


**Mathematical Statistics**
**Stockholm University**

# Maskininlärningsmetoder inom prissättning av skadeförsäkring

Mathias Lindholm
Filip Lindskog

## Research Report 2024-04-01

ISSN 1650-0377

## Postal address:

Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

## Internet:

http://www.math.su.se

Mathematical Statistics
Stockholm University
Research Report **2024-04-01**,
http://www.math.su.se

# Maskininlärningsmetoder inom prissättning av skadeförsäkring

Mathias Lindholm och Filip Lindskog

**Sammanfattning**

En väl fungerande försäkringsmarknad bygger på möjligheten för individer att gemensamt dela på ekonomisk risk. Detta möjliggörs genom försäkringskontrakt som låter individen dela sin risk med kollektivet genom att betala en fast premie mot att kollektivet, eller ett försäkringsbolag, betalar den faktiska kostnaden om en skadehändelse inträffar. Förenklat gäller att riskdelningsprincipen kommer fungera om det är osannolikt att många som är med i riskdelningen drabbas av samma eller samtida skadehändelser. En naturlig fråga är hur man bör prissätta försäkringskontrakt, d.v.s. bestämma försäkringspremien. Projektet "Maskininlärningsmetoder inom prissättning av skadeförsäkring" som finansierats av Länsförsäkringars Forskningsfond har bidragit till betydande nya insikter och förbättrade metoder för att besvara denna fråga. Fokus för forskningsprojektet har främst varit den riskbaserade premien som tas fram med försäkringsmatematiska metoder och är tänkt att återspegla den (typiskt) ettåriga förväntade framtida kostnaden för en försäkrad. Projektet har fokuserat på att utveckla och tillämpa maskininlärningsmetoder för prissättning inom skadeförsäkring med fokus på robusthet och prisstabilitet över tid, grad av prisdifferentiering, skattningsfel och återbäring.

# Innehåll

**1 Introduktion** **3**

**2 Metod** **4**

**3 Resultat** **9**

**4 Diskussion** **12**

**5 Slutsatser och rekommendationer** **13**

**6 Appendix: forskningsartiklar** **15**

# 1 Introduktion

Maskininlärningsmetoder är en del i ett pågående och accelererande teknikskifte som väsentligt kommer påverka många delar av ett försäkringsbolags verksamhet. Traditionella analysmetoder kommer gradvis ersättas av maskininlärningsmetoder och många arbetsuppgifter som till stor del har utförts manuellt kommer att automatiseras. De försäkringsbolag som väljer att inte utforska de möjligheter som det pågående teknikskiftet innebär kommer successivt tappa konkurrenskraft. Moderna maskininlärningsmetoder är inte nödvändigtvis svåra att använda, men betydligt mer än så krävs för att metoderna ska kunna användas säkert som beslutsunderlag. En djup förståelse för försäkringsrisker har i alla tider varit grunden för lönsam försäkringsverksamhet och inget talar för att detta ändras av ett teknikskifte. Vårt projekt syftar till att stödja försäkringsbolag att bättre förstå sina försäkringsrisker i en tid då nya analysverktyg successivt ersätter de gamla.

Projektet har fokuserat på metodutveckling och metodanalys inom prissättning av försäkringskontrakt, vilket tillhör kärnverksamheten för försäkringsbolag. Syftet med projektet har varit att utforma, tillämpa och utvärdera olika maskininlärningsmetoder inom prissättning av skadeförsäkring med avseende på robusthet och prisstabilitet över tid, grad av prisdifferentiering, skattningsfel och säkerhetspåslag. Projektets resultat är av hög relevans för allmän prissättningsmetodik inom skadeförsäkring, men i synnerhet för ömsesidiga försäkringsbolag.

Denna rapport ger bakgrund till och sammanfattar resultat och slutsatser från de 8 artiklar som skrivits inom ramen för projektet. De 8 artiklarna, varav en del är publicerade och övriga under granskning av tidskrifter, bifogas för att lätt kunna läsas för en läsare som vill fördjupa sig i specifika delar. 4 av artiklarna studerar och utvecklar specifika maskininlärningsmetoder motiverade utifrån olika behov av bättre metoder för prissättning baserat på försäkringsdata. De övriga 4 artiklarna behandlar frågeställningar såsom värdering och prissättning av försäkringsprodukter med långa avvecklingstider, statistiska metoder för justering och förbättring av en given prissättningsfunktion, hantering och konsekvenser av kontraktsdurationer som kan avvika från försäkringsperiodens (typiskt 1 år) längd, samt analys av och jämförelse mellan olika definitioner av rättvis (diskrimineringsfri) prissättning. Dessa 4 artiklar behandlar frågeställningar där maskininlärningsmetoder naturligt används som verktyg, men frågeställningarna är inte kopplade till specifika statistiska beräknings- och analysmetoder.

