Tilsynsførelse af modeller baseret på Machine Learning

e-nettet A/S
April 2019

Indholdsfortegnelse

Indledning

Denne vejledning anviser, hvordan man kan beskrive modeller baseret på Machine Learning (ML) indenfor den finansielle sektor på en sådan måde, at finanstilsynet kan behandle ansøgninger om dispensationer og/eller godkendelser af disse. Målgruppen for vejledningen er dataanalytikere, forretningsudviklere, jurister og andre, der arbejder med Machine Learning indenfor den finansielle sektor.

Vejledningen er udviklet i forbindelse med e-nettets deltagelse i Finanstilsynets *Fintech Lab* henover efteråret 2018. Forløbet er nærmere beskrevet i Finanstilsynets pressemeddelelse af 20. august 2018[1], og nedenfor gengives et udsnit af meddelelsen, der forklarer det overordnede formål:

> *"Der ligger et stort potentiale i anvendelsen af Machine Learning, og Finanstilsynet vurderer, at teknologien i fremtiden vil finde bredere anvendelse på tværs af erhvervslivet, herunder også inden for den finansielle sektor. Finanstilsynet ønsker at bidrage positivt til dette, og i samme ombæring udbygge sin forståelse af teknologien. Derved understøtter testen, at blandt andet regulatorisk stillingtagen vil blive foretaget på et oplyst og betryggende grundlag.*
>
> *Testen af machine learning i FT Lab tjener således flere formål. Ud over en forståelse af metodens anvendelse til det specifikke formål, ønsker Finanstilsynet at opnå indsigt i, hvordan anvendelse af machine learning samt valg truffet på baggrund heraf kan beskrives og dokumenteres."*

[1] Link: https://www.finanstilsynet.dk/Nyheder-og-Presse/Pressemeddelelser/2018/E-nettet-FT-Lab-200818

Vejledningen indeholder en række anbefalinger til god dokumentation af ML-modeller, som kan anvendes på en bred vifte af forskellige typer af teknikker.

Beskrivelser bør så vidt muligt fremstå således, at en lægmand, uden forhåndskendskab til området, kan tilgå materialet. Hvis der er behov for beskrivelser i fagtermer, vil det fremgå af afsnittet.

Definitioner

I nedenstående tabel ses en oversigt over definitioner på termer, der bruges i denne vejledning.

Kunstig intelligens	Kunstig intelligens er et bredt begreb, der bruges om alle typer af algoritmer, der er i stand til at udføre intelligente handlinger.
Machine learning	Machine learning er et underområde af kunstig intelligens, der dækker over statistiske metoder til at udvikle intelligente systemer og træne disse til at lære fra store eller små datasæt.
Deep learning	Deep learning er et underområde af machine learning, der fokuserer på brugen af dybe neurale netværk eller andre komplicerede algoritmer, der med store mængder data kan trænes til at udføre intelligente handlinger.
Model	En model er en algoritme, der er trænet til at udlede sammenhænge mellem et på forhånd specificeret input og output.
Governance	Governance dækker over en organisations strategier, strukturer, systemer, bemanding og arbejdsgange i forhold til arbejdet med udvikling, træning, evaluering, deployment og monitorering af ML-modeller.
Deployment	Deployment refererer til frigivelse af en model i et produktionsmiljø.

Redegør for modellen

Beskriv udførligt hvilket arbejde modellen udfører, så det er tydeligt at forstå for en lægmand. I denne sammenhæng er der en række underpunkter, der bør beskrives.

Formål

Redegør for modellens formål, samt hvorfor man har investeret i at udvikle modellen. Tydeliggør i den sammenhæng hvilken værdi, modellen bringer til både modelanvenderen samt de, der påvirkes af modellen.

Anvendelse

Redegør for, hvordan modelanvenderen har planlagt at ibrugtage modellen herunder

- hvem den skal bruges af
- hvilke processer, den skal bruges i
- hvilke nuværende processer, den eventuelt erstatter

Interessenter

Redegør for, hvilke interessenter, der tager beslutninger om, interagerer med eller bliver påvirket af modellen. Interessenter kan i denne kontekst inkludere, men er ikke begrænset til de, som er nævnt herunder.

