Bidragssats prædiktion

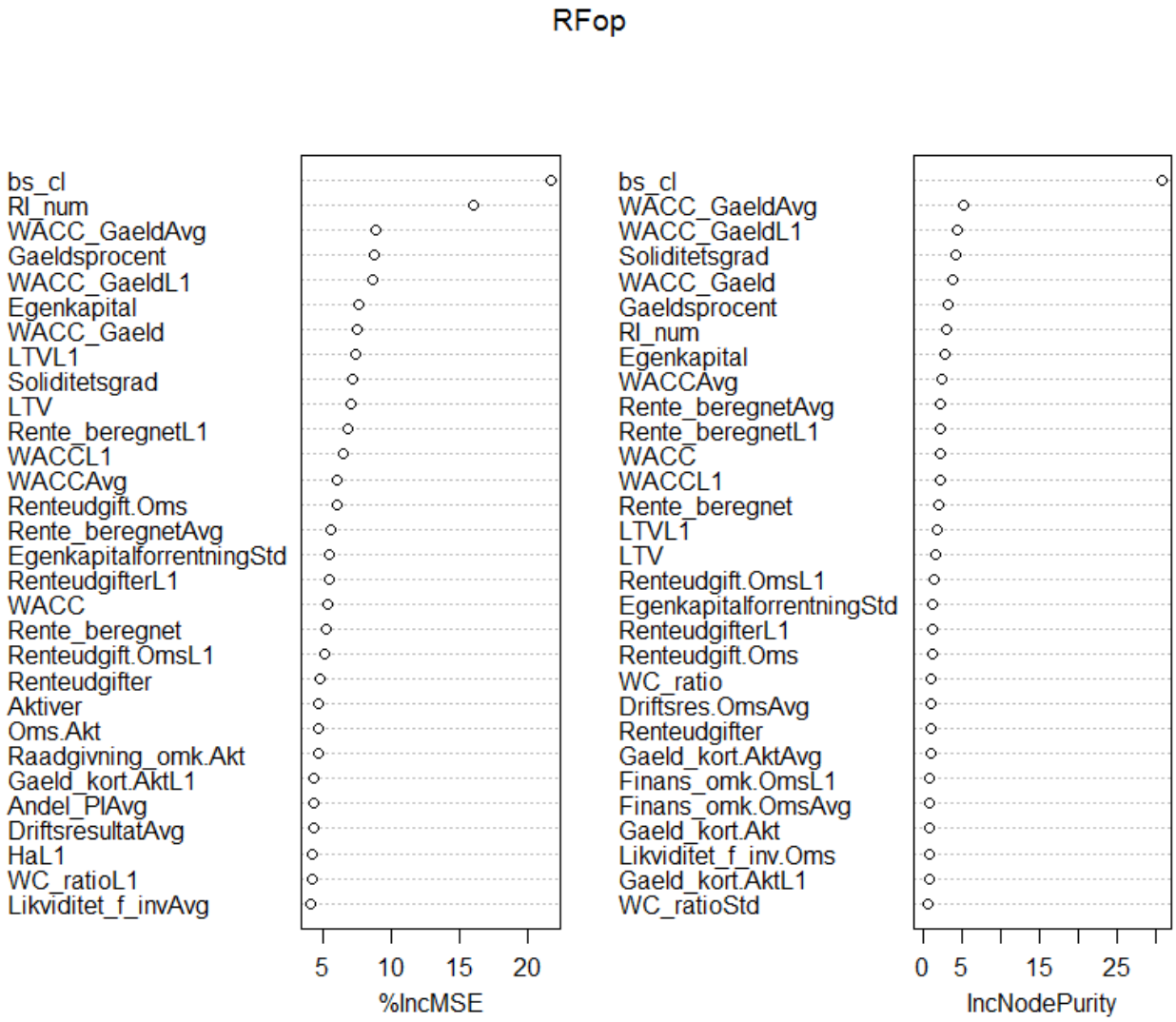
Målet med modellen er at man på baggrund af regnskabsdata kan sige hvilken bidragssats som sammenlignelige bedrifter har opnået.
Formegentlig er punktprædiktionsnøjagtigheden ikke tilfredsstillende og derfor dannes også et prædiktionsinterval på baggrund af et antal bootstrap samples (her 1000).

Data

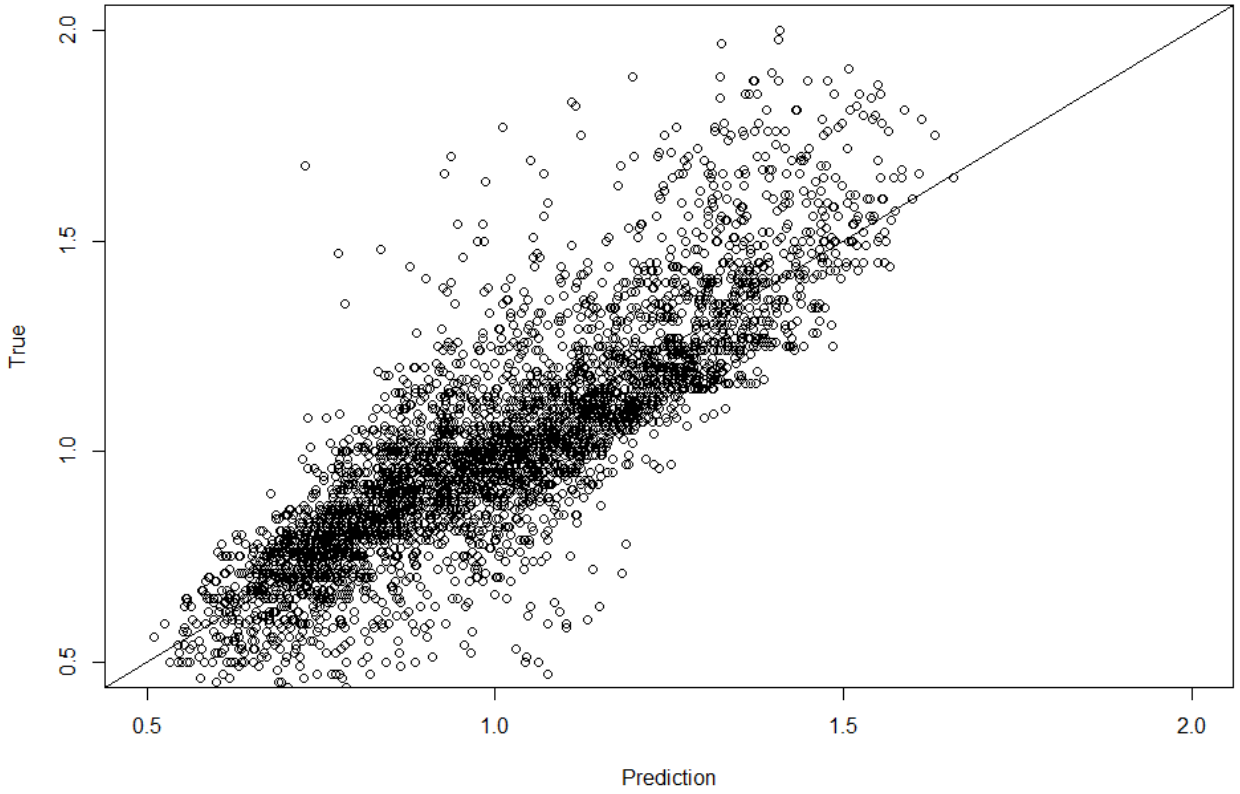
Datagrundlaget er alle heltidsbedrifter i OEDB med mere en 2 års data i perioden (2013-2015) og en bidragssats mellem 0,5 og 2 procent. Der anvendes en lang række økonomiske nøgletal/multipler (prædiktionsvariable) som danner grundlag for prædiktionen.
År med manglende data for de enkelte bedrifter som indgår imputeres med gennemsnittet for den enkelte variabel, da de bedrifter der indgår alle blot mangler et enkelt års data.
Det færdige datasæt indeholder således en linje for hver bedrift (CVR-nummer) alt i alt 3948 bedrifter med 530 variable for hver. For hver bedrift kendes den sande bidragssats i 2016 og det er den man ønsker at ramme med modellen.
For at sikre at hver skov der dannes er så tæt på uafhængig som mulig, fjernes en af variablene i et korrelationspar hvis korrelationen mellem dem er over 0,975.
Der dannes et samlet datasæt, samt et datasæt for hvert realkreditinstitut.