Statistisk analys av försäkringsdata för att kvantifiera försäkringsrisker har i alla tider krävt specifik statistisk och matematisk ämneskunskap. Den tekniska utvecklingen inom beräkningsstatistik och maskininlärning, samt

större tillgång till specialiserade system och programpaket för automatiserad dataanalys, leder till att analys av försäkringsdata är lättare att genomföra men svårare att fullt ut förstå. Ett av projektets mål har varit att underlätta för specialister i försäkringsbranschen att förstå och vid behov förbättra den analys som de moderna tekniska verktygen erbjuder. Med specialister avses aktuarier och mer allmänt "insurance data scientists", d.v.s. de som ska tillhandahålla underlag för hur ett försäkringsbolag ska agera mot försäkringstagare och tillsynsmyndigheter.

Projektets forskningsfrågor syftar alla till praktisk användning i försäkringsverksamhet, men är tekniska och teoretiska till sin natur. En läsare som vill ta del av de möjligheter som forskningsresultaten erbjuder behöver därför ha ett visst ämnesdjup i moderna metoder (inklusive maskininlärning) för statistisk analys. Projektet har genomförts i dialog med, och till viss del i nära samarbete med, aktuarier verksamma vid Länsförsäkringar. Projektets resultat har presenterats både för Länsförsäkringars specialister och för specialister inom maskininlärning och försäkringsmatematik vid internationella konferenser.

# 2 Metod

Forskningen inom ramen för projektet tillhör vetenskapsområdena beräkningsmatematik, matematisk statistik och sannolikhetsteori, motiverad av frågeställningar inom försäkringsverksamhet. Maskininlärning är naturligt inkluderad som ett delområde inom beräkningsmatematik och modern matematisk statistik. Forskningen är av teoretisk karaktär, med fokus på metodutveckling, matematisk analys av maskininlärningsmetoder och matematisk analys av centrala frågeställningar inom försäkringsverksamhet. Publika datakällor har använts för att illustrera resultat och jämföra metoder. Forskningen är inte empirisk och ingen data har samlats in inom ramen för projektet.

Aktuariella prissättningsfunktioner beskriver hur den förväntade skadekostnaden beror på kovariater som beskriver försäkringstagaren och det försäkrade objektet. Sådana funktioner bestäms ofta genom regressionsanalys utförd på skadedata. Den traditionella metoden är baserad på generaliserade linjära modeller (GLM) som baseras på teori för s.k. exponentiella familjer och anpassning med maximum likelihood. Moderna maskininlärningsmetoder erbjuder betydligt mer flexibla regressionsfunktioner än de s.k. log-linjära modellerna inom ramen för GLM. En konsekvens, för att undvika överanpassning till data, är modellval baserat på tränings- och testdata, där prediktionsförmågan på osedd testdata avgör vilken modell som bör väljas. Kombinationer av

många regressionsträd ger en flexibel klass av regressionsfunktioner. Gradient boosting maskiner (GBM) har visat sig vara en klass av regressionsfunktioner som presterar mycket väl. Vid analys av t.ex. skadefrekvens är det naturligt att låta Poissonfördelningens medelvärdesparameter bero på kovariater genom att modellera den med en GBM. Men vid användning av modeller som, till skillnad från Poissonfördelningen beror på flera parametrar, är det önskvärt att utvidga en vanlig endimensionell GBM till en flerdimensionell GBM. I artikeln [1] introduceras en generell flerdimensionell GBM. Utgångspunkten är en traditionell endimensionell GBM, som generaliseras till högre dimensioner genom att använda cyklisk s.k. coordinate descent-metod för parameteruppdatering i sökning efter optimala parametervärden. Detta möjliggör olika kovariatberoenden i olika dimensioner. Givet svaga antaganden kan metoden visas konvergera om den negativa log-likelihoodfunktionen är konvex, vilket t.ex. är fallet för flerparametriska exponentiella familjer. För denna typ av fördelningsfunktioner är det dessutom viktigt att utforma lämpliga metoder för att stoppa den iterativa anpassningen (s.k. early stopping) för att undvika att en flexibel regressionsfunktion överanpassas till träningsdata. Ett enkelt dimensionsvis förfarande för early stopping introduceras och mer avancerade metoder diskuteras.

