Prædiktion af bidragssatser

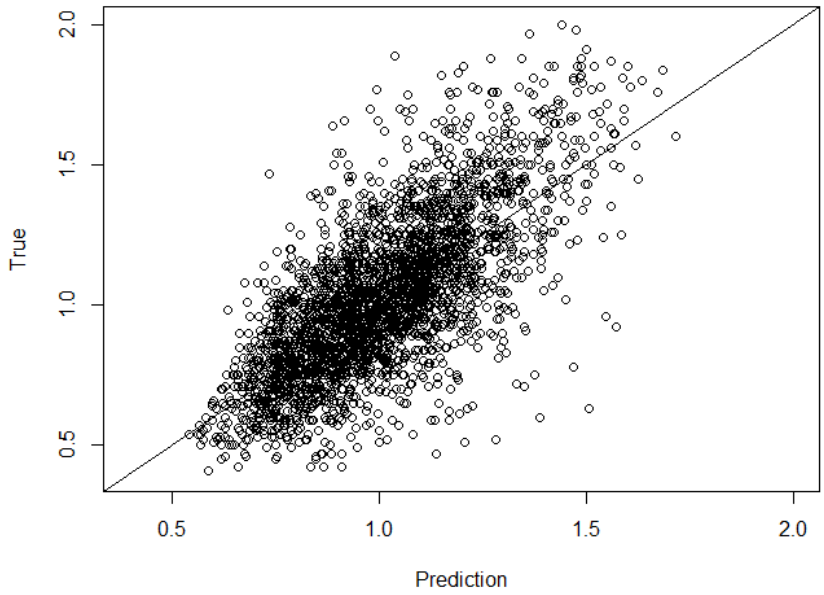
Formålet med denne analyse er at belyse hvorvidt det er muligt at prædiktere næste års bidragssats baseret på forrige års regnskabsinformationer for på den måde at kunne sige noget om hvilken bidragssats som sammenlignelige bedrifter har opnået. Ideelt set skal analysen altså gøre det muligt at sige om en given bedrifts bidragssats er for høj eller for lav ift. lignende bedrifter.
Baseret på regnskabsdata fra bedrifter med regnskaber fra 2014-2015 prædikteres bidragssatsen for 2016. Datasættet indeholder således ikke noget information om hvorvidt enkeltbedriftens opnåede bidragssats er fair eller ej, men håbet er at alle de regnskabstekniske tal som inkluderes i analysen tilsammen kan være med til at forklare dette.

Databehandling

Alle heltidsbedrifter med regnskaber for 2014-2015 og bidragssatser for 2016 mellem 0,3 og 2,0 medtages i træningssættet. Missing values imputeres vha. random forest imputering. For hver bedrift regnes der desuden en middelvariabel mellem de to år samt en udvikling/difference variabel. I det endelige datasæt er der således 317 variable for 2999 bedrifter.

Random Forest 1

Første Random Forest
I den første modelkørsel bruges alle variable til at prædiktere logaritmen til bidragssatsen med standardværdier for de forskellige tuningparametre i Random Forest (se R-koden). For at analysere modellens nøjagtighed ses i første omgang udelukkende på training error i form af mean of squared residuals og andelen af variansen i (log) bidragssatsen som modellen kan forklare.
Træning MSE er 0,040 tage s kvadratroden af dette fås 0,20 hvilket indikerer hvor langt fra modellen er i hver prædiktion i gennemsnit. 0,20 er ret neget når bidragssatserne kun variere mellem 0,4 og 2,0, så denne model er ikke god. Dette ses også ved at andelen af den forklarede varians i (log) bidragssatsen kun er 45,5%. Det vil sige, at denne model kun kan forklare under halvdelen af variationen i bidragssatserne, hvilket er alt for lavt ift. at lave en nøjagtig prædiktion.
Et punktdiagram der holder de prædikterede værdier op mod de sande værdier af bidragssatsen ses nedenfor.



Modellens dårlige prædiktionsevne er umiddelbart svær at forbedre til et niveau der er acceptabelt for en mere nøjagtig prædiktion, men der kan dog foretages nogle justeringer som kan forbedre prædiktionerne en smule.
Det første der kan gøres er at fjerne variable der er "næsten perfekt korreleret" da de i princippet indeholder præcis den samme information. Dette gøres ved at fjerne en af variablene i et variabelpar hvis korrelationen overstiger 0,975. Dette føre til at der slettes 51 variable.

