191102exam

R Markdown

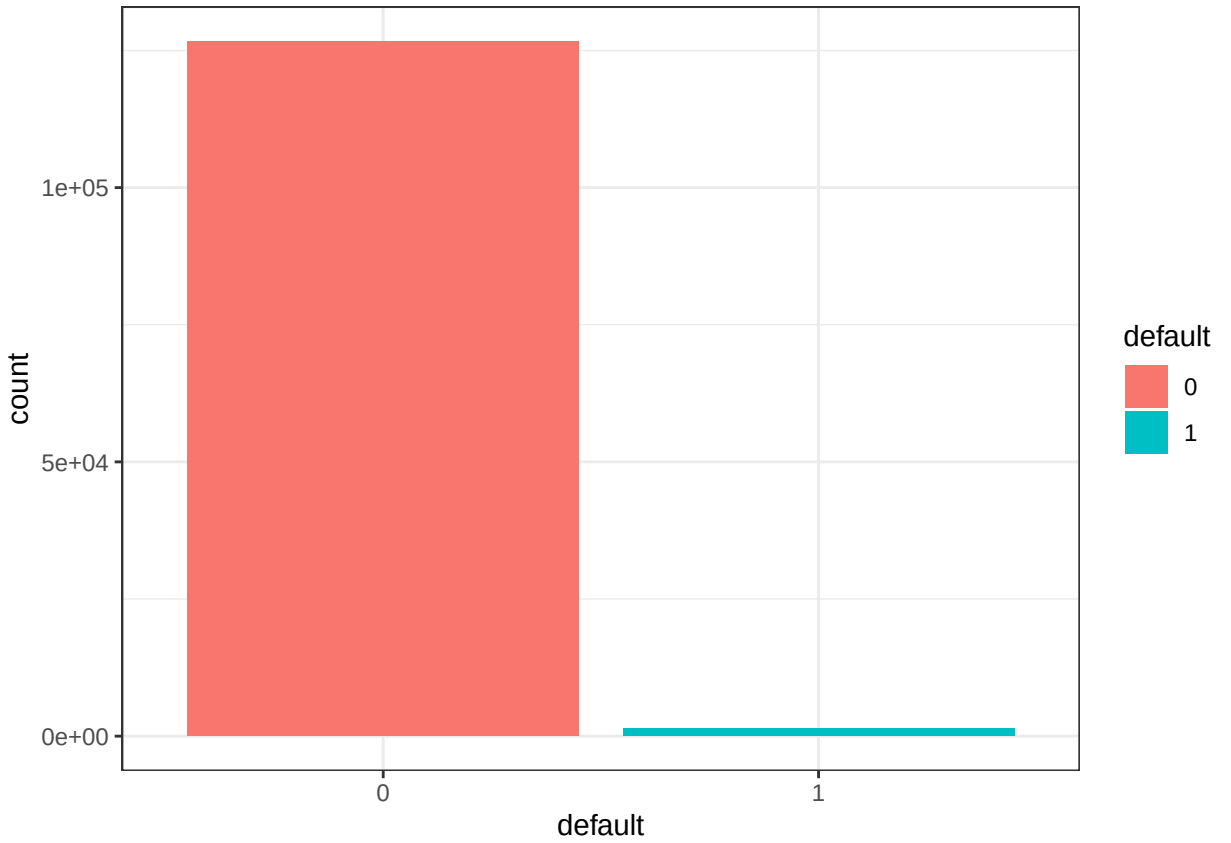
1. Setting up the data

An overview of the data shows us tha we have to change the structure of some of the variabels. Moreover, there are many extreme values in the data set and observations of defaulting companies are higly under-represented in the data set. We will deal with all these problems prior to implementing our prediction models.