Artikeln [8] introducerar en trädbaserad varierande koefficientmodell (VCM) där de varierande koefficienterna modelleras med hjälp av den cykliska gradient-boosting-maskinen (CGBM) från [1]. Modellering av koefficientfunktionerna med hjälp av en CGBM möjliggör dimensionsvis tidig stoppning (för att förhindra överanpassning) och möjlighet att utvärdera till vilken grad olika kovariater bidrar till att förklara responsvariabeln. Dimensionsvis tidig stoppning minskar inte bara risken för dimensionsspecifik överanpassning, utan avslöjar också skillnader i modellkomplexitet mellan dimensioner. Modellen utvärderas på samma simulerade och verkliga dataexempel som de som används i artikeln Local GLMnet av Richman och Wüthrich (2023), och resultaten visar på en prediktionsförmåga som är jämförbar med deras neurala nätverksbaserade VCM LocalGLMnet.

I artikeln [3] studeras prissättning av skadeförsäkringar. Specifikt presenteras en allmän metod för att fördelningsfritt konstruera en lokalt väntevärdesriktig skattning av riskpremien baserat på en godtycklig initialt föreslagen prediktor. Den resulterande prediktorn är styckvis konstant, motsvarande en partition av kovariatrummet, och per konstruktion auto-kalibrerad. Begreppet auto-kalibrerad betyder att ingen kontraktsgrupp subventionerar en annan eller blir subventionerad av en annan kontraktsgrupp. Två viktiga frågor är vad som är en lämplig uppdelningen av kovariatrummet i delar som svarar mot grupper av kontrakt med identisk riskprofil, samt hur slumpmässigt varierande kontraktsdurationer (till följd av eventuella tidiga uppsägningar

av kontrakt) bör hanteras. En grundläggande idé i analysen är att partitionera förutsägelserna från den initialt givna prediktorn, vilket leder till en partition av kovariatrummet. Två olika tillvägagångssätt för att skapa partitioner diskuteras i detalj med hjälp av (i) durationsviktad binning med lika sannolikhet och (ii) binning med durationsviktade regressionsträd. Givet en partitioneringsprocedur erhålls storleken på den partition som ska användas med hjälp av korsvalidering. På detta sätt får vi en automatisk datadriven tarifferingsprocedur, där antalet tariffceller motsvarar partitionens storlek. Vi illustrerar förfarandet baserat på både simulerade och verkliga försäkringsdata, med både enkla GLM och GBM som initiala prediktorer. De resulterande tarifferna visar sig ha ett ganska litet antal tariffceller samtidigt som de bibehåller eller förbättrar den prediktiva prestandan jämfört med de initiala prediktorerna.

I artikeln [4] diskuteras effekter p.g.a. varierande kontraktsdurationer på konsistens (konvergens) av skattare av medelvärdesparameter och överspridningsparameter i exponentiella spridningsfamiljer (EDF) som utgör standardmodeller för prissättning med GLM inom skadeförsäkring. Artikeln fokuserar på att analysera standardantaganden för GLM-analys där både betingat medelvärde och varians är linjära funktioner med avseende på duration. Dessa standardantaganden tas alltför ofta som givna av prissättningsaktuarier. Vi härleder enkla konvergensresultat som belyser konsekvenserna när de linjära antagandena om betingade väntevärden och varianser inte är uppfyllda. Resultaten visar att: (i) De resulterande medelvärdesestimatorerna har alltid en meningsfull asymptotisk tolkning i termer av den durationsjusterade aktuariellt rättvisa premien. Men, den premien överensstämmer endast med den försäkringstekniska standardpremien (svarande mot en ettårig kontraktsduration) givet att det betingade väntevärdet är linjärt i durationen; (ii) Deviansbaserade skattningar av spridningsparametern i en EDF bör undvikas till förmån för Pearson-estimatorer; (iii) Om inte standardantagandena för GLM-analys är uppfyllda kan konsistens hos spridnings- och variansskattare inte kan garanteras och kan resultera i falsk överspridning som därför leder till felaktiga tolkningar. Resultaten ger explicita villkor som kan användas för att undersöka om den underliggande datagenererande processen kommer ge upphov till till falsk överspridning. Detta illustreras med hjälp av verkliga försäkringsdata, där man kan dra slutsatsen att antagandena om linjäritet i durationen för betingade väntevärden och varianser inte har stöd i skadedata och således ger upphov till icke-försumbar falsk överspridning.