Modificeret Random Forest
Den anden modelkørsel foretages med præcis de samme specifikationer som den første men nu på det reducerede datasæt fra korrelationsanalysen. Der opnås ingen betydelig forbedring.

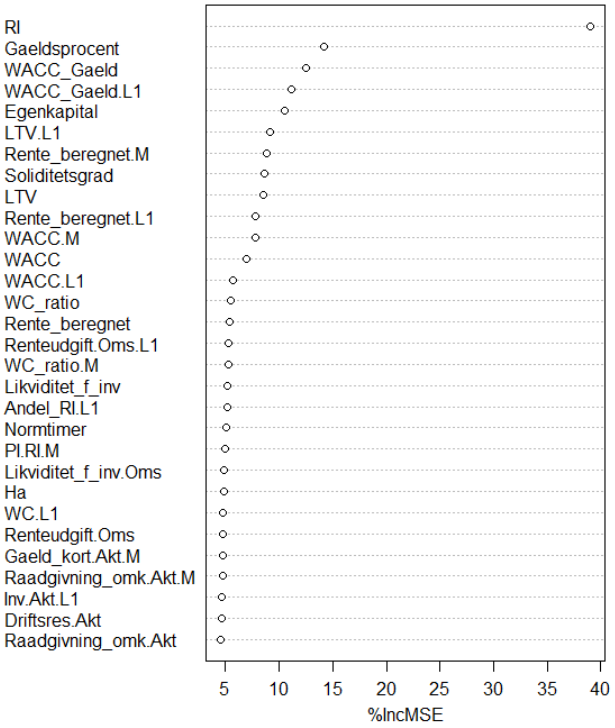
For at forbedre modellen er det muligt at foretage en parameter tuning. Ideen med en parameter tuning er at vælge de optimale modelparametre men samtidig sikre at træerne i skoven er ukorrelerede, hvilket er fundamentet i en god Random Forest model. En måde at sikre dette på er ved at kun at lade modellen vælge mellem et meget begrænset antal af variable i hvert split. Derfor sættes antallet af variable i hvert split til kvadratrode af antallet af variable i datasættet (hvilket er et standard valg i Random Forest modeller med højt korrelerede variable) mens antallet af træer sættes til 250. Der foretages nu en parameter tuning over de resterende variable.

Efter denne tuning køres modellen igen med de tunede parametre. Resultatet af dette er en trænings MSE på 0,037. Tages kvadratroden til dette fås 0,19 mens variansforklaringsgraden stadig blot er på 49%. Der er altså ikke sket en betydelig forbedring af modellen.

Da både korrelationskorrektionen og parametertuningen ikke har resulteret i betydelige forbedringer af modellen konkluderes det at modellen baseret på de tilgængelige data ikke kan opnå en tilfredsstillende punktnøjagtighed. Det sprængende punkt er at de anvendte variable kun er i stand til at forklare halvdelen af variationen i bidragssatsen på tværs af bedrifterne. Der er altså ikke tilstrækkelig med information i de inkluderede variabler til at vi kan sige noget fornuftigt om næste års bidragssats.

Variable importance

Selvom prædiktionsevnen er for dårlig er det stadig muligt at anvende modellens resultater hvad angår variable importance. Variable importance viser hvilket af de anvendte variable der er vigtigst ift. at inddele bedrifterne i undergrupper for at forstå bidragssatserne. Nedenfor ses et "variable importance plot" som viser de vigtigste variable i rangeret rækkefølge.



Her ses det tydeligt at variable RI er den klart vigtigste variabel. Denne variabel fortæller hvilket realkreditinstitut den enkelte bedrift har. Det tyder altså på at kendskabet til en bedrifts realkreditinstitut er første skridt i at få en ide om niveauet af dennes bidragssats. De efterfølgende variable er især nøgletal for bedriftens balance, hvilket er naturligt. WACC_Gaeld er en slags Weighted-Average-Cost-of-Capital på gælden alene. Da bidragssatsen er omkostningerne til realkreditgælden er det ikke mærkeligt at WACC_Gaeld er en vigtig variabel. Disse informationer kan evt. anvendes i en simpel regressions model.