Det skal være klart for lægmand, hvem der har indflydelse på modellen og dens resultater, og hvem der påvirkes af modellen.

Interessent	Beskrivelse
Dataleverandør	Dataleverandør leverer de data, der indgår i træningen af modellen. Der er ofte flere dataleverandører. Dataleverandør er underleverandør til Modelleverandør.
Modelleverandør	Modelleverandør indsamler og forædler modellens inputdata. Modelleverandør er ejer og rettighedshaver til modellen og udvikler, drifter og vedligeholder modellen. Modelleverandøren er underleverandør til modelanvenderen og leverer model eller modelresultater til modelanvenderen.
Modelanvender	Modelanvender udbyder et produkt hvori modellen indgår eller anvendes. Modelanvenderen er i denne vejledning en finansiel virksomhed, der er under tilsyn.
Slutkunde	Slutkunde er en forbruger eller virksomhed, som aftager et produkt fra modelanvenderen, hvori modellen anvendes - f.eks. låntager eller investor i realkreditobligationer
Tilsynsmyndighed	Tilsynsmyndighed fører tilsyn med modelanvenderen, herunder dennes anvendelse af modellen – f.eks. finanstilsynet.
Øvrige interessenter	Øvrige interessenter er de interessenter, påvirkes indirekte af modellen, f.eks. ved reducerede arbejdsopgaver som følge af anvendelsen af modellen

Beskrivelse af modellen

Beskriv, både i lægmandstermer og i fagtermer, hvordan modellen fungerer.

- I sektionen rettet mod lægmænd er det tilstrækkeligt at beskrive modellen i overordnede vendinger.

- I sektionen rettet mod fagmænd bør anvendes den terminologi, der normalt anvendes inden for den givne felt, sådan at andre eksperter indenfor området kan opnå en klar forståelse af, hvordan modellen fungerer.

Afhængig af hvilke teknikker, der er brugt til at udvikle modellen, bør angives væsentlige detaljer om modellens design - f.eks. hvilke tekniske komponenter, der indgår i modellen, hvilke typer af (hyper)parametre, modellen består af, mv.

Valg af modeltype

Hvis der er andre oplagte typer af modeller, man kunne have udviklet til at tjene det overordnede formål, bør man redegøre for disse andre typer samt for årsagerne til, at disse er fravalgt. Har man i udviklingsprocessen udviklet flere forskellige typer af modeller og til sidst besluttet sig for at arbejde videre med én type, er det relevant at inkludere statistikker eller andre former for delresultater, man måtte have opnået med de øvrige modeltyper, man har arbejdet med i processen.

Redegør i den sammenhæng for, hvorfor man har valgt den modeltype, man har, samt hvorfor man mener, at denne modeltype tjener formålet bedst.

Udviklingsforløb

Beskriv udviklingsforløbet for modellen, herunder

- det oprindelige rationale for at påbegynde modeludviklingen
- hvilke beslutninger der er taget undervejs i forløbet, samt
- hvilke konsekvenser beslutningerne har haft for det endelige resultat.

Beslutningerne bør understøttes af saglige argumenter, så det klart fremstår, at den model, man er nået frem til, er den bedste til formålet.

Etik

Redegør for de etiske overvejelser i forbindelse med udvikling og ibrugtagning af modellen. Hvis en eller flere af interessenterne, påvirkes i enten positiv eller negativ retning, bør konsekvenserne klarlægges, og moralen i disse uddybes.

Misbrug

Kortlæg de etiske/moralske risici, der kan opstå ved, at ondsindede brugere udnytter modellen på en måde, der ikke var tiltænkt. Man skal derigennem kunne få et klart billede af, hvem kunne have en interesse i at udnytte modellen til egen fordel, hvordan den kan udnyttes, samt hvordan man vil reducere disse risici.

Begrund at de etiske/moralske konsekvenser, som ibrugtagningen af modellen vil medføre, ikke vil svække tilliden til den finansielle sektor.

Datagrundlag

Redegør for, hvordan de data, der ligger til grund for træning af modellen, er opsamlet, hvad de består af, hvor høj kvaliteten er, hvordan de opbevares, og hvordan de opdateres. På baggrund af denne redegørelse bør det for en lægmand være let at forstå, hvilke datapunkter, der bruges som afhængige variable, og hvilke der bruges som uafhængige.