I artikeln [5] introduceras en metod för att skapa en kategorisk GLM baserad på information från en given black-box-prediktor, t.ex. utifrån en godtyckling maskininlärningsmetod. Förfarandet för att skapa GLM:en är som följer: För varje kovariat, inklusive samspel, skapas en kovariatpartition

med hjälp av partiella beroendefunktioner (PD) som beräknas baserat på den givna black-box-prediktorn. För att förbättra den prediktiva prestandan används ett auto-kalibreringssteg för att bestämma vilka delar av varje kovariatpartition som ska behållas och vilka delar som ska slås samman. Givet kovariat- och samspelspartitionerna anpassas en traditionell kategorisk GLM med hjälp av regularisering med ett lassostraff. Den föreslagna metodens prestanda illustreras med hjälp av ett antal verkliga försäkringsdatamängder där GBM-modeller används som black-box referensmodeller. Av dessa exempel framgår att de guidade GLM:ernas prediktiva prestanda ligger mycket nära de motsvarande referens-GBM:ernas. I exemplen har de guidade GLM-modellerna dessutom få parametrar, vilket gör de resulterande modellerna lätta att tolka. Det framgår också att de guidade GLM:erna ofta tenderar att ligga nära motsvarande GBM:er när det gäller s.k. fidelity.

Artikeln [6] fokuserar på begreppen indirekt diskriminering och rättvisa som för närvarande diskuteras både kopplat till prissättning och kopplat till potentiella problem med användning av maskininlärningsmetoder för beslutsunderlag i stort. Indirekt diskriminering och rättvisa är viktiga problem i algoritmiska modeller. Detta gäller särskilt inom försäkring, där skyddade försäkringstagarattribut inte får användas för prissättning av försäkringar. Att helt enkelt bortse från skyddade försäkringstagarattribut är inte en lämplig lösning, eftersom detta fortfarande gör det möjligt att härleda de skyddade attributen från icke-skyddade kovariater. Denna slutledning leder till så kallad proxy- eller indirekt diskriminering. Även om proxy-diskriminering kvalitativt skiljer sig från grupprättvisebegreppen i maskininlärningslitteraturen, har dessa grupprättvisebegrepp föreslagits för att kontrollera effekten av skyddade attribut på beräkningen av försäkringspriser. Syftet med denna artikel är att diskutera skillnaderna mellan direkt och indirekt diskriminering i försäkringsbranschen och de mest populära axiomen för grupprättvisa. I synnerhet visar vi att det ena inte implicerar det andra, eftersom dessa begrepp är kvalitativt olika. Vidare diskuterar vi metoder för förbehandling av indata och metoder för efterbehandling av modeller som uppnår både diskrimineringsfria försäkringspriser och demografisk paritet. Det huvudsakliga verktyget för dessa metoder är teorin om s.k. optimal transport.

I artikeln [7] betraktar vi ett ömsesidigt försäkringsbolag som hanterar ett stokastiskt dynamiskt kontrollproblem med målet att kunna erbjuda försäkringstagarna låga premier som är stabila över tid utan att hamna i tillstånd av låg solvens. Sådana kontrollproblem har traditionellt endast kunnat lösas för kraftigt förenklade modeller som på sin höjd ger en fingervisning hur ett bolag bör agera i en mer realistisk stokastisk miljö. Målet här är att genom välutformad s.k. reinforcement learning (RL) kunna finna optimala prissättningsstrategier utifrån simulerad data från godtyckligt komplexa mo-

deller för skadedata och avkastningsdata för bolagets investerade kapital. RL är en central del av maskininlärning som är svår att analysera teoretiskt, samtidigt som den rönt stora framgångar i en mängd olika tillämpningar, t.ex. algoritmer för avancerade brädspel som människor inte lyckats vinna över.

