktene i det samme formatet, og Finterai ønsket å drøfte hvor stort handlingsrom de har for å legge til rette for en slik standardisering uten å ta på seg behandlingsansvar.

Behandlingsansvar

Personvernforordningen benytter begrepene behandlingsansvarlig i artikkel 4 nr. 7, databehandler i artikkel 4 nr. 8 og felles behandlingsansvarlige i artikkel 26 for å plassere ansvaret for å følge reglene. Av ansvarlighetsprinsippet går det frem at hovedansvaret for å sikre at behandlingen av personopplysninger er i tråd med personvernforordningen ligger hos den behandlingsansvarlige.

En behandlingsansvarlig er den som utøver avgjørende innflytelse på formålene med (dvs. hvorfor) og midlene for (dvs. hvordan) behandlingen av personopplysninger, mens en databehandler behandler personopplysninger på vegne av den behandlingsansvarlige. Felles behandlingsansvar inntrer hvor partene i fellesskap fastsetter formålene med og midlene for behandlingen av personopplysninger.

Les mer om hva en behandlingsansvarlig er

For det andre er det en viss mulighet for at modellene inneholder personopplysninger. Finterai ønsket derfor å drøfte hvilke konsekvenser dette kan få for dem når de skal kvalitetssikre modellene.

I sandkassen identifiserte vi i fellesskap ulike behandlinger av personopplysninger som skjer i, eller forutsettes av, løsningen til Finterai. Vi valgte ut tre forskjellige behandlingsaktiviteter som vi skulle se nærmere på:

1. Standardisering av transaksjonsdata
2. Innhenting av tredjepartsdata
3. Kontroll av sårbarheter i modellene

Vi valgte å se nærmere på disse behandlingsaktivitetene fordi de er sentrale behandlinger i, eller er forutsatt av, Finterais løsning, og fordi det er behandlinger som også er relevante for sammenlignbare virksomheter eller pliktsubjekter under hvitvaskingsloven.

Datatilsynet har ikke konkludert om hvilken rolle henholdsvis Finterai og bankene har i tilknytning til de tre utvalgte behandlingsaktivitetene. Dette er fordi vi ikke har tatt stilling til hva som eventuelt vil være rettslig grunnlag for Finterais eller bankenes *behandling av personopplysninger*. Diskusjonene i sandkassen har derfor i hovedsak handlet om hva som er relevante momenter i vurderingen av roller på bakgrunn av konkrete behandlingsaktiviteter, og hvilken retning de ulike momentene trekker i. Vi understreker at Datatilsynets vurderinger av roller kun er veiledende. Finterai og bankene må selv ta stilling til sin egen rolle basert på alle faktiske forhold.

1. Standardisering av transaksjonsdata

For at bankenes lokale data skal være kompatible med Finterais løsning for føderert læring, må dataene standardiseres og struktureres før de kan tas i bruk. Finterai tilbyr bankene programvare for standardisering som bruker kunstig intelligensdrevet, naturlig språkprosessering. Hensikten med denne programvaren er å standardisere data, herunder SWIFT-meldinger, slik at de er i samme format som tilsvarende data i andre banker, som også vil være med på den fødererte læringen.

I sandkassen har vi diskutert roller i tilknytning til standardisering av SWIFT-meldinger. Momentene i vurderingen er imidlertid relevante også for vurdering av roller ved standardisering av andre kategorier data, som for eksempel KYC-opplysninger og tredjepartsdata.

Det mest aktuelle formålet med å standardisere data i denne sammenhengen, vil være å tilpasse formatet på dataene, slik at de kan anvendes i føderert læring på tvers av bankene som deltar i samarbeidet. Det er den enkelte bank som selv bestemmer om de vil benytte Finterais tjeneste for føderert læring. Videre er det også bankene selv som bestemmer hvilken metode de vil benytte for å konvertere dataene til formatet som kreves for å kunne delta. Det er altså ikke et krav at de må benytte akkurat Finterais programvare for standardisering.

Videre skal programvaren installeres og kjøres i bankens eget IT-miljø, uten at Finterai eller de andre involverte bankene har tilgang på dataene som standardiseres. Trening av algoritmene i programvaren skal også skje internt hos den enkelte bank, og det er banken selv som har ansvar for dette. Resultatet av treningen skal ikke deles med Finterai eller de andre bankene som benytter samme programvare.

Momentene som er trukket frem ovenfor tilsier at den enkelte bank selv har avgjørende innflytelse både på beslutningen om å ta i bruk føderert læring som ledd i deres anti-hvitvaskingsarbeid, og hvilke midler de skal bruke for å oppnå formålet.

I diskusjonene i sandkassen har vi ikke identifisert noen formål knyttet til bankenes bruk av programvaren for standardisering, der Finterai har avgjørende innflytelse. Finterai skal som nevnt heller ikke få tilgang til personopplysningene som behandles av bankene, eller resultatet av behandlingen (standardiserte opplysninger eller læringen fra kunstig intelligens-drevet naturlig språkprosessering).