Random Forest analyse - Alle

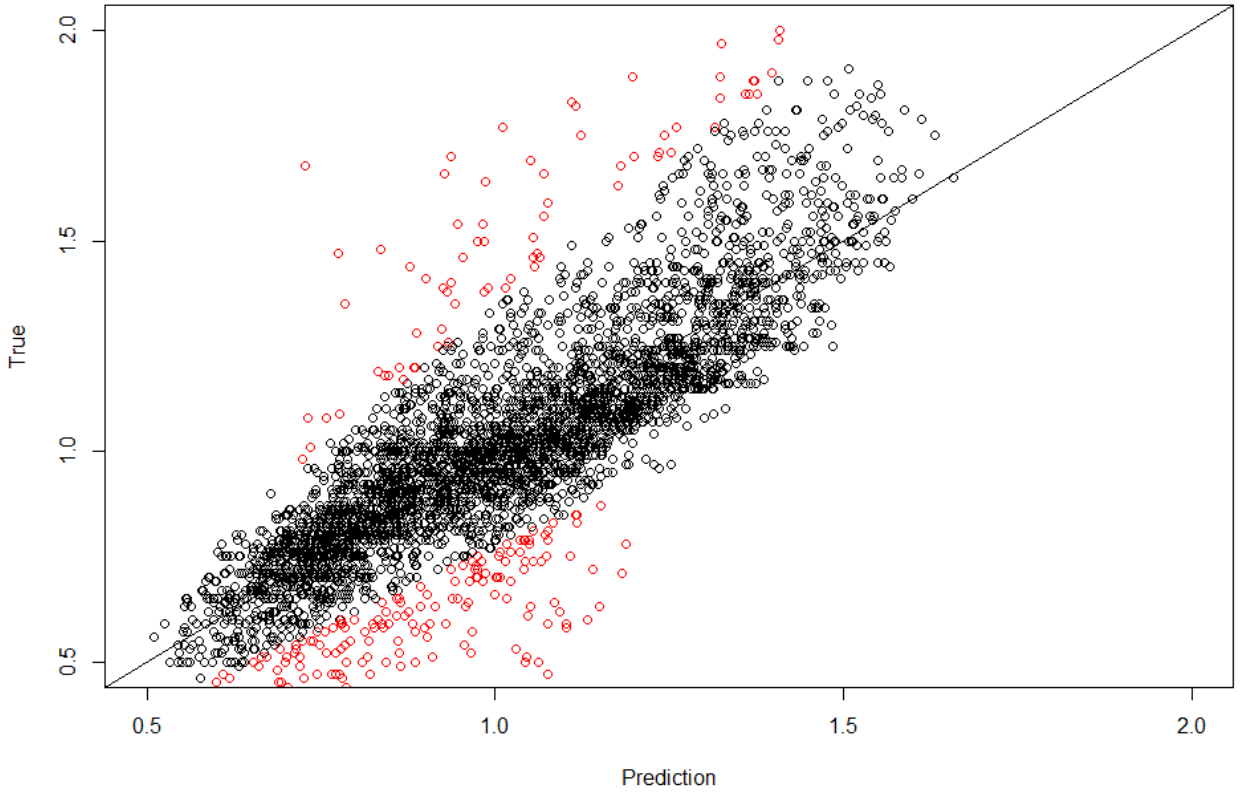
Modellen prædikterer logaritmen til bidragssatsen på baggrund af de valgte variable. Igen vælges modelspecifikationen således at der opnås uafhængighed mellem træerne i skoven.
Der foretages både en analyse for alle samlet og for hver realkreditinstitut for sig.
Med modellen kan ”variable importance” beregnes. Ved brug af alle observationer samtidig opnås for en kørsel:



De to med afstand vigtigste er bs_cl og RI_num. bs_cl indikerer et interval for den sande bidragssats.
RI_num er en numerisk indikator af de forskellige realkreditinstitutter. Resten af variabelnavnene taler for sig selv. Det er tydeligt at se de vigtige variable er dem der beskriver balancen på bedrifterne.
Det er nu muligt at holde den sande bidragssats op mod den prædikterede ved følgende plot:



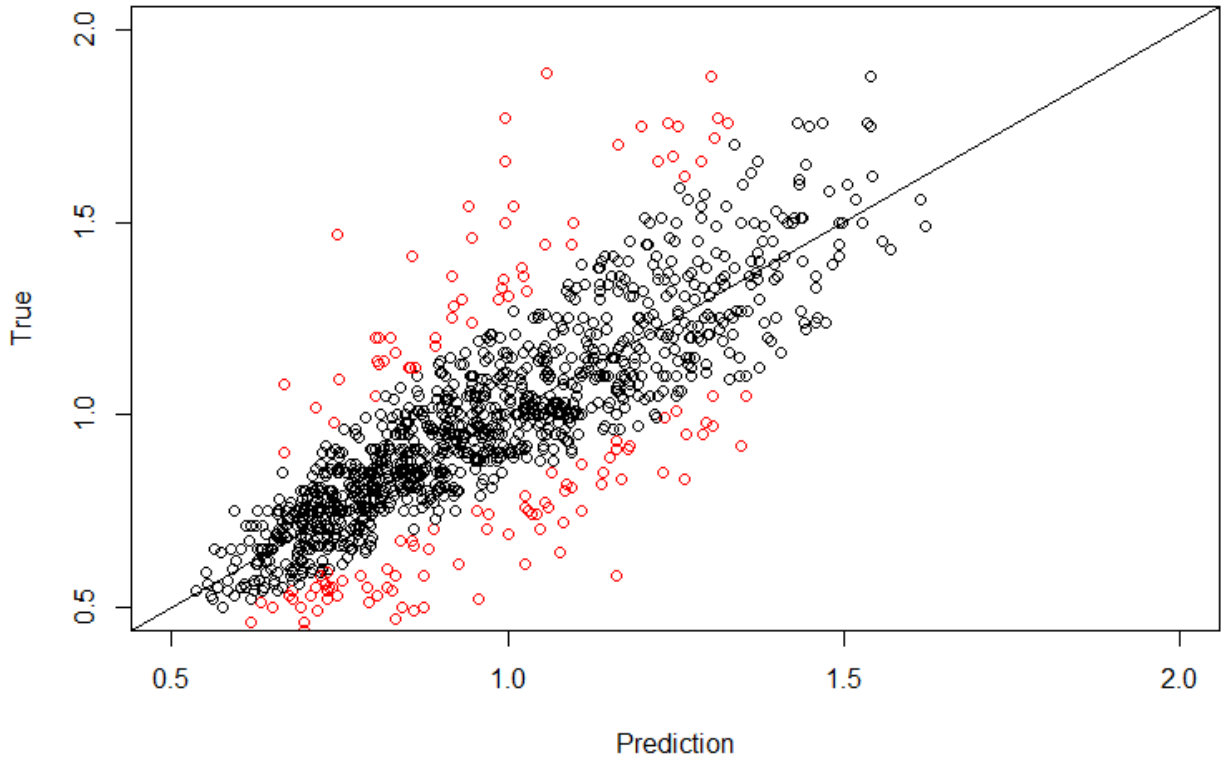
Som det fremgår af plottet er prædiktionen langt fra perfekt. Modellen kan forklare omkring 71 % af variansen og giver en root-mean-square-error (RMSE) på 0,14 % ved en kørsel.
Det skyldes formegentlig at modellen mangler den afgørende information omkring hvem der har fået en fair hhv. ikke fair bidragssats og derfor prøver at fitte alle over en kam, hvilet ikke er muligt.
For at imødekomme dette antages det at residualerne fra modellen er en blanding af to normalfordelinger. Dvs. en fordeling for de bedrifter der har fået en fair bidragssats og en anden fordeling for dem der har fået en unfair bidragssats. Det er nu muligt at identificere hvilke bedrifter der hører til hvilken fordeling og dermed sige om modellen indikerer at de har fået en fair eller unfair bidragssats. En unfair bidragssats vil ofte være en som modellen har svært ved at prædiktere.
I plottet nedenfor viser de røde prikker de bedrifter som ifølge modellen har fået en unfair bidragssats. Som det ses identificeres hovedparten af de bedrifter som modellen har meget svært ved at prædiktere. Det ses også at det er sværere at identificere bedrifterne desto større deres sande bidragssats er. Det skyldes det naturlige nulpunkt.