Forklar desuden hvorfor disse datapunkter er udvalgt. Det bør fremstå overbevisende, at de uafhængige variable, man har udvalgt, har en signifikant forklarende effekt på den afhængige variabel, samt at der

ikke er åbenlyse variable, man har udeladt, som ellers kunne have styrket modellens forklaringsevne.

Datakilder

Redegør for alle de variable, man har inkluderet i sin model, herunder

- Datatypen (f.eks. nominal-, ordinal-, interval-, billed- tekstdata eller andet)
- Formatet som dataene er gemt i
- Leverandøren af dataene
- Måden hvorpå dataene er indsamlet
- Konsistensen i data på tværs af leverancer

Det bør ud fra redegørelsen af datakilderne, der ligger til grund for træning af modellen, være overbevisende, at man kan stole på integriteten og kvaliteten af dataene.

Udvælgelse og præprocessering af data

Beskriv dataudvælgelsesprocessen detaljeret. Har man i løbet af udviklingsforløbet tilføjet yderligere data eller udeladt data, der enten har styrket eller svækket modellens forklaringsevne, bør dette beskrives samt begrundes.

Hvis man har anvendt en eller flere teknikker til præprocessering af dataene, f.eks. på grund af særlige forhold i dataenes format(er), bør man beskrive disse teknikker udførligt. Sådanne teknikker kan f.eks. inkludere:

- Normalisering af datapunkter
- Udeladelse af ekstreme observationer (*outliers*)
- Transformation til andre distributioner
- Kodning af kategoriske variable (f.eks. via *embedding)*
- Diskretisering, generering af afledte variable mv.

Beskriv hvad de forskellige teknikker, der er anvendt i forbindelse med præprocesseringen, bidrager med i forhold til modellen.

Hvis der er væsentlige mangler eller huller i de datasæt, der ligger til grund for træning af modellen (enten regelmæssige eller ikke-regelmæssige), bør man forklare, hvordan man tager hånd om disse. Begrund desuden, hvorfor man har valgt den givne metode, samt hvilke konsekvenser dette valg har på modeltræningen.

Skævheder i datagrundlaget

Redegør for eventuelle skævheder i de data, der ligger til grund for modellen. Afhængigt af modellens natur bør det forklares, at datagrundlaget ikke indeholder bias, der kan have negative konsekvenser for de, som påvirkes af modellens anbefalinger.

Til at støtte denne forklaring kan det være en fordel at visualisere forskellige aspekter af datasættet, så en lægmand vil føle sig tryg ved, at datasættet rent faktisk afspejler populationen, der forsøges modelleret, samt der ikke forefindes signifikante skævheder.

Forklar desuden, hvordan man undgår selvforstærkende, negative *feedback loops* i forbindelse med anvendelse og træning af modellen (på engelsk: *runaway feedback loops*).

Annotering af data

Hvis man har annoteret sine datasæt med informationer, der bruges i modeltræning, bør denne annotering beskrives, og processen for annoteringen forklares trin for trin. Redegør for de instruktioner, som de personer, derudfører annoteringsarbejdet, har modtaget, og

tydeliggør, at disse ikke introducerer utilsigtede skævheder i datasættet.

Opdeling i trænings-, validerings- og testsæt

Opdel datasættet i et eller flere trænings-, validerings- og testsæt inden påbegyndelse af modeltræningen.

- Træningsdatasæt bruges til træning af modellen
- Valideringsdatasæt bruges til at udvælge den model, man mener, er bedst til at tjene det overordnede formål
- Testdatasæt bruges til at evaluere og udvælge den endelige model efter endt træning

Det afhænger af modellens natur samt det overordnede formål, hvordan man opdeler sit datasæt. Der er frihed til at vælge, hvordan man gør dette, men forklar og begrund valget, så det for en lægmand er forståeligt, hvorfor man har valgt den givne metode.

Domæneviden

Hvis modellen bygger oven på eller gør brug af domæneviden, enten til træning af modellen eller i forbindelse med anvendelse af modellen, skal denne viden beskrives, og man bør redegøre for, hvordan den er kommet til veje.