Ett försäkringsbolags skadekostnader och intäkter från investeringar fluktuerar slumpmässigt över tiden. Försäkringsbolaget måste bestämma premierna innan försäkriningsperioderna börjar, det vill säga innan man vet vilka skadekostnader som kommer att uppstå och utan att veta hur det investerade kapitalet kommer att utvecklas. Försäkringsbolaget står alltså inför ett dynamiskt stokastiskt kontrollproblem. Problemet kompliceras av fördröjningar och återkopplingseffekter: premier betalas innan skadekostnaderna blivit kända och premienivåerna påverkar om försäkringsbolaget attraherar eller förlorar försäkringstagare.

Vi betraktar, i [7], ett prissättningskontrollproblem i diskret tid, formulerat i termer av en Markovsk beslutsprocess. I en förenklad miljö kan den optimala premieregeln härledas med dynamiska programmeringsmetoder. Dessa klassiska metoder är dock inte genomförbara i en mer realistisk miljö på grund av tillståndsrummets dimension och bristen på explicita uttryck för övergångssannolikheter. Vi utforskar tekniker för reinforcement learning, med hjälp av funktionsapproximation, för att lösa premiumkontrollproblemet för realistiska stokastiska modeller. Vi illustrerar hur lämplig den approximativa optimala premieregeln är jämfört med den sanna optimala premieregeln i en förenklad miljö och visar vidare att den approximativa optimala premieregeln överträffar referensregler i mer realistiska situationer där klassiska tillvägagångssätt misslyckas.

Artikeln [2] motiveras av problem som uppstår vid flerperiodsvärdering inom försäkring. Flerperiodsvärdering är kopplat till försäkringsprodukter med långa avvecklingstider, d.v.s. skadeersättning till försäkringstagare betalas i många fall flera år efter försäkringsperiodens slut. Eftersom en del av premien måste reserveras för utbetalning av skadeersättning långt senare behöver pengars tidsvärde liksom kapitalkostnader beaktas för en sund värdering av det framtida skuldkassaflödet. Aggregerade kassaflöden för försäkringsåtaganden motsvarar vanligtvis betalningar flera år in i framtiden. Men försäkringsregelverket Solvens 2 bestämmer kapitalkrav som beräknas för en ettårig horisont, med beaktande av kassaflöden under året och värdet vid årets slut av kassaflöden bortom den tidpunkten. Detta innebär att värdena för försäkringsåtagandena måste beräknas rekursivt, bakåt i tiden, med början från året för de mest avlägsna skadeersättningarna. Att lösa sådana bakåtriktade rekursioner leder till stora beräkningsutmaningar. Därför har enkla och explicita (standard)formler med tvivelaktig begreppsmässig

sundhet blivit etablerad praxis. Nackdelen med dessa förenklade formler är att de är inkonsistenta med sund försäkringsekonomi. I artikeln visas att det finns alternativ till vanligt förekommande ad-hoc-formler som är lätta att använda i praktiken och som bibehåller både ekonomisk tolkningsbarhet och och den konceptuella sundheten hos de ursprungliga principerna för försäkringsvärdering. Matematiskt följer de explicita formlerna av att de ursprungliga principerna för försäkringsvärdering med allmänt tillämpliga approximationer för stora försäkringsportföljer.

# 3 Resultat

Prissättning inom sakförsäkring med regressionsmetoder, t.ex. GLM, är sedan länge vedertaget. Modern regressionsanalys tillåter flexibla regressionsfunktioner som beskrivs av artificiella neuronnät eller kombinationer av regressionsträd. Samtliga regressionsmetoder syftar till att skatta ett betingat väntevärde för en responsvariabel (som t.ex. beskriver skadekostnad) givet kovariater (som beskriver det försäkrade objektet, försäkringstagaren, etc.). I artikeln *Local bias adjustment, duration-weighted probabilities, and automatic construction of tariff cells* [3] studeras justering av lokal bias, effekter av varierande kontraktduration samt automatisk konstruktion av tariffceller, för regressionsbaserad prissättning inom sakförsäkring. För prissättning av försäkringskontrakt söks en premiefunktion så att försäkringsbolagets premieintäkt svarar mot den förväntade skadekostnaden, givet kovariatvärden som beskriver försäkringskontraktet. Inom prissättningslitteraturen har det faktum att premien sätts initialt medan kontraktsdurationen först är känd senare inte noterats i tillräcklig omfattning. Vi visar att korrekt hantering av okänd kontraktsduration vid prissättningstillfället svarar mot att lösa ett regressionsproblem med avseende på durations-viktade sannolikheter, samt att detta är konsistent med traditionell prissättning med GLM under ett vanligt förekommande antagande som bör och kan testas. En autokalibrerad prissättningsfunktion är sådan att premien för varje kontraktgrupp svarar mot den förväntade skadekostnaden för kontraktgruppen. För både traditionella prissättningsmetoder och moderna metoder baserade på maskininlärning är de resulterande prissättningsfunktionerna inte automatiskt auto-kalibrerade. Det innebär att vissa delbestånd subventionerar andra, alternativt blir subventionerade av andra. Av rättviseskäl och för att undvika moturvalseffekter är detta något som bör undvikas. Vi beskriver hur man automatiskt kan justera för sådan lokal bias för en initial prissättningsfunktion. Detta leder även till metoder för automatisk konstruktion av tariffceller som ger en auto-kalibrerad prissättningsfunktion. En mycket intressant ob-

