DANMARKS NATIONALBANK

21. JANUAR 2019 — NR. 3

Tekstbaseret machine learning forbedrer konkursmodellering

Machine learning forbedrer beregning af konkurssandsynlighed

Machine learning giver mulighed for nye modelleringsmetoder og for at benytte ustrukturerede data. Det øger nøjagtigheden af de beregnede konkurssandsynligheder for virksomheder.

Tekstbaserede data tilføjer brugbar information

Inddragelse af ustrukturerede data i form af tekst fra virksomheders regnskaber øger informationsgrundlaget og forbedrer muligheden for at forudse konkurser.

Revisorpåtegninger har størst effekt

Det er særligt revisorpåtegninger, som indeholder brugbar information, når man beregner konkurssandsynligheder.

Evnen til at forudse virksomheders konkurser forbedres, hvis man ikke kun baserer sin vurdering på regnskabstal, men også på tekster fra regnskaberne. Det viser resultaterne i et working paper[1] offentliggjort i november 2018, hvor machine learning-metoder anvendes til at beregne konkurssandsynligheder på baggrund af regnskabstekster.

I papiret undersøges det, om to tekststykker (revisorpåtegninger og ledelsesberetninger), som findes i størstedelen af danske virksomheders regnskaber, kan bidrage med brugbar information om en virksomheds konkursrisiko. Det skal vel og mærke være information, som ikke allerede er givet i regnskabstallene.

I denne analyse gennemgås hovedresultaterne fra undersøgelsen. Det er især revisorpåtegninger, som bidrager med brugbar information, når det kommer til at forudsige en virksomheds konkursrisiko.

Formål med konkursmodel

Formålet med en konkursmodel er at beregne sandsynligheden for, at en virksomhed kommer i vanskeligheder og i sidste ende går konkurs. En sådan beregning kan have mange forskellige anvendelsesmuligheder. Set ud fra synspunktet om finansiel stabilitet kan det give et billede af, om nogle banker er mere eksponerede over for virksomheder, der er, eller er på vej, i økonomiske vanskeligheder.

En klassisk konkursmodel er typisk baseret på virksomheders regnskabstal samt nogle virksomhedskarakteristika som fx branche og alder. Ud over diverse regnskabstal indeholder virksomhedernes regnskaber også to tekststykker: revisorpåtegninger og ledelsesberetninger, der kan betegnes som ustrukturerede data.

I boks 1 ses et uddrag af en revisorpåtegning. Som det fremgår, har revisorerne i dette tilfælde en negativ holdning til virksomhedens fortsatte drift. Det kan indikere, at virksomheden har en forhøjet konkurssandsynlighed. Det er dog ikke sikkert, at informationen er fuldt afspejlet i regnskabstallene. På den måde er der potentiale for, at en konkursmodel kan forbedres ved at inddrage information fra den type tekststykker.

Eksempel på revisorpåtegning Boks 1

… "Det er vores vurdering, at der ikke er realistiske muligheder for at fremskaffe finansiering, og vi tager derfor forbehold for, at årsregnskabet er aflagt under forudsætning af fortsat drift. " …

Hvad er machine learning?

Machine learning er ikke et nyt fænomen, men har eksisteret siden 1940'erne, hvor de første teoretiske modeller blev udviklet. Det er dog først i de seneste år, med den hurtige udvikling i computerkraft og kraftigt øgede mængde af tilgængelige data, at potentialet og interessen for at anvende og udvikle machine learning rigtig har taget fart.

Machine learning adskiller sig fra mere konventionelle statistiske metoder, først og fremmest ved at gøre det muligt at modellere mere komplekse sammenhænge i data, som analytikeren ikke nødvendigvis kender på forhånd. Mens konventionelle metoder går ud på at "fitte" data til præspecificerede forhold mellem input- og outputvariable, tillader machine learning-metoder en "friere" tilgang til modelleringen af data, da de i højere grad lader data "tale", frem for at forsøge at tvinge data til at tilpasse sig en bestemt funktionel form. Derudover giver machine learning-metoder nye muligheder for at bruge såkaldt "ustrukturerede data", som fx billeder eller tekst.

Som en del af den tekstbaserede konkursmodel indgår der tre former for neurale netværk: et "convolutional neural network", et "recurrent neural network" og et klassisk neuralt netværk. En beskrivelse af, hvordan et neuralt netværk fungerer, er angivet i boks 2.