På bakgrunn av dette vurderer Datatilsynet det som lite sannsynlig at Finterai har avgjørende innflytelse på formålet eller midler til behandlingen. Finterai vil i så fall ikke få status som behandlingsansvarlig. Dersom Finterai ikke behandler personopplysninger på vegne av bankene, vil de heller ikke få status som databehandler.

2. Innhenting av tredjepartsdata

Det er ikke bare dataenes format som må være likt på tvers av aktører som vil bruke føderert læring, men også datakategoriene som benyttes. Dette er for at bankene skal ha tilgang på samme type data, og at inngående data skal kunne tolkes av modellen.

For å sikre at alle bankene som benytter tjenesten har tilgang på de samme data-kategoriene, må Finterai sette rammer for hvilke typer data som kan benyttes i systemet for føderert læring. I tillegg til data fra SWIFT-meldinger og KYC-data, som innhentes fra kundene og eventuelt transaksjonsmotpartene, krever Finterai at bankene innhenter det de kaller tredjepartsdata. Disse vil ofte inneholde personopplysninger.

I sandkassen har vi diskutert hvilken konsekvens det kan få for Finterais rolle at de definerer datakategoriene som deltagerne i den fødererte læringen må ha tilgang på. Vi har lagt til grunn at bankene allerede er i besittelse av SWIFT-meldingene og KYC-data. Diskusjonene har derfor primært knyttet seg til innhenting av tredjepartsdata.

Formålet med standardiseringen er å sikre at de deltagende bankene har tilgang på samme type data. Det er imidlertid et annet formål som er avgjørende for valget av datakategoriene, nemlig at bankene skal kunne bygge og trene modeller som er godt egnet for å avdekke mistenkelige transaksjoner.

Det er bankene som er ansvarlig for å overholde hvitvaskingsreglene. Finterai har ikke noe selvstendig ansvar etter dette regelverket. Som drøftet i avsnittene om standardisering av transaksjonsdata, er det bankene selv som har avgjørende innflytelse på om de som ledd i sitt anti-hvitvaskingsarbeid vil ta i bruk Finterais løsning for føderert læring. Det er ikke et krav at bankene må benytte Finterais programvare for innhenting av tredjepartsdata.

Den enkelte bank bestemmer selv om de vil ta i bruk Finterais tjeneste og dermed hvilke kategorier tredjepartsdata de følgelig må innhente, samt hvilken metode de vil benytte for å innhente dataene. Dersom den enkelte bank er uenig med Finterai om hvilke datakategorier som er nødvendig å innhente til formålet, kan banken velge bort Finterais løsning som ledd i deres anti-hvitvaskingsarbeid. Med andre ord kan det argumenteres for at den enkelte bank har avgjørende innflytelse på hvilke midler som skal brukes for å oppnå formålet.

Datakategorier kan sies å være essensielle midler (hvilke og hvem sine personopplysninger som skal innhentes), eller 'kjernen' i hvordan formålet skal oppnås. Videre skal ikke dataene deles med Finterai. Hvem som har avgjørende innflytelse på hvilke kategorier tredjepartsopplysninger som skal innhentes (midler) for å kunne trene modellene som er godt egnet for å avdekke mistenkelige transaksjoner (formål), er et viktig moment i vurderingen av roller.

Momentene som er trukket frem ovenfor, tilsier at det er den enkelte bank selv som har avgjørende innflytelse på formål og midler knyttet til innhenting av tredjepartsdata. Vi har imidlertid også sett på eventuelle egeninteresser som Finterai kan ha knyttet til beslutningen om hvilke data kategorier som skal benyttes i løsningen.

I vurderingen av roller ved standardisering av transaksjonsdata, la vi vekt på at Finterai hverken skal ha tilgang på personopplysningene som behandles eller resultatet av behandlingen. Når det gjelder innhenting av tredjepartsdata, skal Finterai heller ikke ha tilgang på denne typen personopplysninger men de vil få tilgang på modellene som er utviklet ved hjelp av personopplysningene.

Disse modellene skal også være tilgjengelig for alle bankene som deltar i den fødererte læringen. Man kan argumentere for at jo bedre disse modellene er, jo mer attraktiv vil Finterai sin tjeneste trolig være. Bankenes tilgang på modeller som er gode og effektive i anti-hvitvaskingsarbeid, kan derfor tenkes å fremme salget av Finterais tjeneste, og de vil derfor få en kommersiell fordel. Formålet med utvelgelsen av hvilke data-kategorier som skal inkluderes i tjenesten, er imidlertid at bankene skal kunne bygge effektive KI-modeller for å oppfylle sine forpliktelser etter hvitvaskingsreglene.