servation är att även om den initiala prediktorn presterar väl leder autokalibreringen av prediktorn till en betydlig reduktion i modellkomplexitet jämfört med den initiala prediktorn, utan att förlora i prediktiv förmåga. Vi illustrerar metoder och effekter för både traditionella prissättningsmetoder (GLM) och maskininlärningsbaserade moderna metoder (GBM) baserat på publikt tillgänglig försäkringsdata. I uppföljningsarbetet *On duration effects in non-life insurance pricing* [4] analyseras i vilka situationer som antagandet om kontraktsdurationer leder till en felaktig överskattning av risk. De teoretiska resultaten illustreras med hjälp av publikt tillgängliga försäkringsdata och vi visar empiriskt att denna typ av överskattning av risk kan vara betydande. Detta är en viktig observation, eftersom felaktig riskbedömning är direkt kopplat till felaktig prissättning.

När prissättning görs med hjälp av maskininlärningsbaserade metoder som t.ex. GBM är grunduppställningen densamma som för mer traditionella GLM. Genom att utgå från en GLM-formulering antas specifika förhållanden mellan t.ex. väntevärde och varians. Dessa antaganden har traditionellt motiverats med beräkningsmässiga förenklingar, men givet billig och lättillgänglig datorkraft är detta ett antagande som kan undvikas. I artikeln *On cyclic gradient boosting machines* [1] generaliseras teorin för GBM-modeller till att kunna hantera mer flexibla familjer av sannolikhetsfördelningar. Detta illustreras på publikt tillgängliga försäkringsdata och speciellt visas att väntevärdet för den individuella skadekostnadsfördelningen är okänslig för förändringar i kovariatvärden, men att t.ex. variansen i skadekostnadsfördelningen påverkas av förändringar i kovariatvärden. Detta är ett exempel på en situation som inte kan fångas med en klassisk GLM eller GBM, där den nya metoden visar hur riskpåslag på premier bör bero på ett kontrakts kovariatinformation.

Maskininlärningsbaserade prissättningsmodeller, som t.ex. GBM-modeller, tenderar att leda till ökad prisdifferentiering, men det är ofta svårt att förklara vilka kovariater det är som driver skillnader i pris och hur kovariaterna samvarierar. I artikeln *A tree-based varying coefficient model* [8] används resultaten från [1] för att skapa en modell som kan ses som en blandning mellan en vanlig GLM och en GBM. Genom att använda denna konstruktion går det att bibehålla förbättringen i prediktiv förmåga som man får från en GBM jämfört med en GLM, men samtidigt enklare kunna förklara vilka kovariater som påverkar priset. Detta är en fördel jämfört med att använda en "black-box" GBM som predikterar väl, men som är svårförklarad. I artikeln visas hur modellen kan tillämpas på publikt tillgänglig försäkringsdata och hur man kan använda modellen för att förklara beroenden i den underliggande skadeprocessen, samtidigt som den prediktiva förmågan fortfarande motsvarar den hos en GBM. En annan ansats som till sin struktur delar

likheter med [3] är artikeln *Black-box guided GLM building with non-life pricing applications* [5] som utgår från en valfri "black-box"-modell och skapar en kategorisk GLM, snarlik de traditionella GLM-baserade tariffmodellerna. I artikeln används en GBM som given "black-box"-modell och vi visar att den föreslagna tolkningsbara metoden presterar jämförbart med den initiala GBM-modellen på publikt tillgängliga försäkringsdata. Liknande det som vi redan observerat i [3] kommer den slutgiltiga prediktorn att ha en reducerad modellkomplexitet, men här baserat på transformerade kontraktskovariater.

