German Credit data
Random Forest Model

Lecture/Practical 16

04/09/2021

RandomForest Model

- library(randomForest)
- library(tree)
- treemodel<-tree (gcredit$Credit.Rating ~Duration.z+Credit.Amount.z+Install_rate + Present.Resident+Age1.z++Num_Credits +gcredit$Balance.in.Savings.A.C+Other.installment,data= gcredit)
- summary(treemodel)

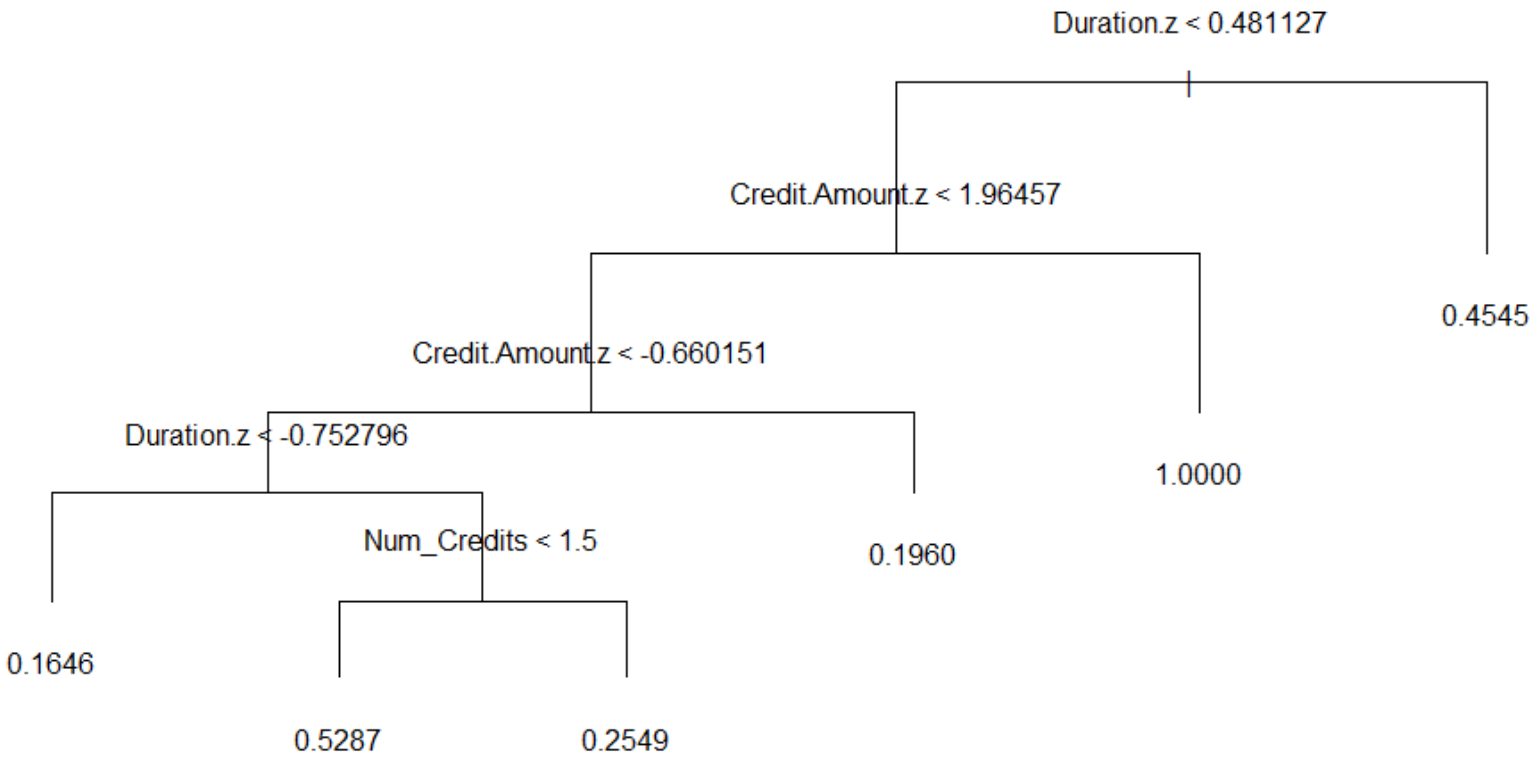
RandomForest Tree summary

```
Variables actually used in tree construction:
[1] "Duration.z"      "Credit.Amount.z" "Num_Credits"
Number of terminal nodes:  6
Residual mean deviance:  0.1871 = 148.6 / 794
Distribution of residuals:
    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-0.5287 -0.1960 -0.1960  0.0000  0.4713  0.8354
> |
```

RandomForest Tree

- plot (treemodel)
- text(treemodel,pretty=0)