Mia: make a nice summary table here

```
##     default        profit_margin        gross_operating_inc_perc
##  Min.   :0.00000   Min.   :-1.000e+19   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.: 0.000e+00   1st Qu.:0.3288
##  Median :0.00000   Median : 0.000e+00   Median :0.6035
##  Mean   :0.01094   Mean   :-1.010e+18   Mean   :0.5879
##  3rd Qu.:0.00000   3rd Qu.: 0.000e+00   3rd Qu.:0.9484
##  Max.   :1.00000   Max.   : 2.555e+04   Max.   :1.0000
##  operating_margin     EBITDA_margin        interest_coverage_ratio
##  Min.   :-1.000e+19   Min.   :-1.000e+19   Min.   :-1.000e+19
##  1st Qu.: 0.000e+00   1st Qu.: 0.000e+00   1st Qu.: 0.000e+00
##  Median : 0.000e+00   Median : 0.000e+00   Median : 4.000e+00
##  Mean   :-9.964e+17   Mean   :-9.935e+17   Mean   : 1.861e+17
##  3rd Qu.: 0.000e+00   3rd Qu.: 0.000e+00   3rd Qu.: 3.400e+01
##  Max.   : 5.860e+02   Max.   : 5.860e+02   Max.   : 1.000e+19
##   cost_of_debt       interest_bearing_debt revenue_stability
##  Min.   : -8131.72   Min.   :-1.000e+19    Min.   :-1.000e+19
##  1st Qu.:     0.28   1st Qu.: 0.000e+00    1st Qu.:-1.000e+00
##  Median :     1.99   Median : 0.000e+00    Median : 0.000e+00
##  Mean   :    13.10   Mean   :-7.574e+15    Mean   :-2.288e+18
##  3rd Qu.:     4.47   3rd Qu.: 1.000e+00    3rd Qu.: 0.000e+00
##  Max.   :165944.57   Max.   : 7.523e+03    Max.   : 4.473e+03
##   equity_ratio        equity_ratio_stability liquidity_ratio_1
##  Min.   :-1.000e+19   Min.   :-1.000e+19     Min.   :-2.474e+03
##  1st Qu.: 0.000e+00   1st Qu.:-1.000e+00     1st Qu.: 1.000e+00
##  Median : 0.000e+00   Median :-1.000e+00     Median : 1.000e+00
##  Mean   :-5.372e+16   Mean   :-1.422e+18     Mean   : 1.886e+17
##  3rd Qu.: 0.000e+00   3rd Qu.:-1.000e+00     3rd Qu.: 2.000e+00
##  Max.   : 1.690e+02   Max.   : 1.680e+02     Max.   : 1.000e+19
##  liquidity_ratio_2    liquidity_ratio_3        equity
##  Min.   :-2.474e+03   Min.   :-2.474e+03   Min.   :  -771200
##  1st Qu.: 1.000e+00   1st Qu.: 0.000e+00   1st Qu.:      114
##  Median : 1.000e+00   Median : 0.000e+00   Median :      483
##  Mean   : 1.886e+17   Mean   : 1.886e+17   Mean   :    32017
##  3rd Qu.: 2.000e+00   3rd Qu.: 1.000e+00   3rd Qu.:     2058
##  Max.   : 1.000e+19   Max.   : 1.000e+19   Max.   :182466000
##   total_assets          revenue          age_of_company
##  Min.   :    -9685   Min.   : -2883588   Min.   : 2.00
##  1st Qu.:      784   1st Qu.:      674   1st Qu.: 8.00
##  Median :     2450   Median :     3401   Median :15.00
```

```
##  Mean   :   122995   Mean   :    60941   Mean   :13.17
##  3rd Qu.:     7800   3rd Qu.:    11856   3rd Qu.:18.00
##  Max.   :544267000   Max.   :588422000   Max.   :22.00
##  unpaid_debt_collection paid_debt_collection adverse_audit_opinion
##  Min.   :-5.000e+00     Min.   :-6.000e+00   Min.   :0.000
##  1st Qu.: 0.000e+00     1st Qu.: 0.000e+00   1st Qu.:0.000
##  Median : 0.000e+00     Median : 0.000e+00   Median :0.000
##  Mean   : 8.534e+16     Mean   : 8.144e+16   Mean   :1.341
##  3rd Qu.: 0.000e+00     3rd Qu.: 0.000e+00   3rd Qu.:3.000
##  Max.   : 1.000e+19     Max.   : 1.000e+19   Max.   :6.000
##     industry      amount_unpaid_debt   payment_reminders
##  Min.   : 0.000   Min.   :-1.212e+05   Min.   :0.0000
##  1st Qu.: 0.000   1st Qu.: 0.000e+00   1st Qu.:0.0000
##  Median : 3.000   Median : 0.000e+00   Median :0.0000
##  Mean   : 3.576   Mean   : 7.675e+16   Mean   :0.5841
##  3rd Qu.: 7.000   3rd Qu.: 0.000e+00   3rd Qu.:1.0000
##  Max.   :11.000   Max.   : 1.000e+19   Max.   :3.0000

## 'data.frame':    128070 obs. of  24 variables:
##  $ default                 : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ profit_margin           : num  0.09037 0.06838 0.13838 0.14727 -0.00423 ...
##  $ gross_operating_inc_perc: num  0.631 0.611 0.55 0.617 0.544 ...
##  $ operating_margin        : num  0.0997 0.0873 0.1512 0.1542 0.0114 ...
##  $ EBITDA_margin           : num  0.1091 0.0877 0.1716 0.1802 0.0702 ...
##  $ interest_coverage_ratio : num  8.297 5.288 10.444 15.916 0.757 ...
##  $ cost_of_debt            : num  4.08 4.61 4.19 2.82 3.97 ...
##  $ interest_bearing_debt   : num  0.997 1.405 0.829 1.003 3.312 ...
##  $ revenue_stability       : num  -0.14 0.11 0.184 0.201 0.336 ...
##  $ equity_ratio            : num  0.557 0.369 0.285 0.233 0.142 ...
##  $ equity_ratio_stability  : num  -0.438 -0.623 -0.69 -0.746 -0.864 ...
##  $ liquidity_ratio_1       : num  3.55 1.93 2.14 2.36 2.13 ...
##  $ liquidity_ratio_2       : num  1.41 0.919 1.133 1.689 1.36 ...
##  $ liquidity_ratio_3       : num  0.2865 0.0304 0.2628 0.5415 0.2058 ...
##  $ equity                  : int  8548 6671 4957 3045 1499 1502 1257 1040 1821 1829 ...
##  $ total_assets            : int  15200 17714 15986 12011 11034 9581 8469 9309 3515 3915 ...
##  $ revenue                 : int  27691 32187 28994 24488 20383 15253 13689 16276 16725 19138 ...
##  $ age_of_company          : int  21 20 19 18 17 16 15 14 21 20 ...
##  $ unpaid_debt_collection  : num  3.61e-05 3.11e-05 0.00 0.00 0.00 ...
##  $ paid_debt_collection    : num  3.61e-05 0.00 3.45e-05 4.08e-05 4.91e-05 0.00 0.00 0.00 0.00 0.00
##  $ adverse_audit_opinion   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ industry                : int  3 3 3 3 3 3 3 3 7 7 ...
##  $ amount_unpaid_debt      : num  0 0.0103 0 0 0 ...
##  $ payment_reminders       : int  3 0 0 2 0 0 1 2 0 0 ...
```



Factor variables: ‘ " Then we recategorize som of the factor variables. Adverse audit is coded as a dummy, where 1 indicates that there has been an adverse audit opinion, and 0 indicates no adverse audit.

The tables below show the distribution of the companies along the factor variables, depending on wether they have defaulted or not.