Et nyt datasæt 2014-2016

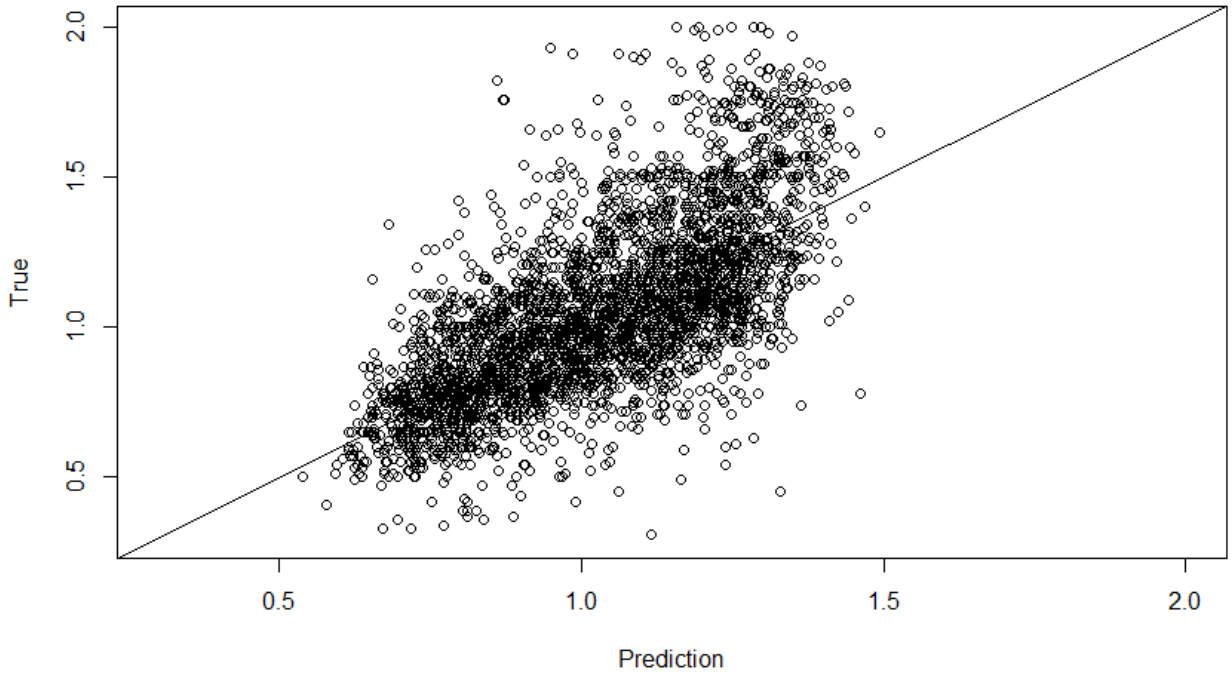
For at sikre at de dårlige resultater i den ovenstående analyse ikke skyldes datasættet foretages samme analyse på et nyt datasæt med forklarende variable fra 2014 til 2016 med det formål at prædiktere bidragssatsen i 2017. Der er således nu tre års data med forklarende variable, hvorfor antallet af variable stiger til 473.

Første Random Forest

I den første modelkørsel bruges alle variable til at prædiktere logaritmen til bidragssatsen med standardværdier for de forskellige tuningparametre i Random Forest (se R-koden). Det er således præcis den samme analyse som ovenfor blot på et år nyere data: 2014-2016 frem for 2014-2015.

For at analysere modellens nøjagtighed ses i første omgang udelukkende på training error i form af mean of squared residuals og andelen af variansen i (log) bidragssatsen som modellen kan forklare. Træning MSE er 0,042 tages kvadratroden af dette fås 0,20. Variansforklaringsgraden er kun 48,2%.

Et punktdiagram der holder de prædikterede værdier op mod de sande værdier af bidragssatsen ses nedenfor.



Sammenlignes disse modelresultater med den forrige model (2014-2015) ses det at der ikke er nogen markant forbedring.