Ved at holde de røde bedrifter ude opnås en RMSE på 0,11%.

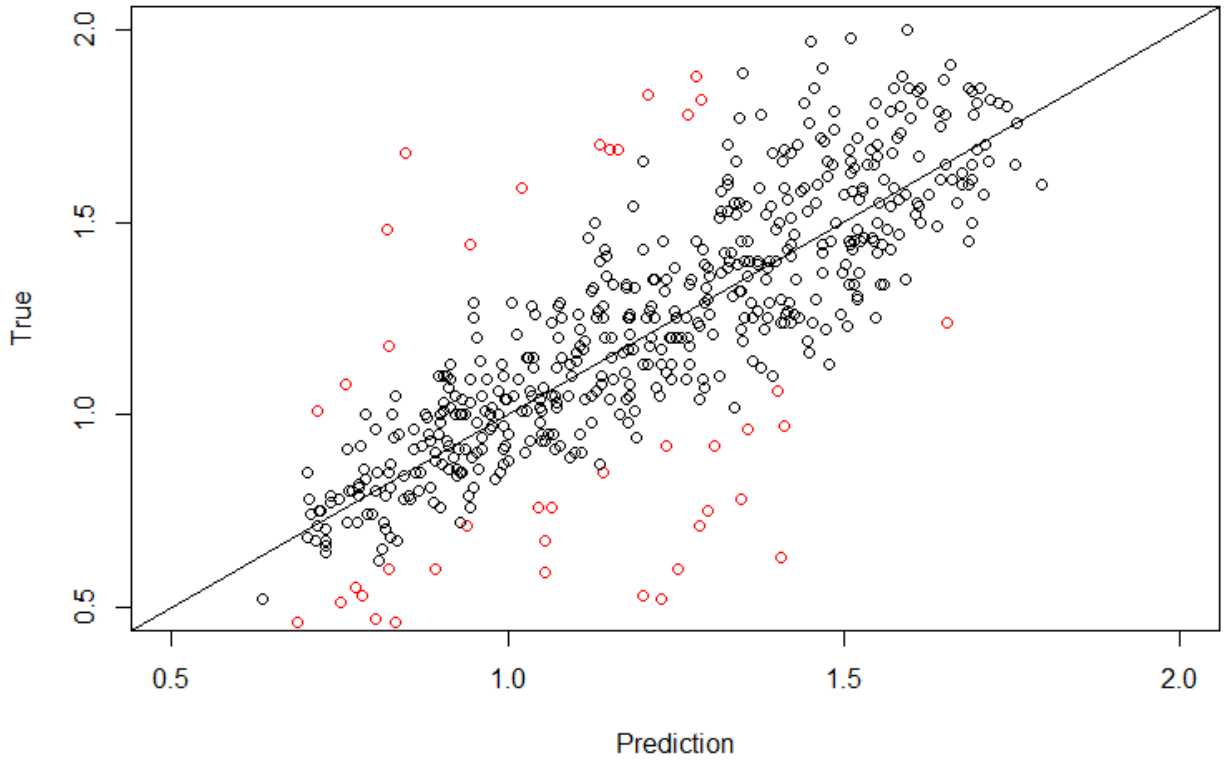
Random Forest analyse – Nykredit

Nedenfor ses resultatet af en kørsel for bedrifter med Nykredit som realkreditinstitut. Det er vigtig at notere sig at de røde prikker, som indikerer en unfair bidragssats, nu kun kan fortolkes i forhold til Nykredits kunder. Det vil sige, at det er de bedrifter der har fået en bidragssats som modellen har svært ved at forklare baseret på de andre Nykredit kunder.



Random Forest analyse – RD

Nedenfor ses resultatet af en kørsel for bedrifter med RD som realkreditinstitut. Det er vigtig at notere sig at de røde prikker, som indikerer en unfair bidragssats, nu kun kan fortolkes i forhold til RD's kunder. Det vil sige, at det er de bedrifter der har fået en bidragssats som modellen har svært ved at forklare baseret på de andre RD kunder.



Bootstrapping

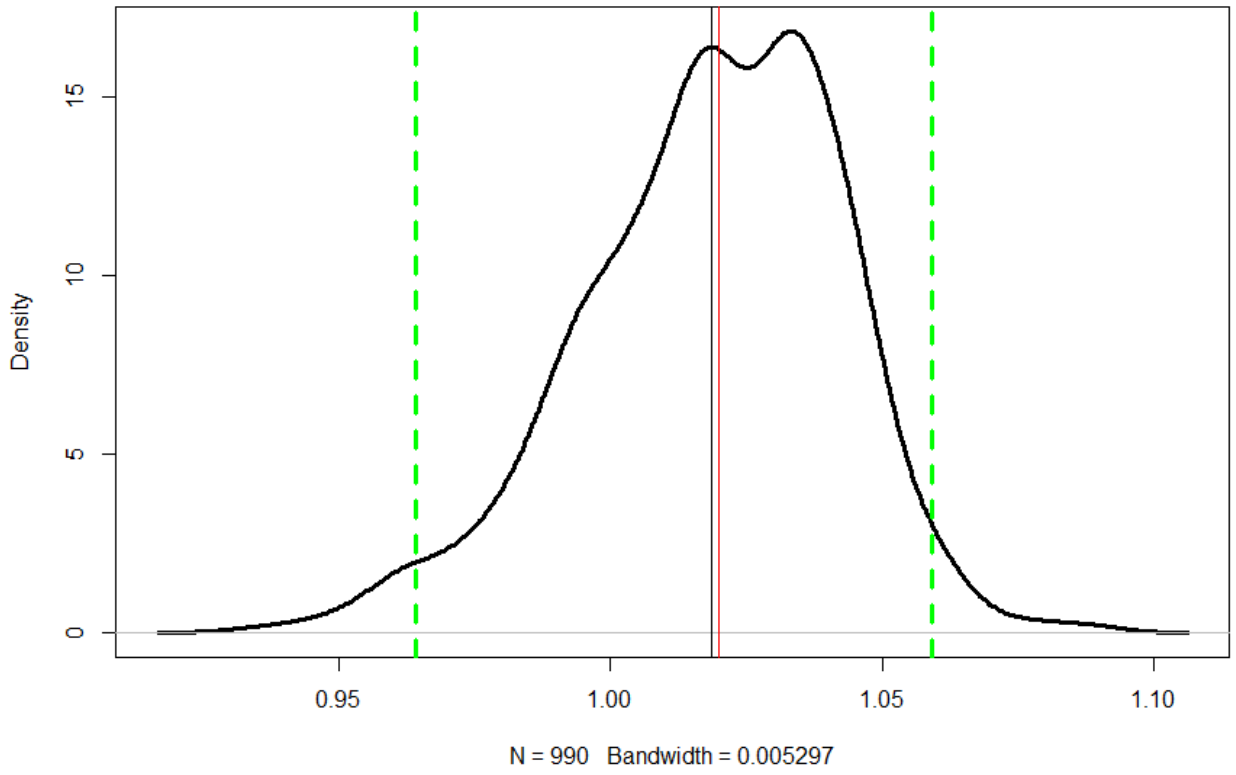
For at komme med prædiktionsintervaller og gøre resultaterne mere robuste udvides analysen med en bootstrapdel. Der dannes således 1000 bootstrapsamples med tilbagelægning, hvorefter hele analysen foretages 1000 gange, således at der opnås 1000 prædiktioner for hver bedrift. Ved denne tilgang er det også muligt at sige en sandsynlighed for at en given bidragssats er fair eller unfair.
Der er meget store udsving i usikkerheden, men den gennemsnitlige prædiktion rammer i nogen tilfælde meget tæt på.
Nedenfor gennemgås resultaterne for bootstrapanalysen baseret på bedrifter som har Nykredit som realkreditinstitut. Dette sammenlignes med resultaterne fra bootstrapanalysen baseret på alle bedrifter over en kam. I datasættet er der 1204 bedrifter der har Nykredit som realkreditinstitut og bootstrap analysen giver altså 1,2 millioner prædiktioner. For Nykredit er procenten af varians forklaret af modellen omkring 84%. I det tilsvarende datasæt for alle bedrifter er der 3948 observationer. For alle bedrifterne i en analyse hvor realkreditinstituttet indgår som forklarende variabel er procenten af varians forklaret omkring også omkring 84%.