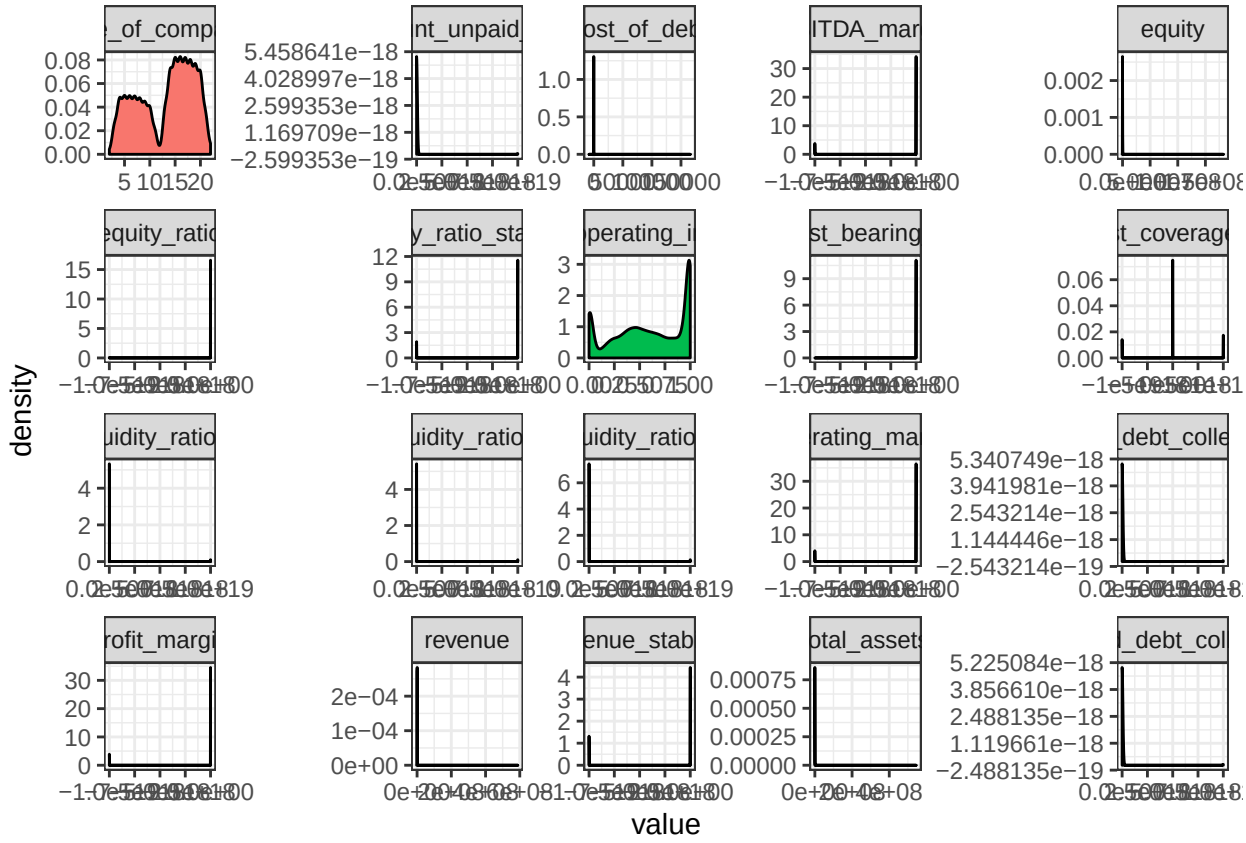
```
##
##         0     1     2     3     4     5     6
##   0 86937  1029    87 12435  3305 22340   536
##   1   293    22     1   147    59   825    54
```

```
##
##         0     1
##   0 86937 39732
##   1   293  1108
```

```
##
##         0     1     2     3     4     5     7     8     9    10    11
##   0 50541   916   842 12627 13255   639 37287  4391  3286  2279   606
##   1   369    10     7   178   171    13   501    82    48     7    15
```

```
##
##         0     1     2     3
##   0 80182 26677 13775  6035
##   1   328   229   292   552
```

Let's have a look of the distribution of the variables in the data set. The figure below contains density plots for all the numeric variables. Due to the presence of outliers, these figures do not provide much information.




```
##  default    profit_margin        gross_operating_inc_perc
##  0:126669   Min.   :-1.000e+19   Min.   :0.0000
##  1:  1401   1st Qu.: 0.000e+00   1st Qu.:0.3288
##             Median : 0.000e+00   Median :0.6035
##             Mean   :-1.010e+18   Mean   :0.5879
##             3rd Qu.: 0.000e+00   3rd Qu.:0.9484
##             Max.   : 2.555e+04   Max.   :1.0000
##
##  operating_margin     EBITDA_margin        interest_coverage_ratio
##  Min.   :-1.000e+19   Min.   :-1.000e+19   Min.   :-1.000e+19
##  1st Qu.: 0.000e+00   1st Qu.: 0.000e+00   1st Qu.: 0.000e+00
##  Median : 0.000e+00   Median : 0.000e+00   Median : 4.000e+00
##  Mean   :-9.964e+17   Mean   :-9.935e+17   Mean   : 1.861e+17
##  3rd Qu.: 0.000e+00   3rd Qu.: 0.000e+00   3rd Qu.: 3.400e+01
##  Max.   : 5.860e+02   Max.   : 5.860e+02   Max.   : 1.000e+19
##
##   cost_of_debt       interest_bearing_debt revenue_stability
##  Min.   : -8131.72   Min.   :-1.000e+19    Min.   :-1.000e+19
##  1st Qu.:     0.28   1st Qu.: 0.000e+00    1st Qu.:-1.000e+00
##  Median :     1.99   Median : 0.000e+00    Median : 0.000e+00
##  Mean   :    13.10   Mean   :-7.574e+15    Mean   :-2.288e+18
##  3rd Qu.:     4.47   3rd Qu.: 1.000e+00    3rd Qu.: 0.000e+00
##  Max.   :165944.57   Max.   : 7.523e+03    Max.   : 4.473e+03
```

```
##
##   equity_ratio        equity_ratio_stability liquidity_ratio_1
##  Min.   :-1.000e+19   Min.   :-1.000e+19     Min.   :-2.474e+03
##  1st Qu.: 0.000e+00   1st Qu.:-1.000e+00     1st Qu.: 1.000e+00
##  Median : 0.000e+00   Median :-1.000e+00     Median : 1.000e+00
##  Mean   :-5.372e+16   Mean   :-1.422e+18     Mean   : 1.886e+17
##  3rd Qu.: 0.000e+00   3rd Qu.:-1.000e+00     3rd Qu.: 2.000e+00
##  Max.   : 1.690e+02   Max.   : 1.680e+02     Max.   : 1.000e+19
##
##  liquidity_ratio_2    liquidity_ratio_3        equity
##  Min.   :-2.474e+03   Min.   :-2.474e+03   Min.   :  -771200
##  1st Qu.: 1.000e+00   1st Qu.: 0.000e+00   1st Qu.:      114
##  Median : 1.000e+00   Median : 0.000e+00   Median :      483
##  Mean   : 1.886e+17   Mean   : 1.886e+17   Mean   :    32017
##  3rd Qu.: 2.000e+00   3rd Qu.: 1.000e+00   3rd Qu.:     2058
##  Max.   : 1.000e+19   Max.   : 1.000e+19   Max.   :182466000
##
##   total_assets          revenue          age_of_company
##  Min.   :    -9685   Min.   : -2883588   Min.   : 2.00
##  1st Qu.:      784   1st Qu.:      674   1st Qu.: 8.00
##  Median :     2450   Median :     3401   Median :15.00
##  Mean   :   122995   Mean   :    60941   Mean   :13.17
##  3rd Qu.:     7800   3rd Qu.:    11856   3rd Qu.:18.00
##  Max.   :544267000   Max.   :588422000   Max.   :22.00
##
##  unpaid_debt_collection paid_debt_collection adverse_audit_opinion
##  Min.   :-5.000e+00     Min.   :-6.000e+00   0:87230
##  1st Qu.: 0.000e+00     1st Qu.: 0.000e+00   1:40840
##  Median : 0.000e+00     Median : 0.000e+00
##  Mean   : 8.534e+16     Mean   : 8.144e+16
##  3rd Qu.: 0.000e+00     3rd Qu.: 0.000e+00
##  Max.   : 1.000e+19     Max.   : 1.000e+19
##
##     industry     amount_unpaid_debt   payment_reminders
##  0      :50910   Min.   :-1.212e+05   0:80510
##  7      :37788   1st Qu.: 0.000e+00   1:26906
##  4      :13426   Median : 0.000e+00   2:14067
##  3      :12805   Mean   : 7.675e+16   3: 6587
##  8      : 4473   3rd Qu.: 0.000e+00
##  9      : 3334   Max.   : 1.000e+19
##  (Other): 5334
```

#Handling missing observations and outliers

We observe that the number xx appears throughout the data set, and assume that these are missing observations. In total, these extreme values account for xx percent of our observations.