En ren kommersiell fordel som Finterai kunne tenkes å få her, er ifølge Personvernrådets retningslinjer ikke i seg selv nok til å kvalifisere som et formål til en behandling. Dette trekker derfor i retning av at Finterai ikke har avgjørende innflytelse på formålene nevnt ovenfor, og behandlingsansvar vil trolig ikke utløses for Finterai.

Les Personvernrådets retningslinjer om rollene som behandlingsansvarlig og databehandler (07/2020)

Det er også viktig å vurdere om man har et rettslig grunnlag for å innhente og bruke tredjepartsdata på denne måten før dataene innhentes og brukes. Selv om rettslig grunnlag ikke er tatt opp som tema i denne rapporten, vil vi likevel minne

leseren om dette. Dataminimeringsprinsippet må også vurderes for innhenting og bruk av tredjepartsopplysninger, noe som blir drøftet nedenfor i kapitlet om dataminimering.

3. Kontroll av sårbarheter i modellene

Den siste behandlingsaktiviteten vi diskuterte, er knyttet til kontroller som skal utføres av Finterai i forbindelse med gjennomføring av den fødererte læringen.

Modellene som utveksles mellom bankene og Finterai, skal i utgangspunktet ikke inneholde personopplysninger. Det er likevel en viss *risiko* for at man kan hente ut personopplysningene som en modell er trent på, dersom modellen har sårbarheter. Dette er omtalt som datalekkasje, og innebærer at det er mulig å reidentifisere enkeltindivider. Slike sårbarheter kan oppstå dersom det har skjedd en feil under treningen.

Finterai planlegger å gjennomføre ulike typer kontroller på modellene som skal trenes ved bruk av føderert læring. Én kontroll som alle modellene må gjennom etter treningen og før de deles videre til andre deltakere, skal avdekke eventuell sannsynlighet for datalekkasje fra modellen. Dersom kontrollen hos Finterai avdekker sårbarheter som kan føre til datalekkasje, vil behandling av den aktuelle modellen kunne være å anse som behandling av personopplysninger.

Det er grunn til å anta at slike feil vil oppstå fra tid til annen, og et spørsmål som dukket opp i sandkassen var hvilken rolle Finterai får dersom de i praksis ender opp med å behandle personopplysninger på grunn av feil fra bankene.

Formålet med sårbarhetskontrollene er å forhindre at modeller som inneholder personopplysninger skal slippe inn i systemet for føderert læring. Man kunne tenke seg at den enkelte bank selv gjorde denne kontrollen før de sendte modellen til Finterai. Det er imidlertid gode grunner som taler for at denne kontrollen bør ligge hos Finterai, blant annet for å unngå at kvaliteten på kontrollen er avhengig av kompetansen i den enkelte bank.

Finterai har opplyst at eventuelle modeller med sårbarheter blir slettet, og beskjed blir sendt til den aktuelle banken som modellen kom fra. Ved å kontrollere modellene for svakheter, og luke ut modeller med feil, bistår Finterai bankene med å forhindre at personopplysninger kommer på avveie.

På bakgrunn av ovennevnte anser vi det lite sannsynlig at Finterai blir behandlingsansvarlig for eventuelle personopplysninger som avdekkes i kontrollen de gjennomfører. Kontrollen må trolig sies å skje på vegne av banken som modellen sist ble trent hos, og Finterai fremstår som databehandler for bankene i den grad personopplysninger behandles som del av kontrollen. Det vil være naturlig å innta retningslinjer i databehandleravtalen om hvilke tiltak Finterai og bankene skal gjennomføre dersom en feil oppstår og Finterai får tilgang til personopplysninger.

For modeller som ikke inneholder personopplysninger, vil hverken kontrollene som gjennomføres eller den videre utvekslingen av modellene, være behandlinger i personvernforordningens forstand, siden personopplysninger ikke behandles. Personvernregelverket kommer derfor ikke til anvendelse i slike tilfeller. Hvis det er noen tvil om hvorvidt modellen inneholder personopplysninger eller ikke, bør den behandles som om den inneholder personopplysninger og at personvernregelverket kommer til anvendelse.

Dataminimering

Drøftelsene i dette kapittelet knytter seg til hvordan Finterai kan legge til rette for dataminimering i tjenesten sin.

Utvikling av kunstig intelligens er ofte avhengig av store mengder personopplysninger. Prinsippet om dataminimering (lovdata.no) stiller imidlertid krav om at opplysningene som brukes skal være adekvate, relevante og begrenset til det som er nødvendig for å oppnå formålet de behandles for. Det betyr at en behandlingsansvarlig ikke kan bruke flere personopplysninger enn det som faktisk er nødvendig for å oppnå formålet, og at opplysningene må slettes når det ikke lenger er bruk for dem. Videre innebærer prinsippet om dataminimering at det må velges opplysninger som er relevante for formålet.

