Prediktiv modell gir høy treffprosent på kontroll av selvangivelser

Skatteetaten utvikler stadig flere prediktive modeller for å kunne gruppere skattytere etter sannsynlighet for feil, og tilpasse saksbehandlingen etter dette. Det er få skattytere som gjør feil i selvangivelsen, og det er utfordrende å velge ut selvangivelser til kontroll. For å velge ut selvangivelser for inntektsåret 2014 til kontroll, har vi laget en prediktiv modell som velger ut selvangivelser etter risikoen for feil i de mest brukte fradragspostene. Kontrollene ble gjennomført i 2015. Resultatene viser at modellen har vært en suksess med treff på 71 prosent av de kontrollerte selvangivelsene. I denne artikkelen forteller vi om hvordan vi har utviklet modellen.

Majken Thorsager,
Skatt øst, analyse

Øystein Olsen,
Skattedirektoratet, analyse

Cecilie Foss,
Skattedirektoratet, analyse/
internasjonal stab

Hvert år håndterer Skatteetaten selvangivelser for 3,7 millioner lønnstakere og pensjonister i Norge. Et av Finansdepartementets krav til Skatteetaten er at skattene skal fastsettes riktig og til rett tid. Vi ønsker derfor å forhindre feil i selvangivelsene gjennom å veilede skattyterne og forenkle den elektroniske innleveringen. Skattytere som ikke har endringer eller tilføyelser behøver ikke aktivt å levere selvangivelsen. Dette utgjør tre fjerdedeler av selvangivelsene. Skattytere som har endringer eller tilføyelser må aktivt levere selvangivelsen. Når selvangivelsene er levert, kontrollerer vi dem for å finne de skattyterne som gjør feil.

For å få en oversikt over alle feil i selvangivelsene kartla vi i 2012 et representativt utvalg selvangivelser og fant at det er feil i cirka fem prosent av selvangivelsene. Risikoen for feil øker betraktelig når skattyter har gjort endringer eller tilføyelser i selvangivelsen, noe som i stor grad gjelder bruk av fradragsposter. De fleste av feilene er da også på fradragsposter. Det er også en tydelig sammenheng mellom antall fradragsposter skattyter bruker og risikoen for at skattyter gjør feil (Foss, Thorsager og Olsen 2015). Når det er få feil i forhold til det store antallet skattytere som leverer selvangivelsene, er det en utfordring å målrette kontrollen av selvangivelsene, slik at vi ikke bruker mer ressurser enn nødvendig. Som en del av kontrollutvelgelsen har vi derfor utviklet en prediktiv modell, som kan hjelpe oss med å finne de skattyterne som har størst sannsynlighet for feil. En prediktiv modell bruker historiske data for å forutsi fremtidige hendelser. I Skatteetaten har vi utviklet flere prediktive modeller (se blant annet Christophersen og Larssen 2013, Hussain, Berset og Paulsen 2015). Hensikten er å bruke den informasjonen vi har om skattyter slik at vi kan målrette saksbehandlingen.

I denne artikkelen forteller vi om bakgrunnen for modellen som skal velge ut riktige selvangivelser til kontroll, hvordan vi har utviklet den, og resultatene etter det første året.

Vi må jobbe smartere

Alle selvangivelser gjennomgår en maskinell kontroll som velger ut selvangivelser for nærmere vurdering av våre saksbehandlere. Hvilke skattytere som velges ut til kontroll bestemmes i hovedsak av opplysninger og beløp i enkeltposter i selvangivelsen, hvor vi erfaringsmessig vet at skattytere fører mye feil. Hvert år vurderer Skatteetaten hvilke poster og beløpsgrenser som skal ligge til grunn for utvelgelsen til manuell kontroll. En konsekvens av et slikt fokus på enkeltposter er at det blir en del kontrollutslag på selvangivelser der skattyter for eksempel har endret noe i en post, men hvor endringen ikke nødvendigvis er feil. Fokuset på enkeltposter kan også innebære at feil i andre poster i selvangivelsen enn den som skal kontrolleres, ikke blir rettet.

Flere fradrag øker risikoen for feil

Som nevnt innledningsvis, viste kartleggingen av feil i selvangivelser at risikoen for feil øker når en skattyter krever flere fradrag. Blant de som har brukt fem poster, er det over 60 prosent som har feil på en av dem. Blant de skattyterne som har brukt tre poster, er det om lag 35 prosent som har feil på minst en av postene (Foss m.fl. 2015). Under en halv prosent av personlige skattytere bruker tre eller flere fradragsposter, og vi har egne kontrollfiltre som velger ut disse til kontroll.