We choose to replace these values with NA to begin with.

```
##      profit_margin gross_operating_inc_perc operating_margin EBITDA_margin
## [1,]         12929                       0            12761         12724
##      interest_coverage_ratio cost_of_debt interest_bearing_debt
## [1,]                   22224            0                   97
##      revenue_stability equity_ratio equity_ratio_stability
## [1,]             29303          688                  18208
```

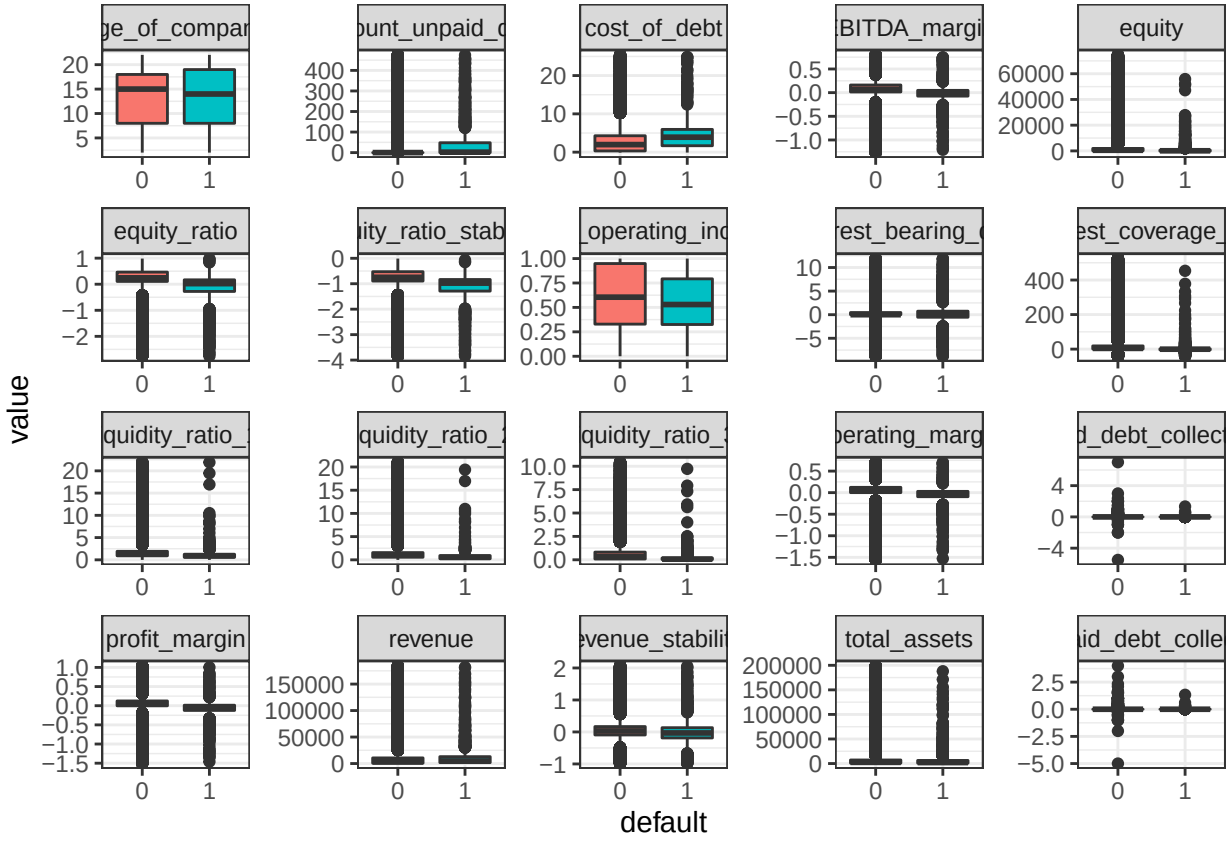
```
##      liquidity_ratio_1 liquidity_ratio_2 liquidity_ratio_3 equity
## [1,]              2416              2416              2416      0
##      total_assets revenue age_of_company unpaid_debt_collection
## [1,]            0       0              0                   1093
##      paid_debt_collection amount_unpaid_debt
## [1,]                 1043                983

##      profit_margin gross_operating_inc_perc operating_margin EBITDA_margin
## [1,]     0.1009526                        0       0.09964082    0.09935192
##      interest_coverage_ratio cost_of_debt interest_bearing_debt
## [1,]               0.1735301            0          0.0007573983
##      revenue_stability equity_ratio equity_ratio_stability
## [1,]         0.2288046  0.005372062              0.1421722
##      liquidity_ratio_1 liquidity_ratio_2 liquidity_ratio_3 equity
## [1,]        0.01886468        0.01886468        0.01886468      0
##      total_assets revenue age_of_company unpaid_debt_collection
## [1,]            0       0              0            0.008534395
##      paid_debt_collection amount_unpaid_debt
## [1,]          0.008143984         0.00767549

##  profit_margin        operating_margin    EBITDA_margin
##  Min.   :-25001.000   Min.   :-8391.346   Min.   :-7259.521
##  1st Qu.:    -0.007   1st Qu.:   -0.002   1st Qu.:    0.011
##  Median :     0.041   Median :    0.046   Median :    0.069
##  Mean   :     0.382   Mean   :   -0.641   Mean   :   -0.536
##  3rd Qu.:     0.130   3rd Qu.:    0.128   3rd Qu.:    0.167
##  Max.   : 25553.333   Max.   :  586.175   Max.   :  586.171
##  NA's   :12929        NA's   :12761       NA's   :12724
##  interest_coverage_ratio interest_bearing_debt revenue_stability
##  Min.   : -41488.8       Min.   :-72702.90     Min.   :-4775.143
##  1st Qu.:      0.4       1st Qu.:     0.00     1st Qu.:   -0.106
##  Median :      4.1       Median :     0.00     Median :    0.030
##  Mean   :     85.1       Mean   :    -1.83     Mean   :    0.682
##  3rd Qu.:     19.9       3rd Qu.:     0.56     3rd Qu.:    0.179
##  Max.   :1037284.8       Max.   :  7523.00     Max.   : 4473.242
##  NA's   :22224           NA's   :97            NA's   :29303
##   equity_ratio       equity_ratio_stability liquidity_ratio_1
##  Min.   :-1.43e+05   Min.   :-1.43e+05      Min.   :-2473.83
##  1st Qu.: 1.00e-01   1st Qu.:-9.00e-01      1st Qu.:    0.88
##  Median : 2.40e-01   Median :-7.50e-01      Median :    1.26
##  Mean   :-2.90e+00   Mean   :-4.15e+00      Mean   :    8.24
##  3rd Qu.: 4.80e-01   3rd Qu.:-5.10e-01      3rd Qu.:    1.96
##  Max.   : 1.69e+02   Max.   : 1.68e+02      Max.   :34126.68
##  NA's   :688         NA's   :18208          NA's   :2416
##  liquidity_ratio_2  liquidity_ratio_3   unpaid_debt_collection
##  Min.   :-2473.79   Min.   :-2473.809   Min.   :-5e+00
##  1st Qu.:    0.53   1st Qu.:    0.072   1st Qu.: 0e+00
##  Median :    1.00   Median :    0.340   Median : 0e+00
##  Mean   :    7.90   Mean   :    3.378   Mean   : 8e-04
##  3rd Qu.:    1.60   3rd Qu.:    0.850   3rd Qu.: 0e+00
##  Max.   :34126.59   Max.   :16330.311   Max.   : 4e+00
##  NA's   :2416       NA's   :2416        NA's   :1093
##  paid_debt_collection amount_unpaid_debt
##  Min.   :-5.5000      Min.   : -121226
```