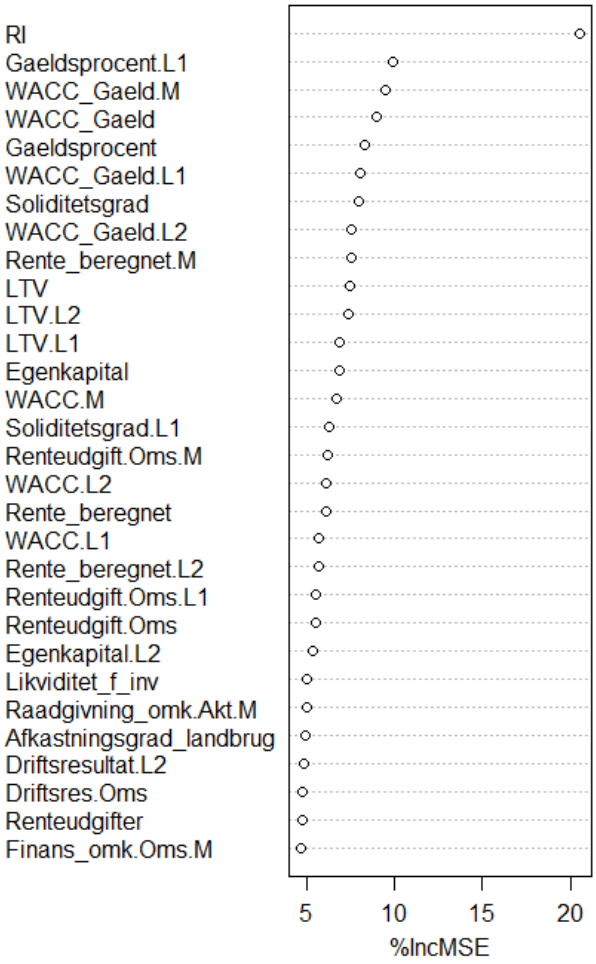
For at gøre analysen helt sammenlignelig med den forrige analyse foretages samme korrelationsanalyse of parameter tuning. Dette resulterer i de endelige modelresultater

- Trænings MSE 0,04 (kvadrat roden af dette er 0,2)
- Variansforklaringsgraden er 49%

Ovenstående resultater bekræfter hvad vi fandt i forrige analyse og medfører således samme konklusion: Det er ikke muligt at prædiktere næste års bidragssats med tilstrækkelig nøjagtighed på baggrund af de tilgængelige regnskabstal.

Variable importance

Som i den forrige analyse er det muligt at se på hvilke variable der er de vigtigste i denne model. Nedenfor ses et "variable importance plot" som er sammenligneligt med det der blev produceret i den forrige analyse.



Konklusionen er igen den samme som i den forrige analyse. Bedrifternes realkreditinstitut er den vigtigste variabel og derefter kommer en række nøgletal som primært omhandler bedrifternes balance.

Konklusion

På baggrund af de ovenstående analyser kan det konkluderes at det ikke er muligt at prædiktere næste års bidragssats med tilstrækkelig nøjagtighed baseret på regnskabstal alene. Dermed er det ikke muligt at identificere de bedrifter som har fået en ”unfair” bidragssats. Dette resultat kan i høj grad tilskrives det faktum at modellen ikke er i stand til at lære hvad en unfair bidragssats er simpelthen fordi denne information ikke er til stede i datasættet. Modellen er altså ikke i stand til at træne sig til en brugbar beslutningsregel ift. at sammenligne enkeltbedrifters bidragssatser med modellens prædiktioner.
Det tyder altså på modellen mangler vigtig information som er afgørende i realkreditinstitutternes fastsættelse af bedrifternes bidragssatser. Dette kunne for eksempel være en menneskelig vurdering af de enkelte bedrifter, hvilket ikke er muligt at inkludere i denne model.
Selvom disse konklusioner umiddelbart er nedslående indeholder de stadig vigtig information om bidragssatsen. For eksempel fortæller den lave prædiktionsevne at realkreditinstitutterne ikke har en systematisk regnskabsbaseret tilgang til bestemmelse af bidragssatserne. Dvs. at bidragssatsen bestemmes ud fra meget andet end fx kombinationen af en lav gældsprocent og en høj afkastningsgrad.