Et konkret CVR-nummer er udtaget til illustration af en bedrift som modellen rammer relativt godt. Denne bedrift har indgået i modellen for Nykredit 990 gange og prædiktionerne af bidragssatserne ses i det første tæthedsplottet nedenfor. Dette plot indeholder desuden nogle vertikale linjer som illustrere vigtige værdier:

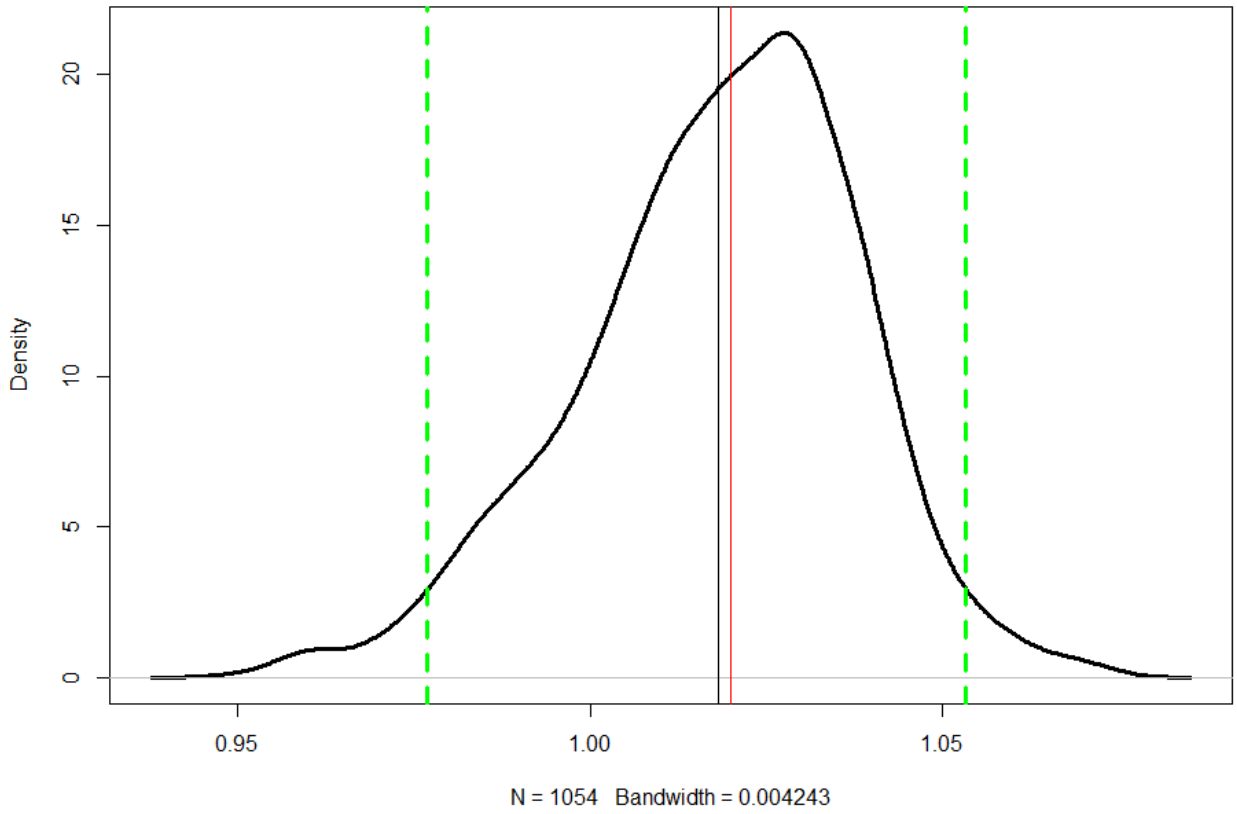
- Den rød linje viser den sande bidragssats på 1,02
- Den sorte linje viser gennemsnittet af alle prædiktioner som er 1,019

- De stiplede grønne linjer, som viser konfidensintervallet på et 5% niveau. Det vil sige, at modellen med 95 % sikkerhed kan sige at den sande bidragssats bør ligge mellem 0,96 og 1,06.

Desuden siger modellen at denne bedrift med har en sandsynlighed på 0% for at have fået en unfair bidragssats. Med andre ord, denne bedrift har en fair bidragssats.



Til sammenligning ser tæthedsplottet ved brug af alle bedrifter på tværs af realkreditinstitutter således ud:

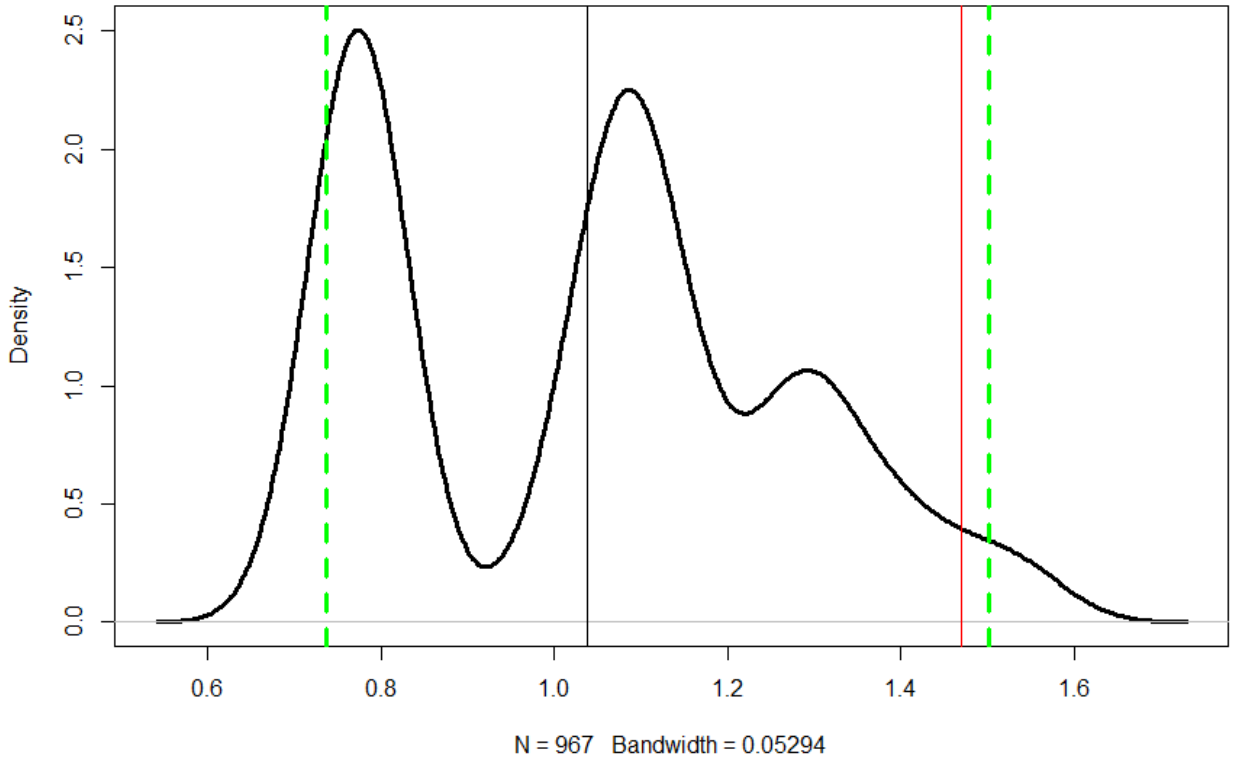


Som det ses er der ikke meget forskel på de to. Den samme prædiktion opnås, mens konfidensintervallerne er en smule mere snævre på hhv. 0,98 og 1,05.