Les mer om dataminimering

I lovkommentaren til personvernforordningen pekes det på at kravene til adekvans og relevans, betyr at personopplysningene som behandles må «ha en nær og naturlig sammenheng med behandlingsformålet, og være egnet til å oppnå formålet». Vurderingen av dataminimering er uløselig knyttet til formålet med behandlingen.

Les lovkommentaren til personvernforordningen (juridika.no)

Den behandlingsansvarlige har ansvar for å overholde prinsippet om dataminimering. En leverandør av programvare, som etter en konkret vurdering ikke er å anse som behandlingsansvarlig, vil i utgangspunktet ikke ha et direkte ansvar for å overholde dataminimeringsprinsippet. Det er imidlertid viktig at programvaren som leveres legger til rette for at den behandlingsansvarlige i praksis kan overholde regelverket. I motsatt fall vil ikke leverandørens kunder lovlig kunne benytte programvaren til *behandling av personopplysninger*. Det er derfor viktig at Finterai har et bevisst forhold til dataminimering ved utvikling av tjenesten sin, uavhengig av om de er å anse som behandlingsansvarlig eller ikke.

I sandkassen har vi diskutert hvordan tjenesten til Finterai kan påvirke mengden personopplysninger bankene benytter i arbeidet med å avdekke mistenkelige transaksjoner, og eventuelle tiltak Finterai kan gjøre for å legge til rette for dataminimering. Diskusjonene har altså primært knyttet seg til innhenting av tredjepartsdataopplysninger som ikke kommer direkte fra transaksjonen, og som innhentes fra andre enn kunden selv. Det kan for eksempel være opplysninger som har kommet frem i media. Tredjepartsdata inneholder ikke alltid personopplysninger. Vurderingene i denne rapporten gjelder kun for behandling av tredjepartsdata som er å anse som personopplysninger.

Forholdet til hvitvaskingsreglene – utfordringer med standardisering

Vi har ikke vurdert rettslig grunnlag i dette sandkasseprosjektet. Bankenes forpliktelser til å bidra til å bekjempe hvitvasking, følger imidlertid av hvitvaskingsloven og hvitvaskingsforskriften. Det er derfor nærliggende å anta at bankene, dersom de vil behandle tredjepartsopplysninger med formål å avdekke hvitvasking, må finne et rettslig grunnlag for behandlingen i hvitvaskingsregelverket.

Hvitvaskingsreglene er risikobaserte. Dette innebærer at den enkelte banks forpliktelser til å innhente informasjon etter dette regelverket, avhenger av risikoen den enkelte kunde representerer i den aktuelle banken. Den samme kunden kan ha ulik risiko i ulike banker. I tillegg vil kundemassen i hver enkelt bank bestå av kunder med ulik risiko.

Den risikobaserte tilnærmingen i hvitvaskingsreglene kan skape utfordringer for Finterai når de ønsker å standardisere datakategoriene bankene må benytte i den fødererte læringen. Dataminimeringsprinsippet innebærer at bankene ikke kan behandle flere personopplysninger enn det som er nødvendig for å oppfylle formålet. Et spørsmål som har dukket opp i sandkassen, er om det innenfor rammene av det gjeldende hvitvaskingsregelverket er mulig å finne et minimumsnivå av opplysninger som alltid kan innhentes uavhengig av risiko, og som dermed kan inngå i en standardisering.

Det er Finanstilsynet som fører tilsyn med de rapporteringspliktiges etterlevelse av hvitvaskingsreglene, og en tolkning av dette regelverket faller utenfor rammene for sandkasse-prosjektet til Datatilsynet. Vi kan derfor ikke besvare

spørsmålet. All behandling av personopplysninger krever imidlertid et rettslig grunnlag, og drøftelsene under forutsetter derfor at det er mulig å oppstille et minimumsnivå av opplysninger som kan benyttes i anti-hvitvaskingsarbeidet, uavhengig av risiko.

Dataminimering og føderert læring – behovet for forhåndsdefinerte datakategorier

Noen banker innhenter allerede i dag tredjepartsdata i forbindelse med sitt anti-hvitvaskingsarbeid, men det er ulik praksis mellom bankene knyttet til hvilke data som hentes inn. For at føderert læring skal fungere etter sin hensikt, er det imidlertid nødvendig å samkjøre hvilke data-kategorier bankene behandler. Bakgrunnen for dette er at en modell som er utviklet i Bank A skal trenes i Bank B og C. Disse bankene må da ha tilgang på de samme data-kategoriene som Bank A benyttet ved utvikling av modellen.