1 Hansen, Casper, Christian Hansen, Rastin Matin og Pia Mølgaard, Predicting distresses using deep learning of text segments in annual reports, Danmarks Nationalbank Working Paper, nr. 130, november 2018 (*link*).

Tekstanalyse

Her gennemgås machine learning-metoden, som estimerer konkurssandsynligheder på baggrund af tekst. Modelleringen består af følgende tre trin:

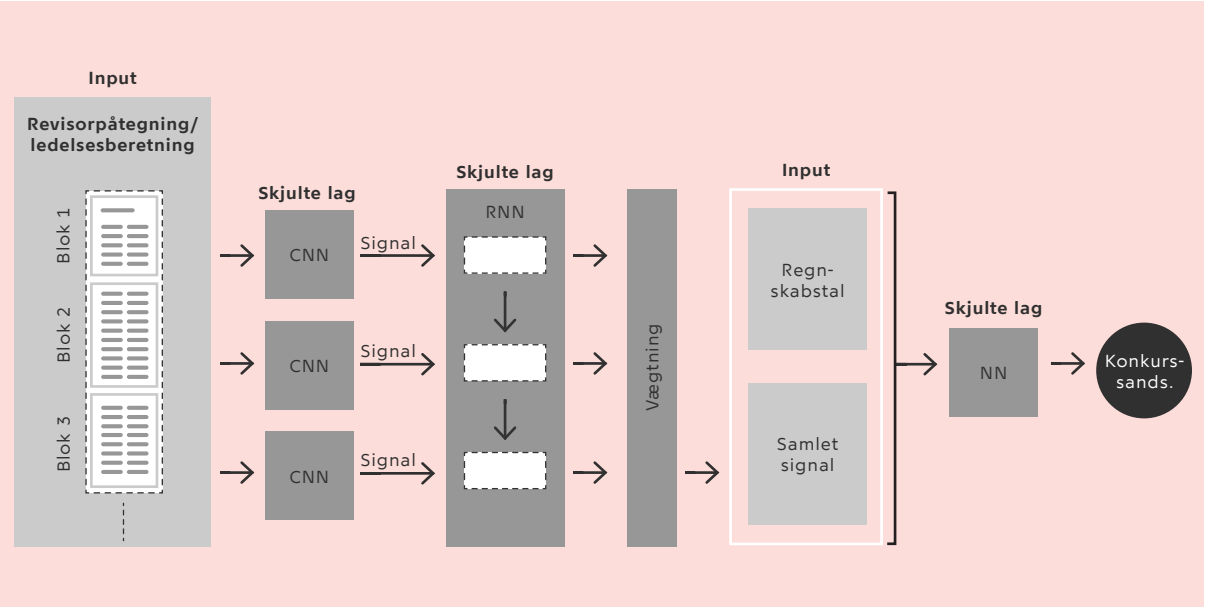
1. ordrepræsentation,
2. semantikken i ordsammensætningen og
3. inddragelse af regnskabstal samt beregning af konkurssandsynlighed.

I det følgende gennemgås disse tre trin kort. I dem alle indgår der en række valg af parametre, som løbende vil blive opdateret simultant, når modellen trænes. Det vil sige, at alle parametre er optimeret til netop vores problem, som er beregning af konkurssandsynlighed. Et diagram over den samlede model kan ses i figur 1.

Trin 1: Ordrepræsentation
Det første, vi skal gøre, er at danne en ordrepræsentation, som gøres ved såkaldt word2vec. Metoden går ud på at omdanne alle ord fra tekst til tal, som er lettere for modellen at arbejde med. Alle (unikke) ord omdannes til talrækker, som unikt identificerer det enkelte ord.

Som en del af ordrepræsentationen kommer ord, der "ligner" hinanden i betydning, også til at ligne hinanden i talrepræsentationen, sådan at der kan laves beregninger på ordene i forhold til deres semantiske betydning. For eksempel vil talrepræsentationen af følgende ordkombination *konge – mand + kvinde* ligge meget tæt op ad talrepræsentationen af ordet *dronning*.

Diagram af tekstanalysemodellen Figur 1



Anm.: I det første inputlag opdeles teksten i blokke, der køres igennem hver deres CNN-netværk, som spytter et signal ud. Signalet fodres til RNN-netværket, som tager højde for tekstblokkenes placering i forhold til hinanden. Signalerne herfra vægtes i forhold til hinanden, og det samlede signal sættes sammen med regnskabstal og køres igennem et klassisk neuralt netværk, som til sidst producerer en konkurssandsynlighed.