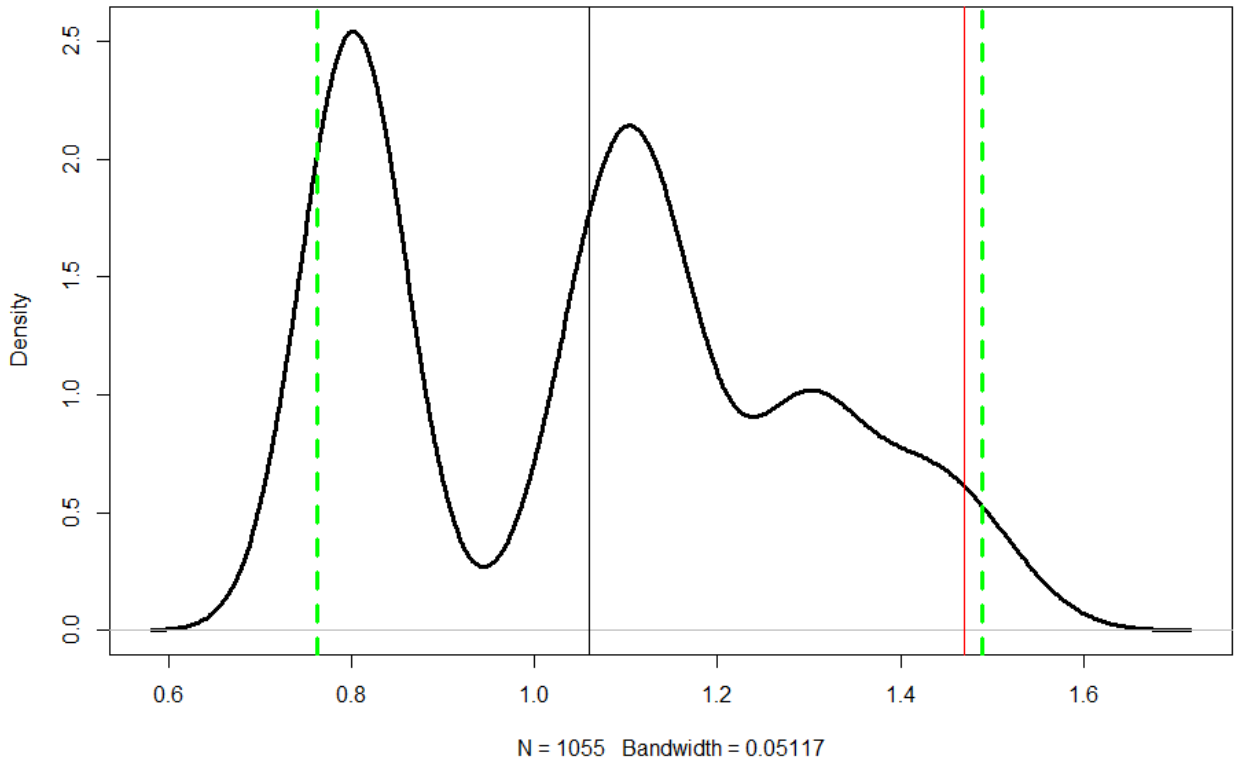
Et andet konkret er udtaget til illustration af en bedrift som modellen rammer relativt dårligt. Denne bedrift har indgået i modellen for Nykredit 967 gange og prædiktionerne af bidragssatserne ses i tæthedsplottet nedenfor. Dette plot indeholder desuden nogle vertikale linjer som illustrere vigtige værdier:

- Den rød linje viser den sande bidragssats på 1,47
- Den sorte linje viser gennemsnittet af alle prædiktioner som er 1,04
- De stiplede grønne linjer, som viser konfidensintervallet på et 5% niveau. Det vil sige, at modellen med 95 % sikkerhed kan sige at den sande bidragssats bør ligge mellem 0,74 og 1,50.

Desuden siger modellen at denne bedrift med har en sandsynlighed på 83% for at have fået en unfair bidragssats. Med andre ord, denne bedrift har fået en unfair bidragssats, selvom den ikke er blandt de højeste.



Til sammenligning ser tæthedsplottet ved brug af alle bedrifter på tværs af realkreditinstitutter således ud:



Heller ikke her er der den store forskel sammenlignet med den institutspecifikke model. Den gennemsnitlige prædiktion er 1,06 og konfidensintervallet er 0,76-1,49.

Angående variable importance er der også for hver bootstrapsample en prioriteret liste. Da det tager lang tid at køre, har jeg valgt kun at se på de 20 første samples for hhv. Nykredit og DLR. For begge realkreditinstitutter er bs_cl den vigtigste parameter i alle bootstrapsamples ligesom det også er vist tidligere. De næste vigtigste variable ligger meget tæt placeret og vigtigheden varierer derfor fra sample til sample. For Nykredit består top 10 hovedsageligt af

- Egenkapital
- WACC_GaeldL1
- Gaeldsprocent
- WACCAvg
- WACC
- WACC_Gaeld
- Soliditetsgrad
- LTV
- GE_forhold
- Rente_beregnet

Disse er i store træk også de vigtigste for DLR.

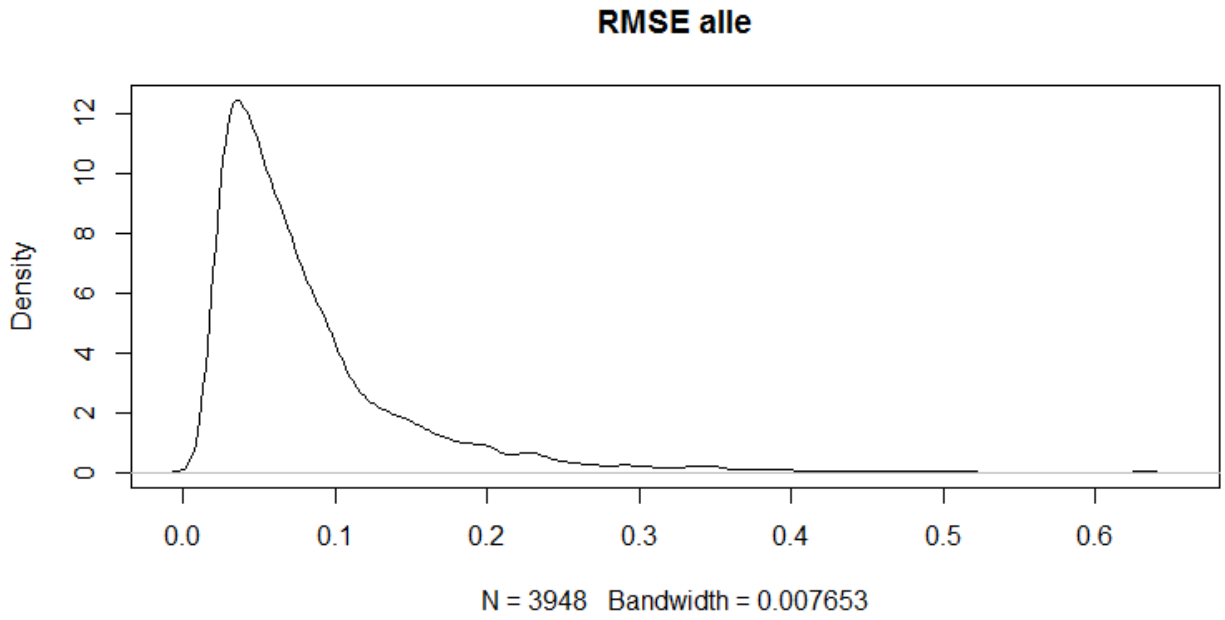
For analysen med alle bedrifter samlet er bs_cl og RI_num altid 1 og 2 i variable importance. De mest betydningsfulde variable på tværs af alle bootstrapsamples er listet neden for. Parantesen indikerer andelen af gange hvor variablen er i top 10 af variable importance ud af de 1000 bootstraps:

- Gaeldsprocent (997/100)

- WACC_GaeldAvg (994/1000)
- WACC_Gaeld (982/1000)
- WACC_GaeldL1 (971/1000)
- Egenkapital (949/1000)
- Soliditetsgrad (726/1000)
- LTVL1 (687/1000)
- LTV (555/1000)
- Rente_beregnet (121/1000)