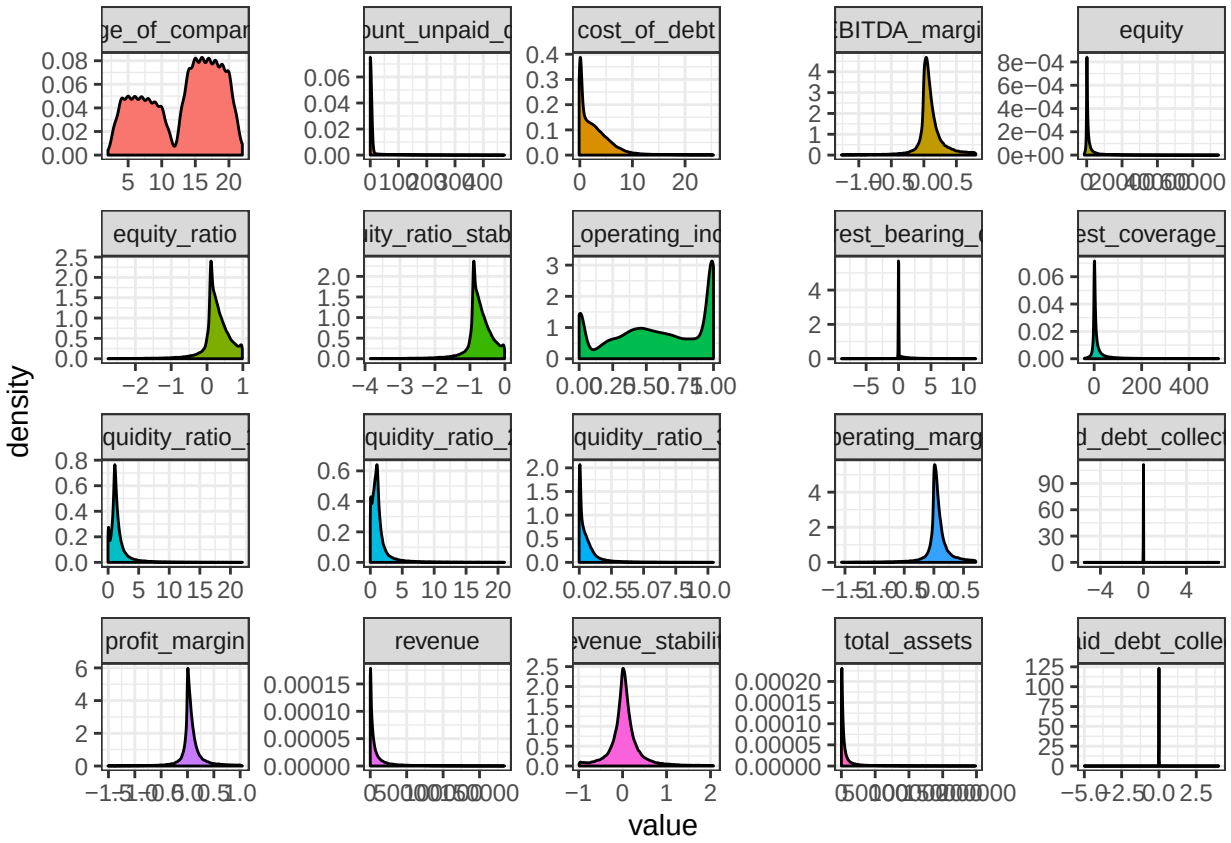
```
##  1st Qu.: 0.0000      1st Qu.:       0
##  Median : 0.0000      Median :       0
##  Mean   : 0.0009      Mean   :     957
##  3rd Qu.: 0.0000      3rd Qu.:       0
##  Max.   : 7.0000      Max.   :45000000
##  NA's   :1043         NA's   :983
```

The figures below show the distribution after replacing xxx with NA. As shown, we still have an issue with outliers.

(Show summary table here??)

We assume that many of these values are error measurements. Applying a treshold of 2,5 percent at each end of the variables' distribution, we replace values exceeding this treshold with NAs.