Aktuariell analys av prissättning från ett ömsesidigt försäkringsbolags perspektiv studeras i artikeln *Premium control with reinforcement learning* [7]. Analysens utgångspunkt är strävan efter en stabil och låg premie utan att riskera solvens. Analysen är därför i grunden ett stokastiskt optimeringsproblem med multipla mål som måste balanseras. Traditionella metoder för att lösa den typen av optimeringsproblem kräver alltför förenklade modeller för att trovärdigt beskriva hur ekonomiska storheter för ett försäkringsbolag utvecklas stokastiskt över tid samt återkopplingsmekanismer mellan sådana värden. Dessutom kräver traditionella lösningsmetoder att de stokastiska modellerna fullt ut kan preciseras, vilket sällan är fallet för modeller av realistisk komplexitet. Under den senaste tiden har utvecklingen inom maskininlärning, och mer specifikt s.k. reinforcement learning, erbjudit tekniker för att effektivt analysera stokastiska optimeringsproblem för komplexa datagenererande mekanismer som inte fullt ut kan preciseras i termer av explicita övergångssannolikheter. I artikeln beskriver vi hur vi med effektivt utformad reinforcement learning kan finna optimala prissättningsregler i realistiska situationer för ömsesidiga försäkringsbolag.

En fråga som är av intresse för både prissättning och återbäring är rättvisa och att undvika diskriminering. Det vill säga, hur ska en prissättnings- eller återbäringsfunktion fungera för att undvika diskriminering eller uppfylla en given form av algoritmisk rättvisa. I artikeln *What is fair? Proxy discrimination vs. demographic disparities in insurance pricing* [6] jämförs en matematisk definition som formaliserar försäkringsdiskriminering i enlighet med EU-lag med olika definitioner av så kallad "algoritmisk rättvisa" som har blivit populära inom maskininlärningslitteraturen. I artikeln tillhandahålls metoder för att undgå försäkringsdiskriminering och för att uppfylla olika definitioner av algoritmisk rättvisa och ett huvudresultat är att vi visar att dessa koncept tenderar att vara kvalitativt olika. I praktiken betyder detta att de olika definitionerna av "diskriminering" och "rättvisa" är oförenliga och kräver tydlighet från beslutsfattare och tillsynsmyndigheter.

Det övergripande projektet har fokuserat på prissättning för produkter med korta avvecklingstider, d.v.s. där skadekostnaden för ett kontrakt kan betraktas som känd vid kontraktstidens slut. För produkter med långa avveck-

lingstider kan det dröja långt efter kontrakttidens slut innan den slutgiltiga skadekostnaden är känd. Detta är även anledningen till att reservsättning är en utmanande och central aktivitet för försäkringsbolag. För kontrakt med långa avvecklingstider, d.v.s. som sträcker sig över flera år, kan inte premien anses omedelbart intjänad och diskontering och kapitalkostnader bör beaktas för en värdering av en grupp av försäkringskontrakt. Även om metoder för värdering av grupper av försäkringskontrakt av slaget ovan har studerats i den försäkringsmatematiska litteraturen så har beräkningskrävande värderingsmetoder hindrat aktuarier att anamma sådana konceptuellt sunda metoder. I artikeln *Approximations of multi-period liability values by simple formulas* [2] visar vi att för väldiversifierade och inte alltför små grupper av kontrakt approximeras värden för grupper av kontrakt, som normalt kräver utmanande numeriska beräkningar, av explicita matematiska uttryck som lätt kan beräknas numeriskt. Artikelns syfte är att ge aktuarier ett enkelt verktyg för prissättning även för kontrakt med långa avvecklingstider.

# 4 Diskussion

Sammanfattningsvis har projektet spänt över ett förhållandevis stort område inom prissättning av sakförsäkring och projektet har resulterat i metodutveckling inom maskininlärningsmetoder med hög relevans för praktisk användning i försäkringsverksamhet. Minst lika viktigt är att projektet resulterat i viktiga insikter om möjligheter och fallgropar vid användning av maskininlärningsmetoder inom av sakförsäkring. Projektet har därför bidragit till både utveckling av metoder och kunskap som stödjer effektiv och säker användning.