Kvaliteten af prædiktionerne – alle samlet

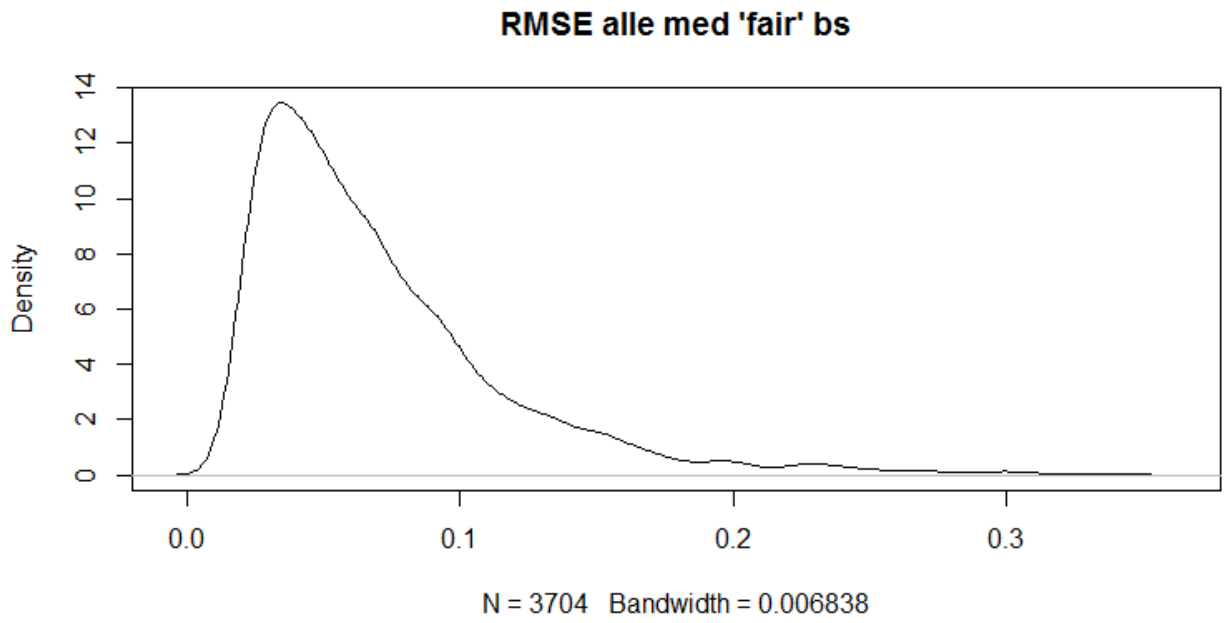
For at se på kvaliteten af prædiktionerne regnes en RMSE for alle bedrifter. De fordeler sig som i plottet nedenfor:



Det er tydeligt at se at hovedparten rammer inden for 0,1 % (over 75%), men også at der findes nogle virkelig store afvigelser. Omkring 25% af prædiktionerne har en RMSE på blot 0,04 %. Den gennemsnitlige RMSE er 0,08 %.

For at se på kvaliteten af prædiktionerne testes det også hvor tæt de gennemsnitlige prædiktioner ligger på de sande værdier. I gennemsnit rammer modellen 0,06 pp forkert. Over 50% af gennemsnitsprædiktionerne rammer indenfor 0,05 pp af den sande værdi.

I forhold til at skældne mellem en fair og en unfair bidragssats er det kun lige godt 6 % af observationerne der har en sandsynlighed over 50% for at have en unfair bidragssats. Holdes disse ude fås følgende RMSE plot som er sammenligneligt med det ovenfor:



Det ses en tydelig forbedring. Den gennemsnitlige RMSE er nu 0,07.

Dette viser også at modellen muligvis kan forbedres ved at holde de unfair behandlede bedrifter ude.

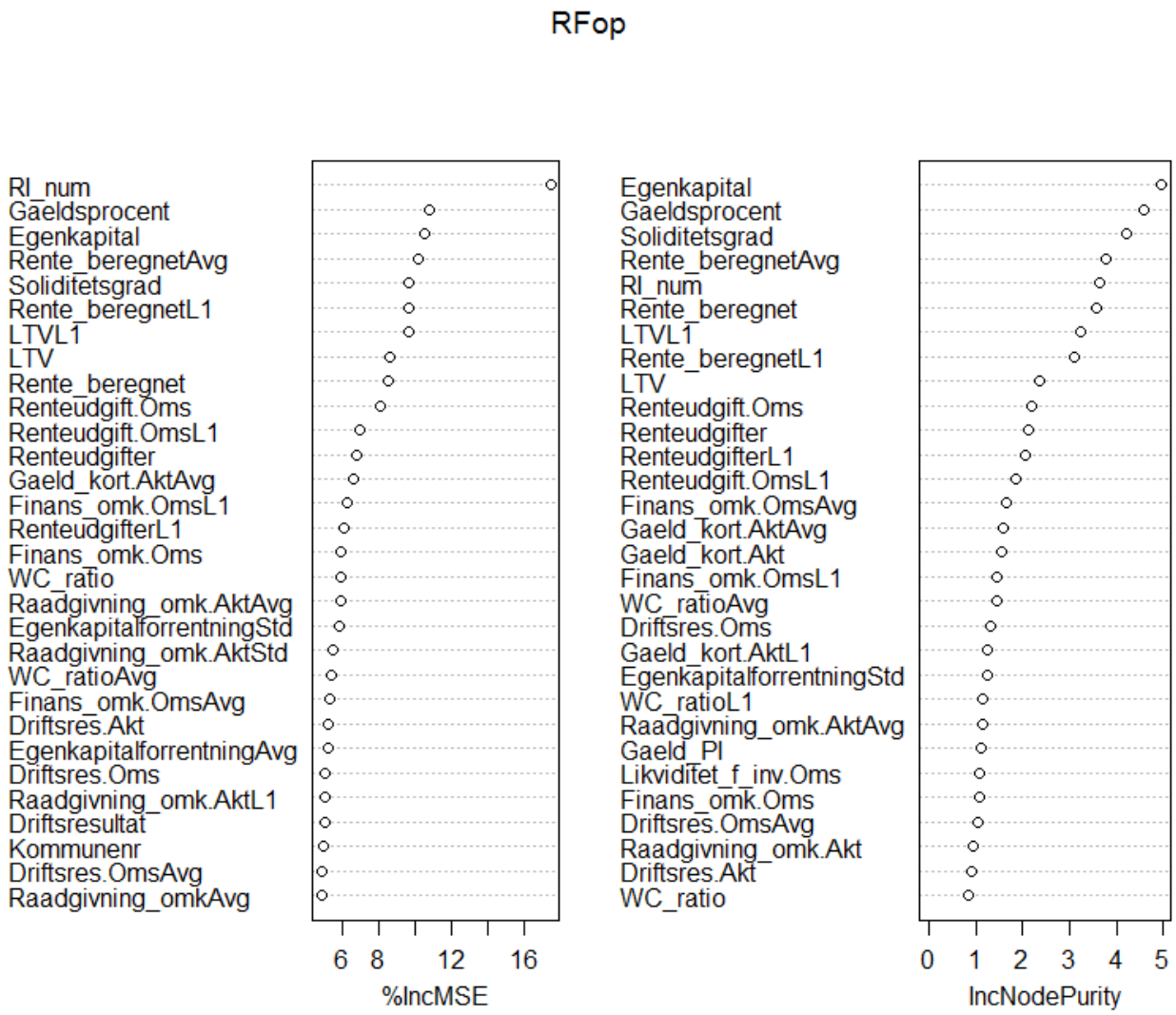
Ses der kun på bedrifter med Nykredit som institut er billedet meget lignende det overfor fremførte.