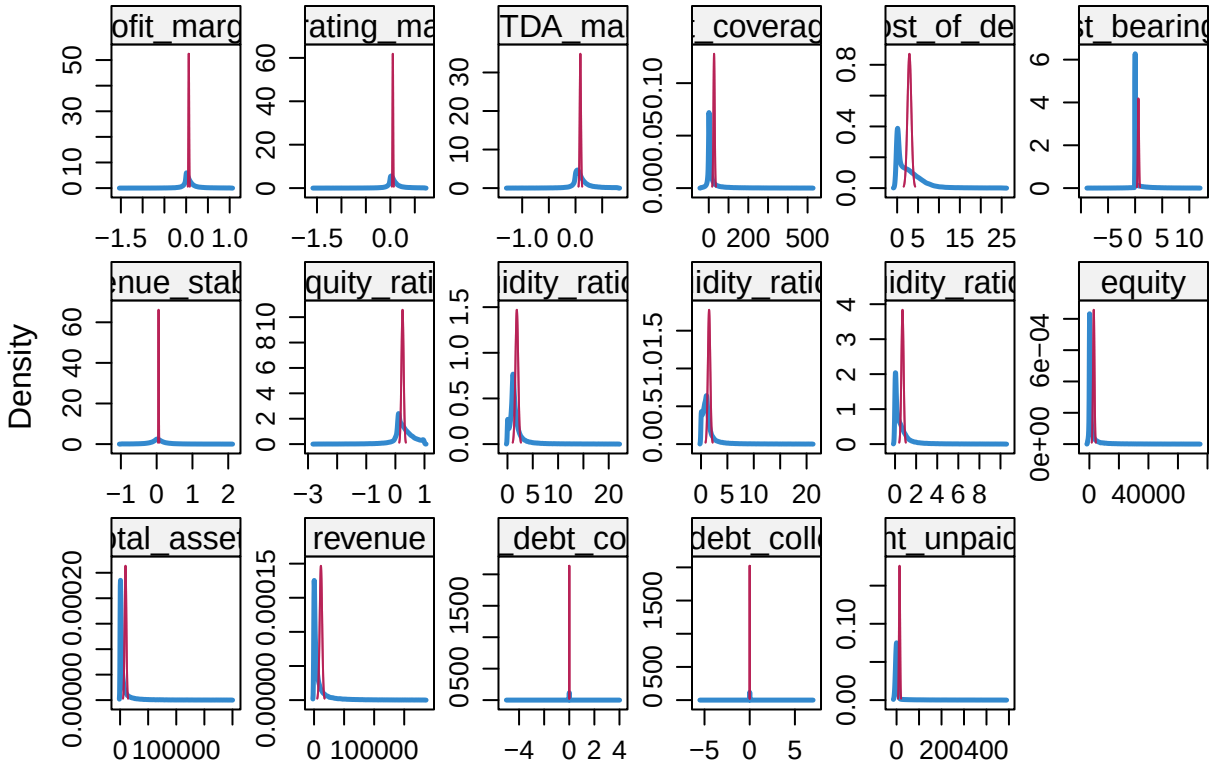
Mia: Might mention that equity ratio stability seems to have exactly same distribution as equity ratio. We test for correlation etc etc and end up removing this variable moving forward. Saves some computation time for r when imputing.

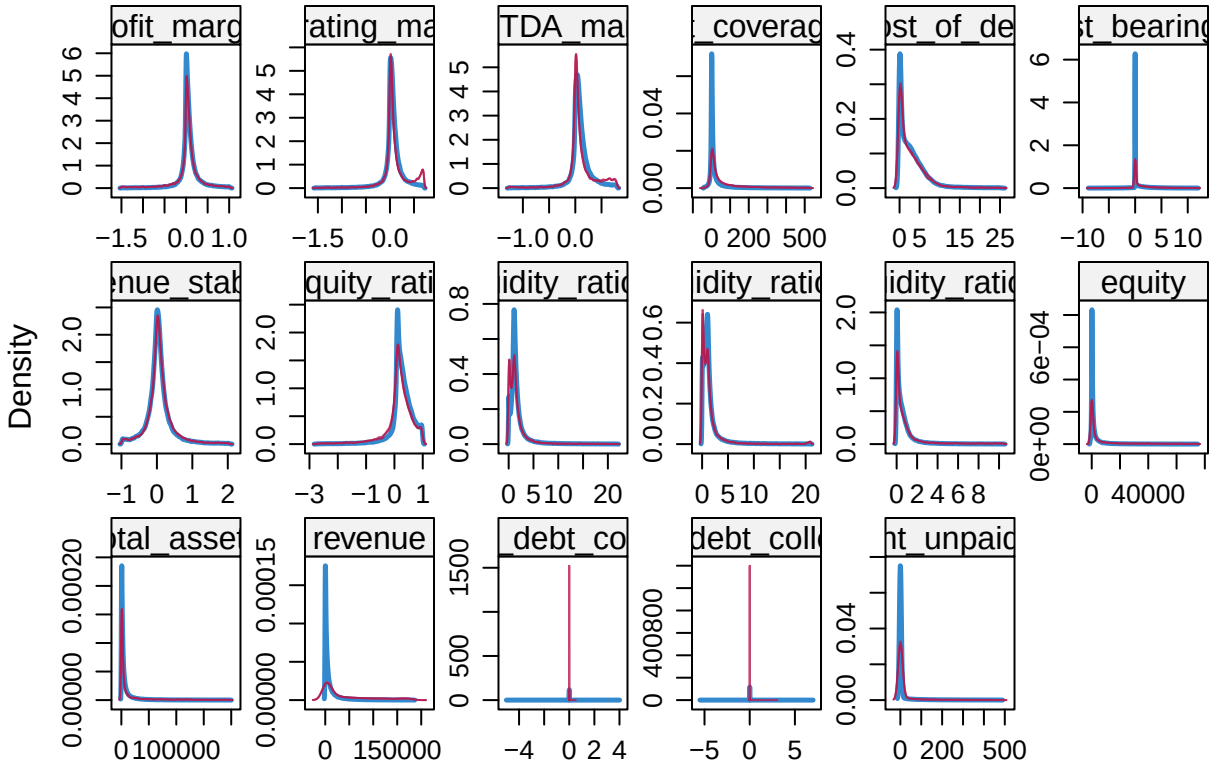
Imputation

As we have as much as xxx NA's, we choose not to delete these values, but rather impute them using the MICE package.