Domæneviden kan være viden, som en bankrådgiver naturligt opbygger i forbindelse med sin interaktion med kunder, eller som en valuar udvikler, når han/hun er ude på besøg for at værdiansætte en ejendom.

Redegør for, at den anvendte domæneviden kun er med til at styrke modellen, samt at der ikke er risiko for, at der gennem domæneviden opstår uhensigtsmæssige bias i datagrundlaget, som kan påvirke modellen i negativ retning.

Træning af modeller

Beskriv hvordan træningen af modellen foregår. Beskriv træningen i to sektioner til henholdsvis lægmænd og fagmænd. Forklar hvordan data bliver brugt i træningen, hvordan træningsprocessen fungerer, samt hvilke konsekvenser, den valgte træningsteknik har for modellens resultater.

Hvis der er flere mulige teknikker for træningen af den valgte modeltype, bør man begrunde, hvorfor man har valgt den anvendte teknik. Hvis man har anvendt andre teknikker i løbet af udviklingen, bør man præsentere resultaterne ved anvendelsen af disse, så det klart fremgår, at man har udvalgt den teknik, der bedst tjener det overordnede formål.

Indstillinger

Hvis den valgte teknik til træning af modellen afhænger af en eller flere hyperparametre eller andre typer af indstillinger, bør disse beskrives. Forklar desuden, hvordan og hvorfor de forskellige værdier for disse er blevet valgt, samt hvilke konsekvenser de givne værdier har for den endelige model.

Hvis man anvender særlige fremgangsmåder til at finde værdierne på disse hyperparametre og/eller indstillinger, bør disse beskrives. Redegør desuden for, at processen for at finde værdierne er konsistent og ikke resulterer i væsentlige ændringer i træningen af modellen fra gang til gang.

Opdatering og vedligeholdelse

I takt med at de grundlæggende data, jf. afsnittet Datagrundlag, ændrer sig eller opdateres, kan det være nødvendigt at gentræne modellen løbende. Beskriv planen for opdatering og vedligehold af modellen. Beskrivelsen bør som minimum indeholde følgende oplysninger:

- Udløser for gentræning (f.eks. periodisk eller på basis af nye data)
- Ved periodisk gentræning skal intervallet specificeres
- Ved gentræning på basis af nye data skal regler for påbegyndelse af træningen forklares
- Udgangspunktet for træning (f.eks. træning fra bunden eller fra tidligere model)
- Hvilke data, ud af det samlede træningssæt, modellen gentrænes på

Hvis det er nødvendigt at ændre på hyper-parametre eller indstillinger for at vedligeholde modellens forklaringsevne ved fremtidig gentræning, bør processen for at ændre disse beskrives med henblik på at sikre, at modellen fortsat er i stand til at leve op til sit formål.

Træningsmål

Forklar hvordan træningsmålet (f.eks. den anvendte *loss*-funktion) for modellen er designet og implementeret, samt hvilken betydning det anvendte mål har for træningen af modellen. Tydeliggør, at træningsmålet er i overensstemmelse med modellens overordnede formål, samt at målet ikke på uhensigtsmæssig vis favoriserer en eller flere interessenter frem for andre.

Modelevaluering

For at vurdere om modellen lever op til sit formål på betryggende vis, bør man for hver iteration af modellen foretage en grundig evaluering af modellens præstation på forskellige datasæt, robusthed samt, hvis det er relevant, forklaringsevne.

Redegør for, hvilke målepunkter, der indgår i modelevalueringen samt hvilke reproducerbare eksperimenter, der afvikles for at genere værdier for disse målepunkter for hver iteration af modellen.

Målepunkter

Der er frie rammer for at vælge/designe de målepunkter, der er relevante for at vurdere, hvor godt en model tjener sit formål. Redegør for, at de målepunkter, der er opsat, i sin helhed giver et komplet billede af modellens evne til at leve op til sit formål. Redegørelsen skal være formuleret på en sådan måde, at en lægmand vil kunne læse, forstå og fortolke modellens evne til at leve op til sit formål. Nedenfor er eksempler på målepunkter, som kunne være relevante at anvende.