Trin 2: Semantikken i ordsammensætningen
Mens det første trin går ud på at forstå og repræsentere de enkelte ord i regnskabsteksten, handler dette trin om at forstå *sammenhængen* i ordene. Vi bevæger os altså væk fra simpel tekstanalyse, der blot forholder sig til de enkelte ord, hen imod at forstå betydningen af hele sætninger eller paragraffer i tekststykkerne.

Specifikt gøres det ved først at opdele revisorpåtegningerne og ledelsesberetningerne i mindre, overlappende tekstblokke illustreret ved inputlaget yderst til venstre i figur 1. Alle blokkene i laget udgør tilsammen én revisorpåtegning eller én ledelsesberetning. Ved hjælp af et såkaldt convolutional neural network, CNN, dannes der en lavdimensional repræsentation af semantikken af hele tekstblokken.

Signalerne fra de enkelte blokke samles herefter til et samlet semantisk signal, der skal repræsentere hele tekststykket. Det gøres i et såkaldt recurrent neural network, RNN, der sekventielt læser de enkelte signaler fra blokkene og samler dem til et endeligt signal. Således "læser" netværket hele teksten og danner sig et samlet signal, hvori rækkefølgen af de enkelte tekststykker kan have betydning. Hvis der fx indgår en negation før et positivt ord, forstår det neurale netværk, at betydningen af sætningen er negativ.

Trin 3: Inddragelse af regnskabstal samt beregning af konkurssandsynlighed
I dette trin kombineres ovenstående samlede signal fra teksten med de "konventionelle" regnskabsdata, og modellen færdigtrænes på både numeriske og tekstbaserede data i et neuralt netværk, NN.

Træning af modellen går ud på at fodre den med en masse eksempler på virksomheder, angivet med en kombination af deres regnskabstal og regnskabstekster, samt en indikator for, om de er gået konkurs eller ej. Baseret på alle disse "kendte" eksempler lærer modellen, hvilke kombinationer af regnskabstal og tekstsignaler en virksomhed under konkurs typisk vil være karakteriseret ved i forhold til en ikke-konkursramt virksomhed. Måden, den lærer dette på, er, groft sagt, ved at prøve sig frem. Baseret på de input den fodres med, angiver den et "gæt" på, om en virksomhed er under konkurs eller ej. Eftersom vi allerede kender svaret, kan modellen revurdere sit gæt i forhold til det rigtige svar ved at ændre på parametrene i netværket og undersøge, om revurderingen bringer den tættere på det rigtige svar. Det gøres mange gange, indtil modellen kommer så tæt på det rigtige svar som muligt.

Træningen af modellen går hele vejen ned igennem systemet. Det vil sige, at når modellen "prøver sig frem" med en ny løsning, vil den også ændre på signalerne, der dannes i trin 2.

Undgå at modellen "over-fitter"
Fordi machine learning-modeller indeholder store mængder data og mange forskellige parametre, er der stor risiko for, at sådanne modeller vil "over-fitte". Det vil sige, at modellen bliver rigtig god til at forudsige konkurssandsynligheden for de virksomheder, som er i træningssættet, men ikke nødvendigvis klarer sig godt på et nyt sæt af virksomheder. For at undgå over-fitting evalueres modellens forudsigelsesevne løbende på et valideringssæt. Træningen af modellen fortsætter kun, så længe den forbedrer sig både med hensyn til træningssættet og med hensyn til valideringssættet. Metoden kaldes krydsvalidering og er almindeligt brugt inden for machine learning.

Resultater og konklusion

Modelevaluering og sammenligningsgrundlag
I alt implementeres tre forskellige neurale netværk, der indeholder tekst: 1. en model, som indeholder både revisorpåtegninger og ledelsesberetninger, 2. en model, som kun indeholder revisorpåtegninger, og 3. en model, som kun indeholder ledelsesberetninger. Ydermere implementeres to modeller kun baseret på regnskabstal: 1. et neuralt netværk (uden tekst), som udnytter de gode egenskaber ved machine learning-modeller – dog kun på konventionelle regnskabstal, og 2. en logistisk regression, som er en konventionel statistisk metode til at forudsige konkurser. Disse to modeller skal bruges som sammenligningsgrundlag for at undersøge, om modellerne, der inddrager tekst, bliver bedre til at forudsige virksomhedernes konkurssandsynlighed.