Bankene som deltar i den føderterte læringen må altså ha tilgang på de samme kategoriene med personopplysninger. Behovet for hver enkelt kategori av personopplysninger oppstår imidlertid først når en bank bygger en modell som benytter de aktuelle personopplysningene. Noen typer opplysninger, for eksempel dataene i SWIFT-meldinger, kan man legge til grunn at alltid vil være aktuelle. Andre kategorier personopplysninger benyttes derimot sjeldnere eller kanskje aldri. Da oppstår spørsmålet om dataminimering. Er det i tråd med prinsippet om dataminimering å innhente personopplysninger som man på tidspunktet for innsamlingen ikke vet om man vil få behov for? Denne problemstillingen vil trolig være aktuell i større eller mindre grad også hos andre aktører som benytter føderert læring på personopplysninger.

I sandkassa har vi diskutert ulike alternativer for tilpasning av Finterais tjeneste som potensielt kan bidra til at bankene ikke trenger å hente inn de ulike kategoriene med personopplysninger før det faktisk er behov for opplysningene.

Det mest realistiske alternativet som ble diskutert, er at bankene først innhenter nødvendige tredjepartsopplysinger når de beslutter å utvikle en modell som inkluderer opplysningene, eller når de får en modell til trening som krever de aktuelle opplysningene. Finterai har pekt på at en slik løsning kan være teknisk utfordrende, samtidig som det vil føre til forsinkelser i treningsprosessen.

Dersom det innhentes personopplysninger som det viser seg at bankene aldri har behov for, vil de aktuelle opplysningene ikke kunne sies å være nødvendige for det konkrete formålet. Datatilsynet foreslår derfor at systemet bør rigges slik at bankene kan vente med å innhente personopplysninger til de vet med sikkerhet at de vil få bruk for opplysningene. Her er det imidlertid viktig å understreke at Datatilsynets innspill er å anse som veiledning, og ikke utgjør noen lovlighetsvurdering av den planlagte tjenesten til Finterai.

Dataminimering i kunstig intelligens

Finterai mener at modellene bankene i dag bruker for å avdekke mistenkelige transaksjoner, er for svake. Selskapet har en teori om at bankene trenger flere datapunkter i modellene sine for å gjøre en tilfredsstillende jobb med å avdekke mistenkelige transaksjoner, noe de ønsker å legge til rette for i tjenesten sin.

Ved bruk av kunstig intelligens kan man bygge systemer som kan lære, finne sammenhenger, gjøre sannsynlighetsanalyser og trekke konklusjoner langt utover det både mennesker og systemer som ikke benytter kunstig intelligens er i stand til. Dette innebærer at systemer basert på kunstig intelligens vil kunne heve kvaliteten i bankenes anti-hvitvaskingsarbeid. Det er en sannsynlighet for at systemene finner sammenhenger i opplysninger som ikke tradisjonelt har blitt benyttet i anti-hvitvaskingsarbeidet, og som i utgangspunktet ikke er ansett for å ha en nær og naturlig sammenheng med bekjempelse av hvitvasking.

Det kan imidlertid være en utfordring at bankene ikke alltid vet i hvor stor grad ulike tredjepartsopplysninger vil bidra til å oppnå formålet om å avdekke forsøk på hvitvasking, før de har testet dataene over tid. Dersom resultatet fra testingen viser at en eller flere kategorier av personopplysninger har hatt lite eller ingen betydning for å oppnå formålet,

vil de aktuelle personopplysningene ikke oppfylle kravet til relevans. Fortsatt behandling av personopplysningene vil da fort være i strid med dataminimeringsprinsippet.

Men hva med behandlingen av de aktuelle personopplysningene som har skjedd frem til tidspunktet når banken (eller Finterai) oppdager at de ikke har tilstrekkelig relevans for å oppnå formålet? Vil den også ha vært i strid med dataminimeringsprinsippet dersom erfaring viser at opplysningene ikke var tilstrekkelig relevant? Dette er spørsmål vi har diskutert i sandkassa, men som det ikke er et klart svar på. Som så mye annet vil svaret bero på en konkret vurdering.

Det er imidlertid neppe grunnlag for å si at det alltid vil være i strid med dataminimeringsprinsippet å behandle personopplysninger som senere viser seg å ikke ha tilstrekkelig relevans for å oppnå formålet. Ved vurderingen er det blant annet relevant å se hen til begrunnelsen for hvorfor den eller de aktuelle personopplysningene ble valgt i utgangspunktet. For eksempel, var utvalget av personopplysninger helt tilfeldig, eller var det basert på saklige og legitime antagelser?

Videre er det viktig å være oppmerksom på risikoen for at en antagelse er feil, og ha effektive tiltak for å kontrollere relevansen av de personopplysningene som benyttes. Jo lengre tid det tar før man fanger opp og stanser en behandling av personopplysninger som viser seg å ikke være tilstrekkelig relevant, jo større er risikoen for at behandlingen er i strid med dataminimeringsprinsippet. Disse problemstillingene er ikke unike for Finterai. Dette er noe alle som bruker kunstig intelligens-verktøy for å behandle personopplysninger bør være særlig oppmerksom på.

