Case 9 - Corporate Bankruptcy Prediction using Accounting Ratios: Using LIME to Explain Black Box Predictions

Situation
Just as doctors check blood pressure and pulse rate as vital indicators of health of a patient, so business analysts scour the financial statements of a corporation to monitor its financial health. A primary issue in financial analysis, then is, how predictable is the health of a company? Using a variety of financial indicators that are available in this dataset of 132 firms (from the manufacturing and retail sectors), 66 of which filed for bankruptcy, your task is to develop a predictive model to identify the likelihood of a given company filing for bankruptcy.

Complication

Key Question
How accurately can we predict the likelihood of a given company filing for bankruptcy. Does it pay to use more fancy black box models such as neural networks? What if a regulator wants us to explain a given decision that was based on a prediction made from a neural network?

Data
CL-bankruptcy.xls

NO   Arbitrary ID number for each firm (we will not import this column)
D   D=0 for failed firms, D=1 for healthy firms. (our outcome variable)
YR   Year of Bankruptcy for failed firm in matched pair

The following financial ratios are available to us
R1   CASH/CURDEBT
R2   CASH/SALES
R3   CASH/ASSETS
R4   CASH/DEBTS
R5   CFF0/SALES
R6   CFFO/ASSETS
R7   CFFO/DEBTS
R8   COGS/INV
R9   CURASS/CURDEBT
R10   CURASS/SALES
R11   CURRASS/ASSETS
R12   CURDEBT/DEBTS
R13   INC/SALES
R14   INC/ASSETS

R15   INC/DEBTS
R16   UBCDEP/SALES
R17   INCDEP/ASSETS
R18   INCDEP/DEBTS
R19   SALES/REC
R20   SALES/ASSETS
R21   ASSETS/DEBTS
R22   WCFO/SALES
R23   WCFO/ASSETS
R24   WCFO/DEBTS

Code
Lets develop this in class using the R code of Case 8, the full blown cancer case study

Discussion points
1. It's best to treat the YR column as a categorical variable, ie a factor. However, this will require additional work if we want to use k-NN. I even found XGBoost to break down if I included the categorical variable YR.

2. Recall we had created 0/1 dummy variables corresponding to the levels of the categorical variables in the Donors case study. There we said...

We will do this by creating dummy variables for each of the levels of each of the categorical variables. For example the gender variables has 4 levels "female" "joint" "male" "UNK" . This will imply that we will have 4 dummy variables corresponding to these 4 levels

```{r }
library( dummies)

donors_x <- data.frame(donors_x) #convert from a tibble to a data frame
donors_x <- dummy.data.frame(data=donors_x, sep="_")

skim(donors_x)

```