For at sammenligne modellernes forudsigelsesevne bruges AUC (Area Under the receiver operating characteristics Curve). AUC måler sandsynligheden for, at en model placerer en højere risiko på en tilfældig virksomhed, som går konkurs i et givet år, end på en tilfældig virksomhed, som ikke gør. AUC siger altså først og fremmest noget om *rangeringen* af virk-

somheder i forhold til deres konkurssandsynlighed. AUC-scoren vil altid ligge imellem 0,5 og 1. Jo tættere scoren er på 1, jo bedre er modellen til at rangere virksomhederne efter deres konkursrisiko.

Resultater

I figur 2 vises AUC for de fem forskellige modeller. Der er flere interessante resultater. Sammenlignes AUC for den logistiske regression med det neurale netværk uden tekst, fremgår det, at forudsigelsesevnen forbedres væsentligt ved at implementere en machine learning-model frem for en konventionel statistisk model, på trods af at de to modeller benytter samme datagrundlag.

Herudover ser vi, at modellerne, som inddrager tekst, klarer sig endnu bedre end det neurale netværk, der kun inddrager regnskabstal. Forskellen i AUC mellem tekstmodellerne og det neurale netværk uden tekst er alle statistisk signifikante på et 1-pct.s niveau. Altså kan vi konkludere, at der er brugbar information i regnskabsteksterne, som kan udnyttes i en konkursmodel til at opnå en bedre forudsigelsesevne.

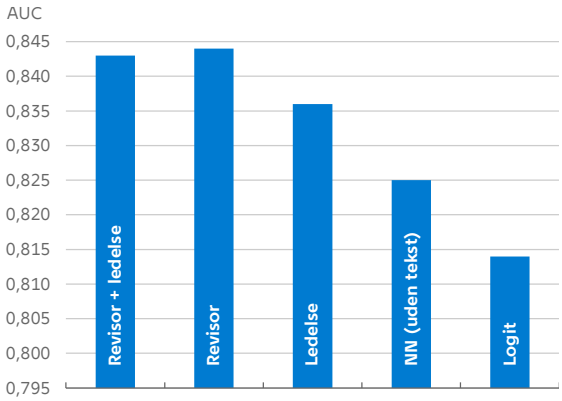
Det er ydermere interessant, at modellen med revisorpåtegninger klarer sig bedre end begge de to andre tekstbaserede modeller, dog ikke statistisk signifikant bedre end modellen, der indeholder både ledelsesberetninger og revisorpåtegninger. Det kan altså ud fra denne model konkluderes, at der ikke er nogen nyttig information at hente i ledelsesberetningerne, ud over det som allerede er at finde i revisorpåtegningerne. Det kan skyldes, at revisorpåtegningerne angiver en mere objektiv vurdering af virksomhedens status, hvorimod ledelsen kan have en tendens til kun at fokusere på, hvad der går godt i virksomheden.

Konklusion

Som det fremgår, er machine learning-metoder brugbare, når det kommer til at forudsige konkurser. Muligheden for at forudse konkurser forbedres især ved at inddrage tekstbaseret analyse af revisorpåtegninger.

Modellernes relative forudsigelsesevne

Figur 2

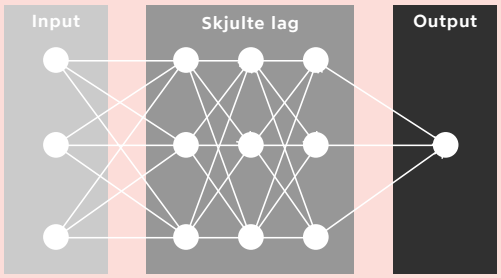


Anm.: AUC for de fem forskellige modeller. AUC er et mål for, hvor godt modellen rangerer virksomheder efter deres konkurssandsynlighed. En høj AUC-score betyder, at modellen har en bedre forudsigelsesevne.
Kilde: Bisnode og egne beregninger.