Klassifikations-modeller	**Regressions-modeller**	**Clusteringmodeller**
Accuracy (Nøjagtighed) Præcision *Recall* *F1-score* *Confusion matrix* *Receiver Operating Characteristic*-kurver (ROC) *Area Under ROC* (AUROC)	Forklaret varians Gennemsnitlig absolut fejl Gennemsnitlig kvadratafvigelse R^2	*Adjusted mutual information*-score (AMI) *Completeness-score* *Fowlkes-Mallows-indeks (FMI)* Homogenitetsscore

Beskriv hvordan disse målepunkter udregnes for hver modeliteration, samt hvordan disse målepunkter fortolkes i kontekst af modellen og dens formål. Redegør for fornuftige niveauer af hvert af målepunkter, og begrund hvorfor netop de niveauer er blevet valgt.

Segmentering

Hvis man arbejder med datasæt, der er påvirket af heteroskedasticitet, bør man (i tillæg til at evaluere modellens præstation på det samlede datasæt) opdele sit testsæt i naturlige segmenter (eller grupper), således at variansen er tilnærmelsesvist konstant på tværs af observationerne i segmentet. De samme evalueringer, som foretages på hele datasættet, bør udføres på hvert segment for at give et indblik i, hvor godt modellen klarer sig på tværs af segmenterne.

Afhængig af modellens natur kan segmenteringen f.eks. være styret af geografi, demografi, indkomstniveau eller andet. Redegør for, hvorfor den givne segmenteringsstrategi er valgt.

Evaluering på validerings- og testsæt

Evaluer hver modeliteration på baggrund af opdelingen af datasæt jf. afsnittet Segmentering. Redegør for, hvor godt den givne model har klaret sig på valideringssættet, som har ført til valg af modellen. Evaluer herefter på testsættet for at give et endeligt indblik i modellens forklaringsevne på data, som den ikke været eksponeret overfor i hverken trænings- eller udvælgelsesprocessen.

Hvis der er væsentlige forskelle på evalueringens resultater af modellen på testsættet mod valideringssættet, kan der være problemer med enten “underfitting” eller “overfitting”. Dette vil i så fald kraftigt svække tiltroen til, at modellen er i stand til at tjene sit formål, og man bør derfor genoverveje de forskellige komponenter, der indgår i modellen, med henblik på at løse dette problem og dermed skabe en mere tillidsvækkende model.

Fortolkning af resultater

Beskriv både hvordan de enkelte målepunkter skal fortolkes, og hvordan den samlede modelevaluering (bestående af flere målepunkter) skal fortolkes. Hvis der er scenarier, hvor en saglig afvejning af de individuelle målepunkter er nødvendig for at producere en overordnet vurdering af modellen, bør man beskrive, hvordan denne afvejning foretages, samt hvis der er særlige forhold, der gør, at denne afvejning skal foretages anderledes under specielle omstændigheder.

Beskrivelsen skal gøre det muligt for lægmand at læse og forstå resultaterne og vurdere om modellen lever op til sit formål på betryggende vis.

Forklaringsevne

Afhængigt af modellens kompleksitet kan det være relevant at udvide modellen eller udvikle værktøjer til at kunne forklare, hvordan eller hvorfor modellen er nået frem til den givne anbefaling for hver af observationerne i validerings- og testsættene.

For modeller, der er umiddelbart forklarlige, såsom eksempelvis lineære regressionsmodeller, logistiske regressionsmodeller, beslutningstræer eller RuleFit-modeller, er det nemmere at fortolke årsagerne til de anbefalinger, som modellen producerer. Hvis modellen er af en sådan karakter, bør man beskrive reglerne bag disse, hvordan de fungerer, og hvorvidt der er nogle undtagelser til dem.

For mere komplekse modeller kan det være relevant at anvende en eller flere teknikker til at analysere modellen efter endt træning. Blandt disse teknikker findes statistiske værktøjer såsom *global surrogate models*, *local surrogate models* (LIME), Shapley-forklaringer mv. I tillæg findes der stærke visualiseringsværktøjer, som bl.a. *Facets Overview* og *Facets Dive*, som kan give en bedre indsigt i modellens datagrundlag og dermed også være understøttende i forhold til at forklare, hvad der har ligget til grund for modellens træning og ultimativt dens vægte. Beskriv og begrund valget af den teknik, der er anvendt til at forklare i modellen – særligt hvis der er flere oplagte teknikker, man kunne have taget i brug.