Sikkerhetsutfordringer

Det er ikke gjort konkrete sikkerhetsrisikovurderinger av Finterai sin løsning i sandkassen, men vi har identifisert hva vi mener er de viktigste overordnede truslene og mulighetene for Finterai sin løsning.

Bruk av føderert læring innebærer både styrker og utfordringer når det kommer til informasjonssikkerhet og personopplysningssikkerhet. At føderert læring legger til rette for at man ikke trenger å dele eller aggregere data, inkludert persondata for trening på tvers av aktørene, er en av de viktigste styrkene til denne teknologien.

Samtidig kreves det da at man deler resultatene av treningen, det vil si selve maskinlæringsmodellene som inneholder parametersettene, på tvers av aktørene for å skape en felles modell. Både delmodellene og den felles modellen kan hypotetisk utsettes for angrep hvor opprinnelig data – inkludert persondata – muligens kan rekonstrueres. Dette kalles "modellinverteringsangrep" (model inversion attacks).

Løsningsarkitektur

Uavhengig av egenskapene til føderert læring, har spesifikke valg av løsningsarkitektur og løsningsdesign naturlig påvirkning på sårbarhetsflaten. Det er ikke gjort konkrete vurderinger av løsningsvalg for Finterai, kun en beskrivelse av overordnede problemstillinger som ble diskutert i prosjektet.

Maskinlæring forutsetter ofte store datamengder, som regel i kombinasjon med spesialisert programvare og maskinvare, som ofte realiseres gjennom bruk av skytjenester. Bruken av skytjenester på generell basis er ikke vurdert i prosjektet. Bruken av en relativt ny metode og teknologi som føderert læring, skaper både utfordringer og muligheter. Utfordringer nettopp fordi metoden er ny og alle potensielle sårbarheter både i *algoritmer*, metoder, verktøy og tjenester muligens ikke er kartlagt godt nok ennå. Muligheter fordi føderert læring egner seg for å redusere klassiske sikkerhetsutfordringer, spesielt knyttet til å redusere behovet for overføring, deling og aggregering av store datamengder. Der hvor skytjenester knyttet til kunstig intelligens ofte har som metode å laste opp og aggregere data for sentralisert behandling, gir føderert læring muligheten til å desentralisere og lokalisere databehandlingen.

Trusselaktører

Finterai er et oppstartselskap med begrensede ressurser for å håndtere cybersikkerhetstrusler fra eksterne ondsinnede aktører, til tross for at løsningen og dets verktøy er realisert i en moderne skyløsning. Skyløsninger har mange potensielle sikkerhetsfunksjoner, men det kreves kompetanse og ressurser for den operative driften av dette, i tillegg til driften av kjerneløsningen. Et viktig tiltak vi diskuterte i sandkassen for å redusere generelle cybersikkerhetstrusler er å minimere mengden verdier man trenger å beskytte.

I Finterais tilfelle gjøres dette ved at man ikke laster opp og aggregerer data. Hver deltagende bank behandler egne data i sine egne instanser av systemene og beskytter dette med sine egne krav, ressurser og kapabiliteter. I stedet lastes kun ferdig trente delmodellpakker opp til den sentrale delen av løsningen for sentralisert styring, samordning og kontroll.

Delmodellpakker beskyttes av kryptering (konfidensialitet) og signering (integritet) under overføring og lagring. Den praktiske gjennomførbarheten og effektiviteten av kryptering avhenger av tiltak og arkitektur som Finterai velger å sette inn. Under selve treningen vil modellpakkene dekrypteres. Her har igjen aktørene muligheten til å benytte sine egne krav, ressurser og kapabiliteter for å legge seg på ønsket og egnet sikkerhetsnivå. Eventuelle konkrete metoder og verktøy for dette er ikke vurdert.

Interne trusselaktører er som regel i form av "utro tjenere" som kan ha privilegert tilgang til interne behandlingssystemer. I dette tilfellet har Finterais løsning de samme utfordringene som andre informasjonssystemer, med krav til tilgangsstyring og autorisasjon av brukere. Føderert lærings desentrale natur begrenser muligheten for en utro tjener hos en aktør til kun å få tilgang egne datasett og ikke andre kunders datasett. Utro tjenere hos Finterai selv har i utgangspunktet ikke tilgang til kunders datasett.

Datatilsynets veileder «Programvareutvikling med innebygd personvern» inneholder generelle retningslinjer for risikovurdering av informasjonssikkerheten som en del av løsningsdesignet.

Tilgjengelighet

Løsningens tilgjengelighet ivaretas delvis ved at løsningens konsept i utgangspunktet er desentral. Man er i utgangspunktet ikke avhengig av den sentrale tjenesten for å gjennomføre lokal modelltrening. Det samme gjelder i en produksjonsfase hvor bankene bruker løsningen for dens primære formål, nemlig å identifisere potensielle hvitvaskingstransaksjoner. Alle disse aktivitetene skjer med Finterais verktøy, men kjøres på den enkelte kundes egen plattform/infrastruktur.