The tables below show how the distributions change when we apply different imputation methods. (Blue line: original data, ???)

```
##
##  iter imp variable
##    1   1  profit_margin  operating_margin  EBITDA_margin  interest_coverage_ratio  cost_of_debt  inte
##    1   2  profit_margin  operating_margin  EBITDA_margin  interest_coverage_ratio  cost_of_debt  inte
##    2   1  profit_margin  operating_margin  EBITDA_margin  interest_coverage_ratio  cost_of_debt  inte
##    2   2  profit_margin  operating_margin  EBITDA_margin  interest_coverage_ratio  cost_of_debt  inte
##    3   1  profit_margin  operating_margin  EBITDA_margin  interest_coverage_ratio  cost_of_debt  inte
##    3   2  profit_margin  operating_margin  EBITDA_margin  interest_coverage_ratio  cost_of_debt  inte
```


Density
ofit_marg
ating_ma
TDA_ma
_coverag
ost_of_de
t_bearing
enue_stab
quity_rati
idity_ratio
idity_ratio
idity_ratio
equity
tal_asset
revenue
_debt_co
debt_coll
nt_unpaid




```
## [1] 128070     23
```

```
## [1] 128070     23
```

As shown in the distribtion plots, there is not much variation in the variables measuring paid and unpaid debt collection. We generate two new dummy variables that provide two binary measures of paid and unpaid debt. Moreover,

The table below shows how this variable is distributed. As shown, defaulting firms are more frequently represented among those with debt collection. Moreover, firms who have reported paying down previous debt are less frequently represented among defaulters.

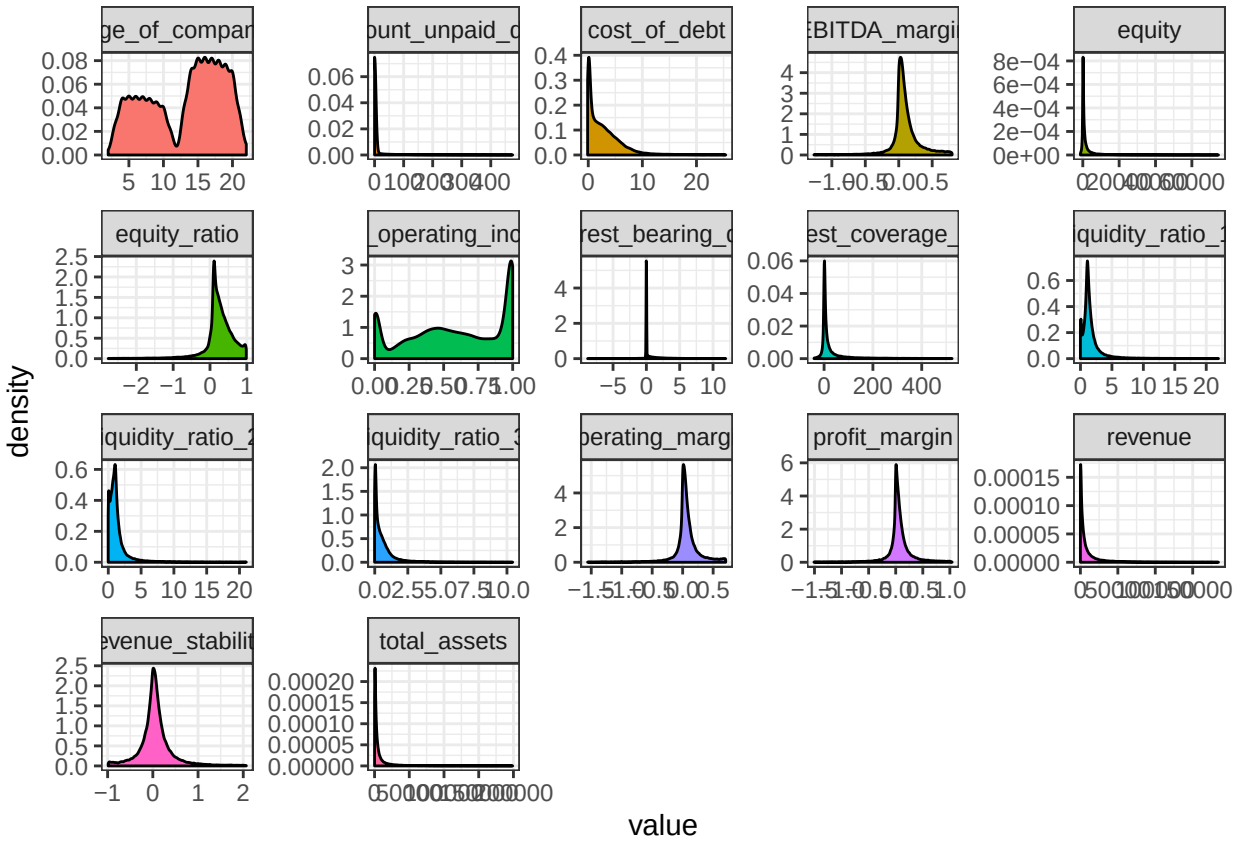
Mia: Might need to work on the reasoning behind generating this dummy a bit more. Could we do without it?

After cleaning, imputing and restructuring our data, our data set is better suited for prediciton modelling.

```
##  default    profit_margin       gross_operating_inc_perc
##  0:126669   Min.   :-1.509086   Min.   :0.0000
##  1:  1401   1st Qu.:-0.004403   1st Qu.:0.3288
##             Median : 0.040623   Median :0.6035
##             Mean   : 0.055478   Mean   :0.5879
##             3rd Qu.: 0.122814   3rd Qu.:0.9484
##             Max.   : 1.040878   Max.   :1.0000
##
##  operating_margin     EBITDA_margin      interest_coverage_ratio
```

```
##  Min.   :-1.5596854   Min.   :-1.26732   Min.   :-41.3653
##  1st Qu.:-0.0000975   1st Qu.: 0.01069   1st Qu.:  0.9261
##  Median : 0.0451302   Median : 0.06502   Median :  5.5595
##  Mean   : 0.0564319   Mean   : 0.09088   Mean   : 35.1157
##  3rd Qu.: 0.1244529   3rd Qu.: 0.16000   3rd Qu.: 26.4953
##  Max.   : 0.7063307   Max.   : 0.79729   Max.   :524.2675
##
##   cost_of_debt     interest_bearing_debt revenue_stability
##  Min.   :-0.1196   Min.   :-8.8111       Min.   :-0.98755
##  1st Qu.: 0.3223   1st Qu.: 0.0000       1st Qu.:-0.09211
##  Median : 1.9742   Median : 0.0000       Median : 0.03403
##  Mean   : 2.9294   Mean   : 0.6325       Mean   : 0.06078
##  3rd Qu.: 4.3025   3rd Qu.: 0.4977       3rd Qu.: 0.17477
##  Max.   :25.4004   Max.   :11.9011       Max.   : 2.06691
##
##   equity_ratio     liquidity_ratio_1  liquidity_ratio_2
##  Min.   :-2.7640   Min.   :-0.01474   Min.   :-0.01378
##  1st Qu.: 0.1050   1st Qu.: 0.88491   1st Qu.: 0.53882
##  Median : 0.2396   Median : 1.24957   Median : 0.99912
##  Mean   : 0.2409   Mean   : 1.83454   Mean   : 1.52981
##  3rd Qu.: 0.4598   3rd Qu.: 1.89696   3rd Qu.: 1.54770
##  Max.   : 0.9852   Max.   :21.94590   Max.   :21.03410
##
##  liquidity_ratio_3      equity       total_assets       revenue
##  Min.   :-0.01214   Min.   :-1326   Min.   :    39   Min.   :     0
##  1st Qu.: 0.07894   1st Qu.:  120   1st Qu.:   824   1st Qu.:   643
##  Median : 0.34118   Median :  490   Median :  2435   Median :  3287
##  Mean   : 0.72943   Mean   : 3223   Mean   : 10079   Mean   : 12190
##  3rd Qu.: 0.82296   3rd Qu.: 1970   3rd Qu.:  7354   3rd Qu.: 11059
##  Max.   :10.43398   Max.   :74642   Max.   :198917   Max.   :184415
##
##  age_of_company  unpaid_debt_collection paid_debt_collection
##  Min.   : 2.00   0:97918                0:110092
##  1st Qu.: 8.00   1:30152                1: 17978
##  Median :15.00
##  Mean   :13.17
##  3rd Qu.:18.00
##  Max.   :22.00
##
##  adverse_audit_opinion    industry     amount_unpaid_debt
##  0:87230               0      :50910   Min.   :  0.00
##  1:40840               7      :37788   1st Qu.:  0.00
##                        4      :13426   Median :  0.00
##                        3      :12805   Mean   : 13.53
##                        8      : 4473   3rd Qu.:  0.00
##                        9      : 3334   Max.   :475.40
##                        (Other): 5334
##  payment_reminders
##  0:80510
##  1:26906
##  2:14067
##  3: 6587
##
##
```