Deployment

Beskriv hvordan modeller deployes i produktion. Hvis man benytter sig af enten automatiske eller manuelle processer til (gen)træning, evaluering og deployment af modeller, bør disse processer beskrives, og man bør begrunde, hvorfor man har opsat processerne på den givne måde, og hvordan processerne er med til at sikre tilliden til modellen, herunder at den tjener sit formål på betryggende vis.

Migreringsstrategi

Hvis der opstår store udsving i modellens karakteristika mellem gentræning af modellen, eller beslutter man at ændre væsentligt på modellens design efter deployment af første version, bør man sikre sig, at udsvingene i modellen ikke på drastisk vis påvirker modellens interessenter på en måde, hvor de ikke har mulighed for at forberede sig på de mulige konsekvenser, den gentrænede/nye model måtte have for dem.

Man bør derfor i forbindelse med gentræning eller videreudvikling af modellen lægge en migreringsstrategi og følge denne i forbindelse med udrulning af modellen. Migreringsstrategien skal sikre, at deployment af modellen sker gnidningsfrit i forhold til modellens interessenter, samt at tilliden til modellen forbliver høj i migreringsprocessen.

Robusthed og sikkerhed

Redegør for, at modellen er robust over for alle de observationer, den måtte blive udsat i et produktionsmiljø. Herunder at der er taget højde for alle tænkelige kombinationer af inputdata, således modellen ikke producerer utilsigtede anbefalinger på baggrund af specielle inputkombinationer.

Beskriv desuden, hvordan modellen er sikret mod angreb fra ondsindede brugere, der kan drage fordel af særlige output fra modellen. Beskriv valgte foranstaltninger mod angreb samt begrund, hvorfor man mener, at disse foranstaltninger er tilstrækkelige til at sikre modellens integritet og korrekte anvendelse.

Monitorering

Redegør for monitoreringssystemer for modellen i produktion samt hvilke alarmer, der er konfigureret for at detektere, hvis modellen ikke kontinuerligt producerer de resultater, man med ret ville kunne forvente på baggrund af evalueringen af denne på validerings- og testsættene (se punkt 5). Beskriv desuden forretningsgangene for, hvordan man reagerer på en alarm, eksempelvis ved at tage modellen ud af produktion eller opdatere modellen, så den igen lever op til sit formål.

Videreudvikling

Redegør for, hvordan modellen vedligeholdes og videreudvikles, så den kontinuerligt tjener sit formål på betryggende vis. Hvis datagrundlaget ændres, eller der er andre eksterne forhold, der gør, at en videreudvikling af modellen er påkrævet, bør man orientere sig om disse forhold og designe forbedringer til modellen. Registrer alle ændringer, som man foretager på modellen, i en log. Redegør for strategien for versionering af modeller og beskriv herunder, hvordan denne strategi udmøntes ifm. deployment, hvordan strategien påvirker modellens interessenter, og om strategien påvirker integration til eksterne systemer.

Governance

IT-systemer, som udfører Machine Learning bør være underlagt samme Governance krav som ethvert andet IT-system. Udover at belyse de teknologispecifikke forhold, som er nævnt i denne vejledning, bør man redegøre for systemets Governance i øvrigt, eksempelvis men ikke udtømmende, i forhold til

- Fortrolighed, integritet, tilgængelighed og robusthed
- Integrationer til eksterne systemer
- Adgangskontrol
- Backup
- Overvågning
- Fallback-strategier
- Exit-strategier

Persondata

Europa-Parlamentets og Rådets forordning (EU) 2016/679 af 27. april 2016 om beskyttelse af fysiske personer i forbindelse med behandling af personoplysninger og om fri udveksling af sådanne oplysninger[2] (Databeskyttelsesforordningen), finder anvendelse i de tilfælde, hvor der behandles persondata i ML-modellen.

Det norske datatilsyn har lavet en vurdering af brugen af persondata i forbindelse med Machine Learning[3] (og AI generelt), som kan tjene til inspiration.

[2] Link: https://eur-lex.europa.eu/legal-content/DA/TXT/?uri=celex%3A32016R0679
[3] Link: https://www.datatilsynet.no/en/about-privacy/reports-on-specific-subjects/ai-and-privacy/