Videreutvikling, etterlæring og utveksling av læring forutsetter tilgang til de sentrale tjenestene. Disse sentrale tjenestene er i begrenset grad viktig for den operative driften og den operative tilgjengeligheten fordi de de ikke er en del av den primære tjenesteproduksjonen.

Angrep på maskinlæringsmodeller

Alle maskinlæringsmodeller som er trent med et datasett, også de som ikke benytter føderert læring, kan utsettes for angrep. Et mål for et slikt angrep kan være å rekonstruere data som er benyttet for treningen, inkludert persondata. Ifølge akademisk litteratur om føderert læring og sikkerhetsutfordringer, er modellinvertering vurdert som en spesielt relevant risiko, fordi man systematisk deler maskinlæringsmodeller mellom flere aktører. Føderert læring kan særlig være sårbar for angrep som truer robustheten til modellen, eller personvernet til de som har personopplysninger lagret hos bankene.

Les forskningsartikkelen "Privacy considerations in machine learning" (engelsk)

Angrep på modellen kan skje i to faser: enten i treningsfasen eller driftsfasen.

- **Treningsfasen**: angrep på dette stadiet kan lære, påvirke eller korrumpere modellen. Angriperen kan også forsøke å påvirke integriteten til data som brukes til å trene modellen. En angriper kan også lettere rekonstruere lokale data hos ulike deltakere i denne fasen.
- **Driftsfasen**: angrep på dette stadiet tar ikke sikte på å endre på selve modellen, men forsøker å endre prediksjonene/analysene til modellen, eller forsøker å samle informasjon om vektingen i modellen. Hvis en angriper får informasjon om vektingen i modellen, er det en hypotetisk mulighet for at dette kan brukes for å rekonstruere (helt eller delvis) de lokale dataene (som kan inneholde personopplysninger) som modellen er basert på.

Forskningslitteratur beskriver flere tiltak for å forebygge slike angrep. De viktigste tiltakene er å benytte modeller og algoritmer som antas å være robuste mot angrep. Andre potensielle tiltak er bruk av metoder som Differential Privacy, homomorfisk kryptering eller Secure Multiparty Computation. Konkrete tiltak er ikke vurdert spesifikt som en del av dette sandkasseprosjektet. Ulempene med enkelte av de eksisterende sikkerhetstiltakene er at de tilføyer støy i modellen, i liten grad er testet i praksis og kan dermed redusere nøyaktigheten, eller belaste systemet med høye beregningskostnader.

Les forskningsartikkelen "A Survey on Differentially Private Machine Learning" (engelsk)

Anvendelse av modellinvertering krever er del forutsetninger. Først (spesielt for eksterne aktører) trengs det tilgang til trente modellpakker som i utgangspunktet er beskyttet gjennom blant annet kryptering og tilgangskontroll. Deretter krever selve prosessen for å gjennomføre modellinverteringsangrep både høy og spesialisert kompetanse. De dataene som rekonstrueres, kan bestå av bare deler av opprinnelige datasett og være av varierende kvalitet.

I tillegg avhenger denne typen angrep i stor grad av om algoritmene som benyttes er sårbare for slike angrep. Det utvalget av algoritmer som Finterai foreløpig ser for seg å bruke i det føderterte læringssystemet, ansees som svært lite

sårbare for modellinvertering, da de matematiske grunnprinsippene bak modellene antas å ikke tillate slike angrep. Algoritmene som benyttes vil ikke være åpent tilgjengelig for eksterne aktører. Interne aktører (typisk deltakende banker) som systematisk får tilgang til hverandres ukrypterte treningsmodeller, vil også være aktører som i utgangspunktet har kontraktsfestede forpliktelser ovenfor hverandre.

Det er i sum betydelige hindre og kostnader knyttet til denne type angrep for eksterne trusselaktører, også i kontekst av Finterais løsning. Finterais egen påstand er at deres fødererte læringssystem ikke er mer sårbart for personopplysningsangrep enn det maskinlæringsmodeller generelt er. Føderert læring er likevel en ung teknologi, og det kan være flere sårbarheter som ikke enda er avdekket, og det er derfor begrenset kunnskap som gjør presis risikovurdering krevende.

Veien videre

I sandkassen har Finterai og Datatilsynet utforsket personvernspørsmål i utviklingen av en løsning for antihvitvasking basert på føderert læring. Rapporten er ikke uttømmende for problemstillinger føderert læring reiser i møte med personvernregelverket, og Datatilsynet vil trekke frem innebygd personvern som et område for videre diskusjon.