De som krever ett eller to fradrag, er en langt større gruppe, og består av om lag 450 000 skattytere. De utgjør 10 prosent av de personlige skattyterne. En relativt liten andel av dem gjør feil, men fordi gruppen er så stor finnes de aller fleste feilene blant skattytere i denne gruppen. De gjør til sammen om lag 100 000 feil. Vi kan ikke manuelt kontrollere alle skattytere som krever ett eller to fradrag, men vi må sikre at vi kontrollerer de som faktisk gjør feil, og unngå å bruke tid på å kontrollere selvangivelser som er riktige.

Målet med vår modell er høy treffprosent

Skatteetaten trengte en modell som kunne hjelpe til med å velge ut hvem av de 450 000 skattyterne som bruker en eller to fradragsposter som skal til manuell kontroll. Hensikten med modellen er å rangere skattyterne i målgruppen etter sannsynligheten for feil på en av 15 utvalgte fradragsposter i selvangivelsen. Da kan vi kontrollere dem som mest sannsynlig har feil, og unngå å bruke ressurser på dem som ikke har feil. Det vi ønsket var å få en høy treffprosent på kontrollene vi gjennomfører. Det var ikke et mål at det skulle være en bestemt fordeling mellom de 15 ulike fradragspostene eller høye beløpsgrenser. Modellen skal heller ikke være i stand til å klassifisere absolutt alle skattyterne, men den skal være god til å finne dem med størst sannsynlighet for feil blant de som bruker en eller to fradragsposter.

Hvordan har vi utviklet modellen?

Å utvikle en prediktiv modell er en lang prosess. Før vi begynte planleggingen høsten 2013, hadde vi kartlagt alle feil i selvangivelsene. Vi hadde derfor et godt grunnlag for å kunne vurdere verdien av å utvikle en prediktiv modell og innenfor hvilket område av selvangivelsen en slik modell var egnet. Innen vår spesifikke målgruppe, skattytere som krever ett eller to fradrag, hadde vi mye informasjon om skattyterne, men vi visste ikke så mye om hva de som gjør feil har til felles. Før vi kunne bygge modellen, trengte vi informasjon om hvilke skattytere i målgruppen som faktisk gjør feil. Vi trakk derfor 15 000 tilfeldige skattytere fra målgruppen og kontrollerte deres selvangivelse i forbindelse med likningen for inntektsåret 2013. Det ga oss oversikt over hvor stor andel av skattyterne i denne målgruppen som har feil på de ulike fradragspostene, og på hva som kjennetegner skattyterne som gjør feil.

Utnytte den informasjonen vi har

Prediktiv modellering handler om å utnytte den informasjonen man har i store datamengder. Det er derfor en forutsetning at man har gode data som er relevante. Det er også viktig å sikre at de dataene en prediktiv modell skal bygges på er enhetlig registrert, og at de vil være tilgjengelige fremover. Vi har ikke data som direkte sier noe om skattemoral eller evnen til å forstå regelverket. Vi har imidlertid informasjon om hendelser som gjør at skattyter må endre på opplysninger i selvangivelsen, og dermed også har økt risiko for å gjøre feil. For eksempel hvis skattyter har byttet jobb og har krav på reisefradrag, eller at ektefellen har endret på rentefradraget i sin selvangivelse og skattyter også selv må gjøre endringer i sin selvangivelse for at selvangivelsene skal bli riktige. Vi har hentet opplysninger fra en rekke ulike registre med opplysninger om skattyterne og deres likningsforhold for året selvangivelsen gjelder og noen år bakover i tid.

Data mining gir muligheter

For å utvikle den prediktive modellen har vi brukt dataminingverktøy. Data mining kan hjelpe oss å oppdage mønstre i dataene som ikke kan gjenkjennes manuelt, og gi oss estimater på forhold vi er interessert i, for eksempel feil i selvangivelsen. Data mining gjør oss i stand til å finne skattytere som gjør feil ut i fra kriterier som ikke er definert av oss på forhånd, eller som vi allerede kjenner til. Dermed kan vi trekke ut ny informasjon fra dataene som vi kan bruke til å lage mer treffsikre prediktive modeller. Data mining hjelper oss å velge ut de viktigste variablene og sette dem sammen på en måte som gjør at skattyterne kan rangeres etter sannsynlighet for feil. Det hjelper oss å utnytte informasjonsmengden vi har i våre registre.