Den skrabede model

I dette afsnit køres modellen på alle bedrifterne (en enkelt kørsel) uden følgende variabler:

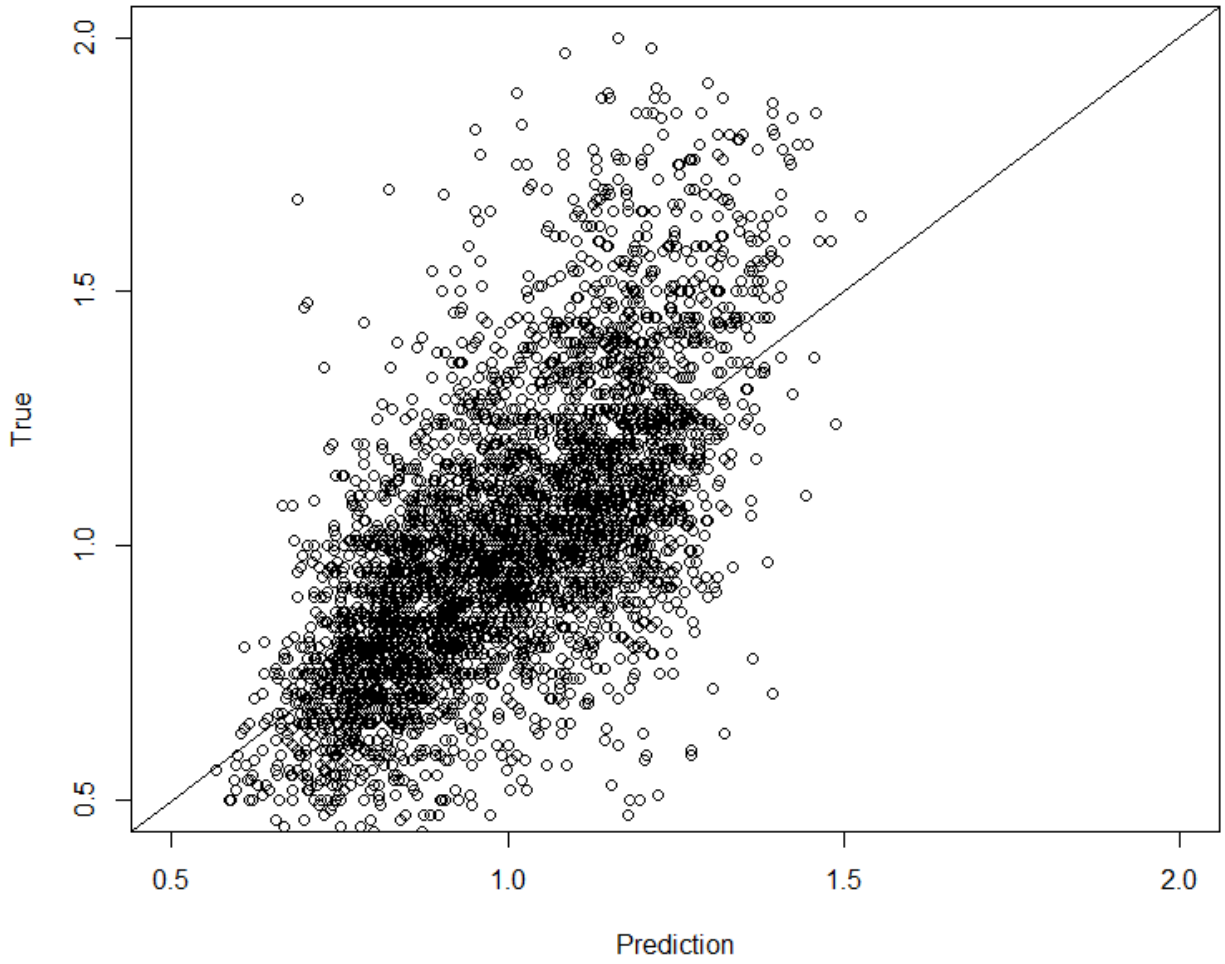
- Bs_cl
- Alle typer af WACC's

Med modellen kan "variable importance" beregnes. Ved brug af alle observationer samtidig opnås for en kørsel:

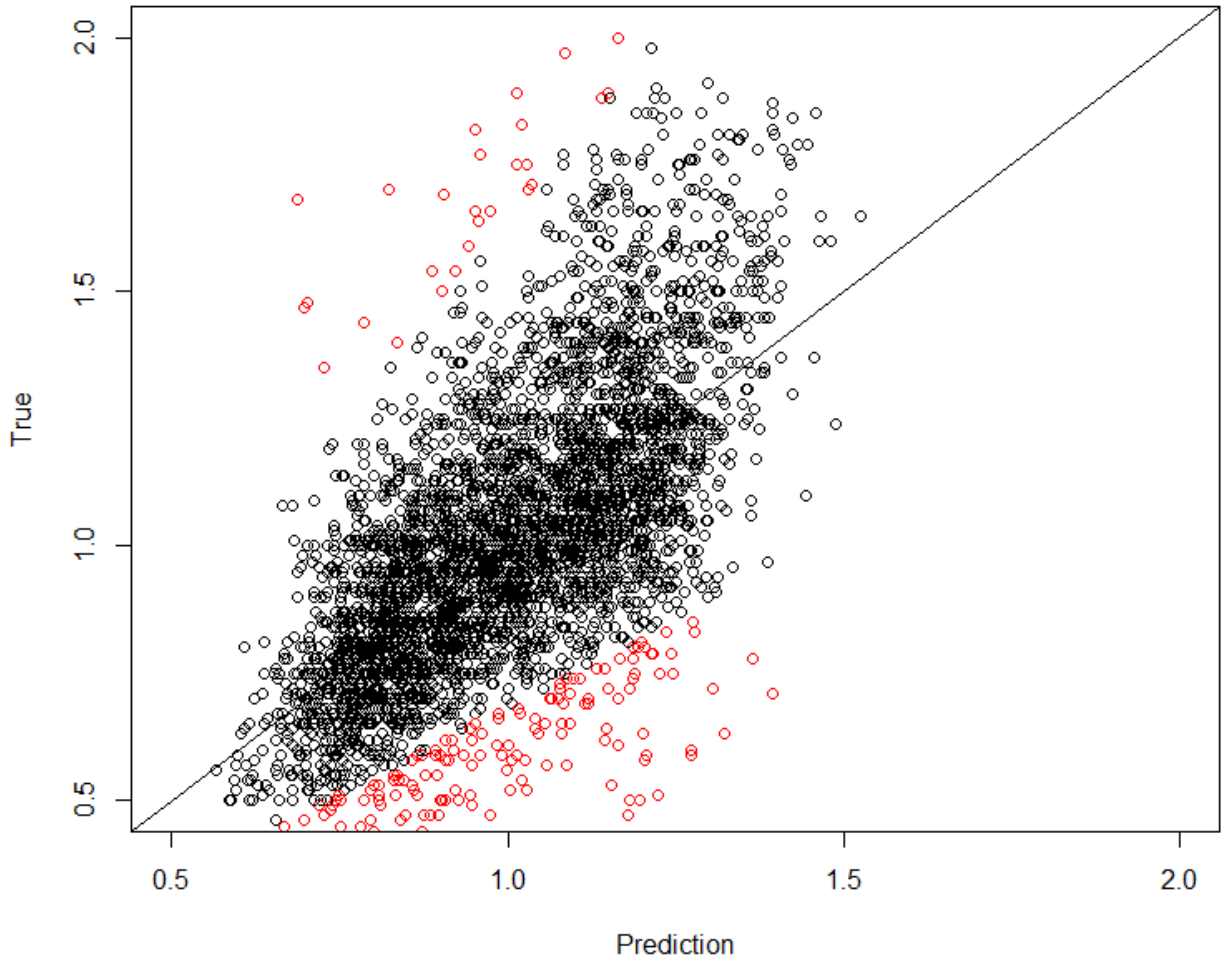


Igen er indikatorvariablen for realkreditinstitutterne med afstand den vigtigste. Efterfølgende kommer igen tal vedrørende balancen:

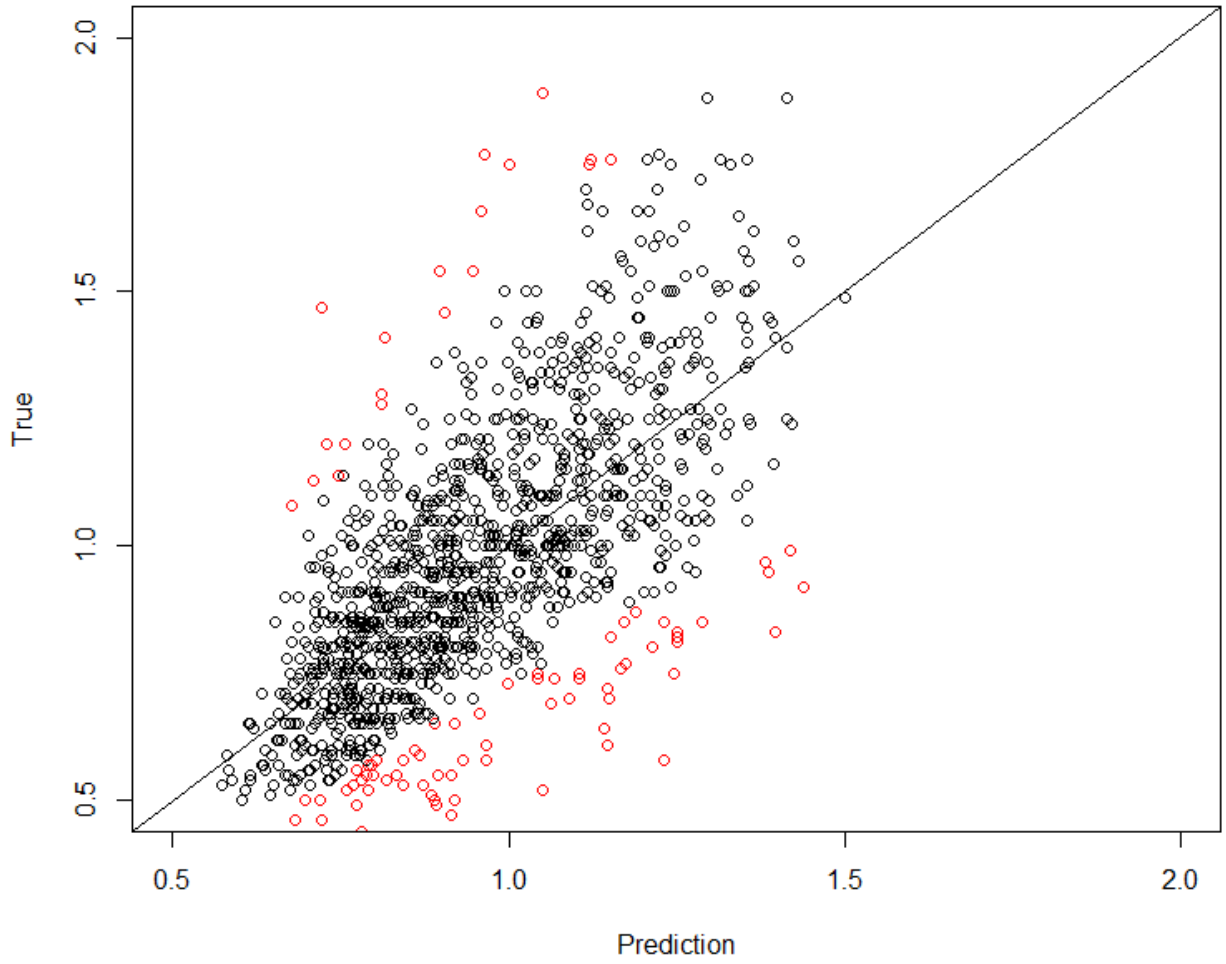
Det er nu muligt at holde den sande bidragssats op mod den prædikterede ved følgende plot:



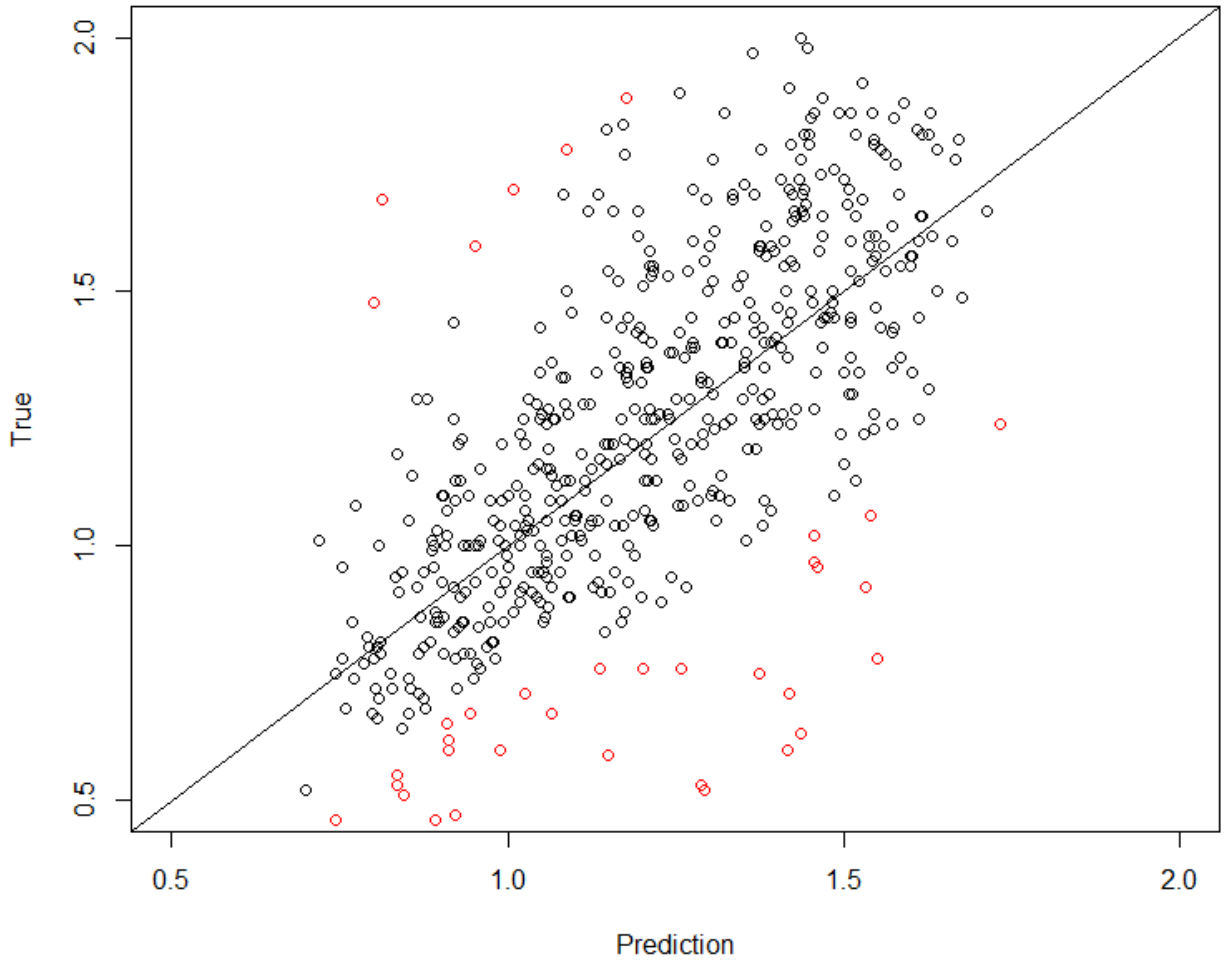
Som det fremgår af plottet er prædiktionsevnen blevet dårligere. Modellen kan forklare omkring 44 % af variansen og giver en root-mean-square-error (RMSE) på 0,20 % ved en kørsel.
Det antages nu at residualerne fra modellen er en blanding af to normalfordelinger. Dvs. en fordeling for de bedrifter der har fået en fair bidragssats (dem der kan forklares ud fra de økonomiske nøgletal) og en anden fordeling for dem der har fået en unfair bidragssats (dem modellen har svært ved at forklare). Det er nu muligt at identificere hvilke bedrifter der hører til hvilken fordeling og dermed sige om modellen indikerer at de har fået en fair eller unfair bidragssats.
I plottet nedenfor viser de røde prikker de bedrifter som ifølge modellen har fået en unfair bidragssats. Som det ses identificeres hovedparten af de bedrifter som modellen har meget svært ved at prædiktere.



For Nykredit ser billedet således ud:



For RD ser billedet således ud:


True
Prediction
0.5
1.0
1.5
2.0
0.5
1.0
1.5
2.0