Innebygd personvern

Prinsippet handler om at det skal tas hensyn til de grunnleggende personvernprinsippene i artikkel 5 i alle faser av livssyklusen til en programvare som behandler personopplysninger slik at de registrertes rettigheter og friheter blir ivaretatt.

Personvern skal integreres i teknologien, og tas med i planleggingsfasen ved utvikling av løsningen. Ivaretakelse av personvernprinsippene skal altså være en naturlig del av utviklingsprosessen, og ikke noe som kommer inn idet en teknisk løsning nesten er ferdig utviklet.

Datatilsynet har utarbeidet en utfyllende veileder om programvareutvikling med innebygd personvern som kan hjelpe virksomheter å forstå og overholde kravene.

Virksomheter som tar personvernet på alvor bygger tillit. Personvernforordningen (artikkel 25) krever at virksomheter skal ta hensyn til de grunnleggende personvernprinsippene i alle faser av livssyklusen til en programvare som behandler personopplysninger, slik at de registrertes rettigheter og friheter blir ivaretatt. Personvern skal integreres i teknologien, tas med i planleggingsfasen ved utvikling av løsningen og være standardinnstilling. Ivaretakelse av personvernprinsippene skal altså være en naturlig del av utviklingsprosessen, og ikke noe som lappes på til slutt. Datatilsynet har utarbeidet en utfyllende veileder om programvareutvikling med innebygd personvern som kan hjelpe virksomheter å forstå og overholde kravene.

Sandkasseprosjektet har ikke hatt kapasitet i dette prosjektet til å utforske i dybden hva innebygd personvern vil si for maskinlæring som er basert på føderert læring, og heller ikke tatt stilling til om Finterai oppfyller artikkel 25. Finterai sin løsning for føderert læring kan i seg selv være en mer personvernvennlig teknologi sammenlignet med mer «tradisjonelle maskinlæringsmodeller», fordi metoden lar deltakerne i det fødererte læringssystemet lære av hverandres

data uten å faktisk dele opplysningene. Det er nettopp denne innebygde begrensningen av videredeling av lokale data som gjør teknologien mer personvernvennlig.

Finterai må likevel forholde seg til kravene i artikkel 25 for å være aktuell for kunder som har en plikt til å velge løsninger med innebygd personvern. Og det ville vært nyttig å utforske hvilke ytterligere tekniske og organisatoriske tiltak i utviklingen av løsningen som kan bygge inn personvernet på en god måte.

Datatilsynet mener også at samspillet mellom personvernregelverket og hvitvaskingsreglene bør kartlegges ytterligere. Det er per dags dato usikkerhet rundt hvordan forholdet mellom personvern- og hvitvaskingsregelverket påvirker hvilke opplysninger virksomheter kan samle inn og bruke for antihvitvaskingsarbeid.

Føderert på andre felt

Fremover vil det være relevant å følge med på nye bruksområder for føderert læring. Metoden er generelt nyttig når:

1. Det finnes få eksempler på minst én klasse data.
2. Mulighetene for datadeling er begrenset.
3. Samarbeid er nødvendig.
4. Det finnes lite relevant data.

Forsikringsselskapenes kamp mot forsikringssvindel har mange likhetstrekk med bankenes antihvitvaskingsarbeid, og er dermed et nærliggende felt for føderert læring.

Løsninger som lærer fra registerdata, kan også være aktuelle for metoden. I Norge samler vi store mengder informasjon om personer i ulike registre, og norske helsedata går for å være blant de beste i verden (www.ehelse.no). Dette gir unike muligheter for å utvikle presise og effektive løsninger, samt forske på sammenhenger om alt fra frafall i videregående skole til pensjonsordninger og helse. Med informasjonsrikdommen følger riktignok et personverndilemma, nemlig at de som behandler registerdata, potensielt kan re-identifisere enkeltpersoner. Men med føderert læring kan ulike aktører trene den samme algoritmen på deres interne datasett, samtidig som data fra den originale kilden ikke deles.

I Norge trenger vi mer kunnskap om personvernvennlig teknologi. At virksomheter som Finterai ønsker å gå i front, og åpent utforsker løsningen i sandkassen, bidrar til å senke fallhøyden og risikoen forbundet med utvikling av nye KI-løsninger, og gir samtidig erfaring av hvordan slik teknologi fungerer i praksis. Vi håper at sandkasseprosjektets vurderinger vil bidra til innovasjon gjennom trygg deling av data og gjøre det enklere for utviklere å etterleve seg til kravene i personvernregelverket.

Datatilsynets regulatoriske sandkasse for ansvarlig kunstig intelligens

Besøksadresse:
Trelastgata 3, Oslo

Postadresse:
Postboks 458 Sentrum
0105 Oslo

sandkasse@datatilsynet.no
Telefon: +47 22 39 69 00

datatilsynet.no/sandkasse
personvernbloggen.no
twitter.com/datatilsynet