RandomForest Tree



Predict Response Variable Value using Random Forest

- gcredit$predicted<-predict(treemodel, data=gcredit)
- gcredit$predicted

```
> gcredit$predicted
  [1] 0.1645570 0.4545455 0.1959799 0.4545455 0.1959799
  [6] 0.4545455 0.1959799 0.4545455 0.1959799 0.4545455
 [11] 0.5287356 0.4545455 0.1959799 0.2549020 0.1959799
 [16] 0.5287356 0.1959799 0.4545455 1.0000000 0.1959799
 [21] 0.1959799 0.1959799 0.1959799 0.1959799 0.1959799
 [26] 0.1645570 0.1645570 0.2549020 0.1959799 0.4545455
 [31] 0.1959799 0.1959799 0.1959799 0.5287356 0.1959799
 [36] 0.4545455 0.4545455 0.1959799 0.1645570 0.1645570
 [41] 0.4545455 0.5287356 0.1959799 0.4545455 0.4545455
 [46] 0.1959799 0.4545455 0.1645570 0.1959799 0.1959799
 [51] 0.1959799 0.4545455 0.5287356 0.1959799 0.4545455
 [56] 0.1645570 0.1959799 0.4545455 0.1959799 0.4545455
 [61] 0.1959799 0.1959799 0.4545455 0.4545455 0.1959799
 [66] 0.4545455 0.1959799 0.5287356 0.4545455 0.4545455
 [71] 0.4545455 0.1645570 0.1645570 0.4545455 0.4545455
```

Prediction

- p1<-predict(treemodel,gcredit)
- prediction1<-ifelse(p1>0.5, 1,0)
- prediction1 # only first 80 observations

```
> prediction1
  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16
  0   0   0   0   0   0   0   0   0   0   1   0   0   0   0   1
 17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32
  0   0   1   0   0   0   0   0   0   0   0   0   0   0   0   0
 33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48
  0   1   0   0   0   0   0   0   0   1   0   0   0   0   0   0
 49  50  51  52  53  54  55  56  57  58  59  60  61  62  63  64
  0   0   0   0   1   0   0   0   0   0   0   0   0   0   0   0
 65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80
  0   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0
```

- head(prediction1)

RandomForestModel

- RandomForestModel<-randomForest(gcredit$Credit.Rating ~Duration.z+Credit.Amount.z+Install_rate + Present.Resident+Age1.z++Num_Credits +gcredit$Balance.in.Savings.A.C+Other.installment,data= gcredit)

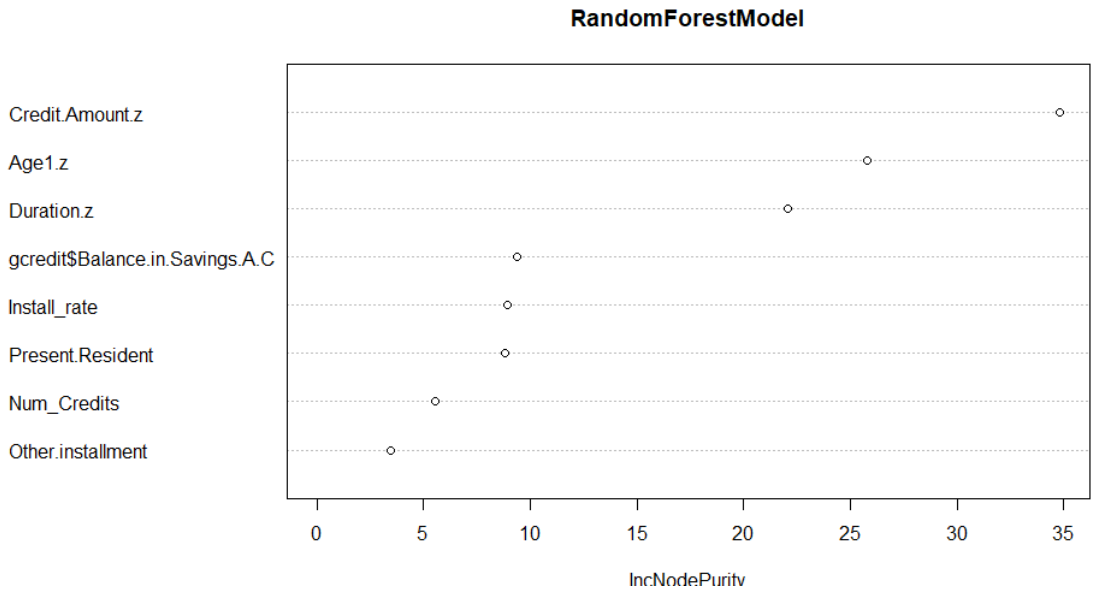
- print(RandomForestModel)

```
No. of variables tried at each split: 2

          Mean of squared residuals: 0.1920612
                    % Var explained: 8.32
```

Importance Plot

- varImpPlot(RandomForestModel)



- If you want to remove any variable as part of model building consider it from the bottom variables.

#Percentage of variation explained importance(RandomForestModel)

```
> importance(RandomForestModel)
                                IncNodePurity
Duration.z                          22.057793
Credit.Amount.z                     34.833609
Install_rate                         8.917006
Present.Resident                     8.792767
Age1.z                              25.801783
Num_Credits                          5.536122
gcredit$Balance.in.Savings.A.C       9.369118
Other.installment                    3.490870
```

- The most important variable is Credit amount followed by Age and Duration.