##



#Modelling preparations A few more steps before we are ready to start modelling:

We split the data frame into a training and test set. The variables total_assets, revenue, industry and paid_debt_collection are removed as they correlate with other independent variables.

```
## [1] TRUE
```

```
##          1
## 0.01094255
```

```
##          1
## 0.01093181
```

```
## profit_margin with operating_margin
## profit_margin with EBITDA_margin
## operating_margin with gross_operating_inc_perc
## liquidity_ratio_1 with gross_operating_inc_perc
## equity with gross_operating_inc_perc
```

Dealing with data imbalance

As mentioned, defaulting firms are highly underrepresentated in the data set. We deal with this by implementing an undersamlpling technique with the MICE package.

Model 1: GLM

set.seed(1)

model_glm <- train(default ~., data = train_data, method = “glm”, trControl = ctrl)

plot(varImp(model_glm))

glm_pred <- data.frame(actual = test_data$default, predict(model_glm, newdata = test_data, type = “prob”))

glm_pred$predict < -ifelse(glm_pred$X1 > 0.5, 1, 0) glm_pred$predict < -as.factor(glm_pred$predict)

cm_glm <-confusionMatrix(glm_pred$predict, test_data$default) cm_glm

summary(model_glm)

ROC curve glm

library(pROC)

result.predicted.prob <- predict(model_glm, test_data, type=“prob”) # Prediction

result.roc <- roc(test_data$default, result.predicted.prob$1) # Draw ROC

plot(result.roc, print.thres=“best”, print.thres.best.method=“closest.topleft”)

result.coords <- coords(result.roc, “best”, best.method=“closest.topleft”, ret=c(“threshold”, “accuracy”)) print(result.coords)#to get threshold and accuracy

Model 2: Random Forest

set.seed(1)

model_rf <- caret::train(default ~ ., data = train_data, method = “rf”, preProcess = c(“scale”, “center”), trControl = ctrl)

saveRDS(model_rf, file = “rf.Rdata”)

model_rf <- readRDS(“rf.Rdata”)

plot(varImp(model_rf))

rf_pred <- data.frame(actual = test_data$default, predict(model_rf, newdata = test_data, type = ”prob”))rf_pred$predict <- ifelse(rf_pred$X1 > 0.5, 1, 0)rf_pred$predict <- as.factor(rf_pred$predict)

cm_rf <- confusionMatrix(rf_pred$predict, test_data$default) cm_rf

ROC curve rf

result.predicted.prob <- predict(model_rf, test_data, type=“prob”) # Prediction

result.roc <- roc(test_data$default, result.predicted.prob$1) # Draw ROC

plot(result.roc, print.thres=“best”, print.thres.best.method=“closest.topleft”)

result.coords <- coords(result.roc, “best”, best.method=“closest.topleft”, ret=c(“threshold”, “accuracy”)) print(result.coords)#to get threshold and accuracy

Model 3: Xgboost

xgb_grid <- expand.grid(nrounds = 300, max_depth = 6, #3 min_child_weight = 1, subsample = 1, gamma = 0, colsample_bytree = 0.8, eta = .4)

set.seed(1)

model_xgb <- caret::train(default ~ ., data = train_data, method = "xgbTree", tuneGrid =xgb_grid, preProcess = c("scale", "center"), trControl = ctrl)

saveRDS(model_xgb, file = "xgb.Rdata")

model_xgb <- readRDS("xgb.Rdata")

model_xgb

plot(varImp(model_xgb))

xgb_pred <- data.frame(actual = test_data$default, predict(model_xgb, newdata = test_data, type = "prob"))

rf_pred$predict < -ifelse(xgb_pred$X1 > 0.5, 1, 0) xgb_pred$predict < -as.factor(xgb_pred$predict)

cm_xgb <- confusionMatrix(xgb_pred$predict, test_data$default) cm_xgb

ROC curve xgb

result.predicted.prob <- predict(model_xgb, test_data, type="prob") # Prediction

result.roc <- roc(test_data$default, result.predicted.prob`1`) # Draw ROC curve.

plot(result.roc, print.thres="best", print.thres.best.method="closest.topleft")

result.coords <- coords(result.roc, "best", best.method="closest.topleft", ret=c("threshold", "accuracy")) print(result.coords)#to get threshold and accuracy

Look at at difference all together

Look at the performance

models <- list(glm = model_glm, rf = model_rf, xgb = model_xgb)

resampling <- resamples(models)

bwplot(resampling)

density plots of accuracy

scales <- list(x=list(relation="free"), y=list(relation="free")) densityplot(resampling, scales=scales, pch = "|", allow.multiple = TRUE)

Other snacks for comparing

splom(resampling)

xyplot(resampling, models=c(“rf”, “xgb”))

summary(resampling)

HALLLAAA