Neuralt netværk

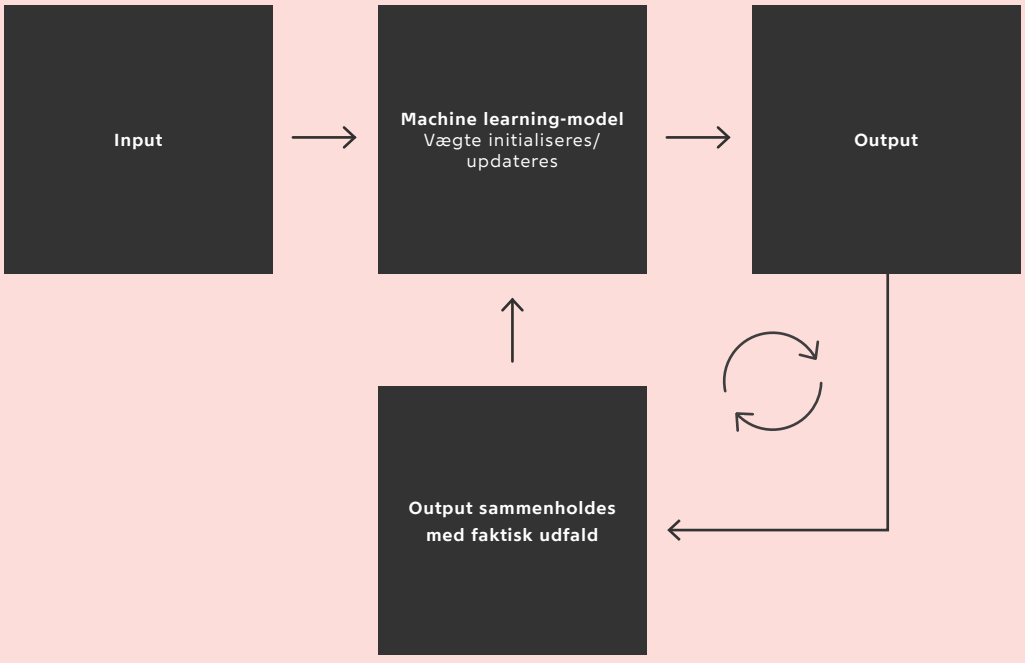
Boks 2

Et neuralt netværk kan minde om strukturen i den menneskelige hjerne, og specielt den måde hjernen behandler information og lærer på. Et neuralt netværk består af et inputlag, et eller flere såkaldt "skjulte lag" og et outputlag, jf. øverste figur. Hvert lag består af en række såkaldte "neuroner", som indeholder unik information. I billedanalyse repræsenterer hvert neuron i inputlaget fx en pixel i et billede. Informationen fra dette lag bliver vægtet og kombineret i det næste lag af neuroner. Sådan propagerer information gennem netværket og ender til sidst i outputlaget, hvor det endelige udfald angives, fx om billedet forstiller en hund eller ej.



Ved at præsentere et neuralt netværk for en masse eksempler på kendte udfald kan det trænes til at genkende mønstre og sammenhænge, der kan bruges til at forudsige nye, og på forhånd ukendte, udfald. Hvis man fx vil lære netværket at genkende en hund på et billede, præsenteres det for en masse forskellige billeder af hunde, en masse forskellige billeder af andre dyr og en indikator for, om hvert enkelt billede viser en hund eller ej.

Modellen starter så med at "gætte" på et udfald (fx hund/ikke hund) baseret på det givne input. Herefter evalueres, hvor tæt gættet er på det faktiske udfald, og vægtene i det neurale netværk opdateres tilsvarende, hvorefter modellen forsøger at gætte på et udfald igen. Processen gentages, indtil netværkets gæt, som gradvist bliver mere og mere kvalificeret, kommer så tæt på det faktiske udfald som muligt. Træningsprocessen er illustreret i nederste figur.



OM ANALYSE

Som en konsekvens af Nationalbankens rolle i samfundet udarbejdes analyser af økonomiske og finansielle forhold.

Analyserne udkommer løbende og omfatter bl.a. vurderinger af den aktuelle konjunktursituation og den finansielle stabilitet.

Analysen består af en dansk og engelsk version. I tilfælde af tvivl om oversættelsens korrekthed gælder den danske version.

DANMARKS NATIONALBANK
HAVNEGADE 5
1093 KØBENHAVN K
WWW.NATIONALBANKEN.DK

Redaktionen er afsluttet
18. januar 2019

Anna Kirstine Hvid
Quantitative Risk Analyst

Pia Mølgaard
Quantitative Risk Analyst

FINANSIEL STABILITET

DANMARKS NATIONALBANK