Vi har testet om lag 500 ulike variabler som sier noe om skattyters demografi, hendelser i livet og øvrige opplysninger i selvangivelsen de har levert. I den endelige modellen har vi med 30 variabler. Det er blant annet opplysninger om bruk av fradragsposter i år og i fjor, alder, økonomiske forhold som inntekt og formue og opplysninger knyttet til enkelte poster. Når vi bygger modellen på denne måten, vet vi ikke nødvendigvis hva det er som gjør at en bestemt skattyter blir rangert til å ha stor risiko for feil. Rangeringen er et resultat av komplekse sammensetninger av dataene i modellen.

Gode resultater det første året

Modellen rangerer de skattyterne som har gjort endringer eller tilføyelser på en eller to fradragsposter i den innleverte selvangivelsen. De som rangeres øverst har størst sannsynlighet for feil. Modellen ble tatt i bruk ved behandlingen av selvangivelser for inntektsåret 2014. Vi valgte da å kontrollere de 8 000 skattyterne som ifølge den prediktive modellen hadde størst sannsynlighet for feil. Kontrollene ble gjennomført i 2015. Resultatene viste at 71 prosent av de skattyterne som ble valgt ut, hadde feil i selvangivelsen. Dette er et svært godt resultat, siden estimater fra året før tilsier at det kun er om lag 17 prosent av alle skattyterne i denne målgruppen som har feil på selvangivelsen.

Vi ser også at de skattyterne som ble valgt ut til kontroll på grunn av rangeringen med prediktiv modell, har brukt et bredt utsnitt av fradragspostene. Det tilsier at modellen er god til å rangere skattytere på alle postene, og at alle fradragspostene blir dekket av modellen. Selv om vi ikke har satt krav til størrelsen på fradragene, er det relativt store beløp som rettes, i gjennomsnitt 62 500 kroner per fradrag. Det betyr at de fleste feil på fradragsposter også innbefatter relativt store beløp. I alt er 440 millioner kroner rettet med denne kontrollsettingen. De fleste feilene går i skattyters disfavør, det vil si at inntektsgrunnlaget økes fordi skattyter har krevd et for høyt fradragsbeløp eller et fradrag han eller hun ikke har krav på.

Erfaringene fra det første året med kontrollutvelgelse basert på den prediktive modellen har vært svært gode. Vi kommer derfor til å videreutvikle den og fortsette å bruke modellen til å velge ut selvangivelser som skal til manuell kontroll i årene fremover.

Forutsetninger for å lykkes

Det var ingen selvfølge at vi skulle lykkes med denne modellen. En viktig forutsetning for at vi har oppnådd høy treffprosent, er at vi på forhånd har kartlagt målgruppen godt og funnet et område hvor en slik modell er egnet. Det har også vært avgjørende at kvaliteten på dataene vi har samlet inn er gode. Vi har gjennomført et stort antall tilfeldige kontroller i målgruppen, som har gitt oss informasjon om hvilke skattytere som gjør feil, og hvem som ikke gjør det. Kontrollørene har fått mye informasjon om hensikten med tilfeldige kontroller og gjort en grundig jobb, slik at vi har hatt gode data å bygge modellen på. Dette har vært en nødvendig investering.

Data mining har hjulpet oss å finne sammenhenger i informasjonsmengden vi har om skattyterne. Det er kompliserte sammenhenger som ville vært vanskelig å finne med andre analysemetoder. Modellen har vist at det er mulig å finne fellestrekk ved de skattytere som gjør feil. Dette har blant annet vært mulig fordi vi har mye informasjon om skattyterne og har jobbet med store datamengder over lang tid.

Den høye treffprosenten henger sammen med at vi har kontrollert de selvangivelsene med størst sannsynlighet for feil. Her treffer modellen godt. Blant de som har lavere sannsynlighet for feil, vil treffprosenten sannsynligvis avta noe. Skatteetatens langsiktige mål er å erstatte andre, mer tradisjonelle, metoder for kontrollutvelgelse med bruk av denne og liknende modeller.

Referanser

Christophersen, Knut (2015) ”Færre skjønn og jevnere arbeidsbyrde med prediktive modeller”, *Skatteetatens Analysenytt*, 1/2015: 20-24.

Christophersen, Knut, Paul Gunnar Larssen (2013) ”Nye modeller gir mer effektiv skjønnsbehandling”, *Skatteetatens Analysenytt*, 2/2013: 19–23.

Foss, Cecilie, Majken Thorsager og Øystein Olsen (2015) ”Ny måte å kontrollere selvangivelsen på – fra fokus på enkeltposter til skattytere som gjør feil”, *Skatteetatens Analysenytt*, 1/2015: 6-10.

Hussain, Shahrukh, Anders Berset og Per Arne Paulsen (2015) ”Modeller for effektiv utvelgelse av omsetningsoppgaver til kontroll”, *Skatteetatens Analysenytt*, 1/2015: 16-20.