Projektet har haft stort vetenskapligt genomslag då det genererat 8 artiklar, varav 2 är publicerade och övriga är under olika stadier av granskning av ledande tidskrifter. Forskningsresultaten har presenterats vid internationella konferenser och vid akademiska seminarier samt föredrag riktade till försäkringsbranschen. Projektets forskningsområde är mycket aktivt, vilket framgår av den stora mängd artiklar som publiceras i internationella topptidskrifter. I ljuset av detta ska projektets bidrag till forskningsområdet inte ses som ett avslut, utan snarare som en utgångspunkt för fortsatt forskning syftande till praktisk användning. För försäkringsbolag som bejakar den pågående teknik- och kunskapsutvecklingen väntar spännande tider.

# 5 Slutsatser och rekommendationer

Som vi beskrivit introduktionsavsnittet är maskininlärningsmetoder en del i ett pågående och accelererande teknikskifte som väsentligt kommer påverka många delar av ett försäkringsbolags verksamhet. Traditionella analysmetoder kommer gradvis ersättas av maskininlärningsmetoder och många arbetsuppgifter som till stor del har utförts manuellt kommer att automatiseras. Men användning av system som tillhandahåller maskininlärningsmetoder kommer inte räcka. Lönsamhet för försäkringsbolag har i alla tider haft sin grund i en djup förståelse för försäkringsrisker. Samma sak gäller nu och framöver men förutsätter en djup förståelse för de analysverktyg som presterar bäst. Dessa analysverktyg är till stor del baserade på maskininlärningsmetoder.

Vi har i 8 artiklar visat på den moderna teknikens möjligheter men även analyserat risker och potentiella fallgropar som felaktigt handhavande kan medföra. Metoderna och resultaten i samtliga av de 8 artiklar som projektet genererat kan användas i daglig försäkringsverksamhet. Vilka delar av projektet som försäkringsbranschen väljer att ta till sig och fortsätta utveckla kan inte vi styra över. Internationellt drivs en stor del av forskningen inom "insurance data science" av försäkringsbranschen i samarbete med forskare från universitetsvärlden. Detta framgår tydligt för den som tar del av tillämpade forskningstidskrifter och ledande internationella konferenser. I Sverige har utvecklingen inte kommit lika långt, men möjligheterna till en positiv kunskapsutveckling är goda: en utveckling där försäkringsbolag istället för att nöja sig med att vara passiva konsumenter av ny teknik deltar aktivt för att anpassa och utveckla tekniken för sina och branschens behov. Genom att finansiera forskningsprojektet har Länsförsäkringars Forskningsfond tydligt visat att Forskningsfonden tror på samverkan mellan försäkringsbolag och forskare från universitetsvärlden som ett naturligt sätt att möta det pågående och accelererande teknikskiftet mot en ny generation analysverktyg. Vår enkla rekommendation är att fortsätta i den riktningen för att säkerställa att Sverige fortsätter vara en kunskapsnation på försäkringsområdet.

# Referenser


[1] Delong L., Lindholm M., Zakrisson H. (2023). On cyclic gradient boosting machines.
Open access: https://ssrn.com/abstract=4352505

[2] Engler N., Lindskog F. (2023). Approximations of multi-period liability values by simple formulas.
Open access: https://arxiv.org/abs/2301.09450

[3] Lindholm, M., Lindskog, F., Palmquist, J. Local bias adjustment, duration-weighted probabilities, and automatic construction of tariff cells Scandinavian Actuarial Journal 2023:10, 946-973.
Open access: https://doi.org/10.1080/03461238.2023.2176251

[4] Lindholm M., Nazar T. (2023). On duration effects in non-life insurance pricing.
Open access: https://ssrn.com/abstract=4474908

[5] Lindholm M., Palmquist J. (2023). Black-box guided GLM building with non-life pricing applications.
Open access: https://ssrn.com/abstract=4691626

[6] Lindholm M., Richman R., Tsanakas A., Wüthrich M.V. (2023). What is fair? Proxy discrimination vs. demographic disparities in insurance pricing.
Open access: https://ssrn.com/abstract=4436409

[7] Palmborg L., Lindskog F. (2023). Premium control with reinforcement learning. ASTIN Bulletin: The journal of the IAA, Volume 53, Issue 2, May 2023, pp. 233-257.
Open access: https://doi.org/10.1017/asb.2023.13

[8] Zakrisson H., Lindholm M. (2024). A tree-based varying coefficient model.
Open access: https://arxiv.org/abs/2401.05982

# 6 Appendix: forskningsartiklar

Artiklarna [1], [2], [3], [4], [5], [6], [7], [8] som forskningsprojektet genererat bifogas här.