Tugas 02 Solution

Roni Yunis

3/8/2021

Import Data

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(readxl)
credit <- read_excel("data/Credit Risk Data.xlsx", sheet = "Base Data")
```

Menampilkan struktur data

```
str(credit)
```

```
## tibble [425 x 12] (S3: tbl_df/tbl/data.frame)
##  $ LoanPurpose   : chr [1:425] "Small Appliance" "Furniture" "New Car" "Furniture" ...
##  $ Checking      : num [1:425] 0 0 0 638 963 ...
##  $ Savings       : num [1:425] 739 1230 389 347 4754 ...
##  $ MonthsCustomer: num [1:425] 13 25 19 13 40 11 13 14 37 25 ...
##  $ MonthsEmployed: num [1:425] 12 0 119 14 45 13 16 2 9 4 ...
##  $ Gender        : chr [1:425] "M" "M" "M" "M" ...
##  $ MaritalStatus : chr [1:425] "Single" "Divorced" "Single" "Single" ...
##  $ Age           : num [1:425] 23 32 38 36 31 25 26 27 25 43 ...
##  $ Housing       : chr [1:425] "Own" "Own" "Own" "Own" ...
##  $ Years         : num [1:425] 3 1 4 2 3 1 3 1 2 1 ...
##  $ Job           : chr [1:425] "Unskilled" "Skilled" "Management" "Unskilled" ...
##  $ CreditRisk    : chr [1:425] "Low" "High" "High" "High" ...
```

Penjelasan variabel:

1. LoanPurpose type data char
2. Checking type data char
3. Savings type data char
4. MonthsCustomer type data integer
5. MonthsEmployed type data integer
6. Gender type data char
7. MaritalStatus type data char
8. Age type data integer
9. Housing type data char
10. Years type data integer
11. Job type data char
12. CreditRisk type data char

```
dim(credit)
```

```
## [1] 425  12
```

Data terdiri dari 425 baris dan 12 kolom

```
summary(credit )
```

```
##  LoanPurpose           Checking        Savings      MonthsCustomer
##  Length:425         Min.   :    0   Min.   :    0   Min.   : 5.0
##  Class :character   1st Qu.:    0   1st Qu.:  228   1st Qu.:13.0
##  Mode  :character   Median :    0   Median :  596   Median :19.0
##                     Mean   : 1048   Mean   : 1813   Mean   :22.9
##                     3rd Qu.:  560   3rd Qu.:  921   3rd Qu.:28.0
##                     Max.   :19812   Max.   :19811   Max.   :73.0
##  MonthsEmployed     Gender           MaritalStatus           Age
##  Min.   :  0.0   Length:425         Length:425         Min.   :18.0
##  1st Qu.:  6.0   Class :character   Class :character   1st Qu.:26.0
##  Median : 20.0   Mode  :character   Mode  :character   Median :32.0
##  Mean   : 31.9                                         Mean   :34.4
##  3rd Qu.: 47.0                                         3rd Qu.:41.0
##  Max.   :119.0                                         Max.   :73.0
##    Housing              Years          Job             CreditRisk
##  Length:425         Min.   :1.00   Length:425         Length:425
##  Class :character   1st Qu.:2.00   Class :character   Class :character
##  Mode  :character   Median :3.00   Mode  :character   Mode  :character
##                     Mean   :2.84
##                     3rd Qu.:4.00
##                     Max.   :4.00
```

```
#6 baris teratas
head(credit)
```

```
## # A tibble: 6 x 12
##   LoanPurpose Checking Savings MonthsCustomer MonthsEmployed Gender
##   <chr>          <dbl>   <dbl>          <dbl>          <dbl> <chr>
## 1 Small Appl~        0     739             13             12 M
```

```
## 2 Furniture          0    1230             25              0 M
## 3 New Car            0     389             19            119 M
## 4 Furniture        638     347             13             14 M
## 5 Education        963    4754             40             45 M
## 6 Furniture       2827       0             11             13 M
## # ... with 6 more variables: MaritalStatus <chr>, Age <dbl>, Housing <chr>,
## #   Years <dbl>, Job <chr>, CreditRisk <chr>
```

```
#6 baris terbawah
tail (credit)
```

```
## # A tibble: 6 x 12
##   LoanPurpose Checking Savings MonthsCustomer MonthsEmployed Gender
##   <chr>          <dbl>   <dbl>          <dbl>          <dbl> <chr>
## 1 New Car          193    2684             13              5 F
## 2 Small Appl~      497       0              7             51 M
## 3 Furniture          0       0             31             53 M
## 4 New Car            0       0             25            103 F
## 5 New Car            0     712             16              6 F
## 6 New Car            0     912              7             39 M
## # ... with 6 more variables: MaritalStatus <chr>, Age <dbl>, Housing <chr>,
## #   Years <dbl>, Job <chr>, CreditRisk <chr>
```

Exploratory Data Analysis

```
colSums(is.na(credit))
```

```
##    LoanPurpose       Checking        Savings MonthsCustomer MonthsEmployed
##              0              0              0              0              0
##         Gender  MaritalStatus            Age        Housing          Years
##              0              0              0              0              0
##            Job     CreditRisk
##              0              0
```

tidak ada data yang missing value

```
credit %>%
  count(LoanPurpose, name = "freq", sort = TRUE)
```

```
## # A tibble: 10 x 2
##    LoanPurpose      freq
##    <chr>           <int>
##  1 Small Appliance   105
##  2 New Car           104
##  3 Furniture          85
##  4 Business           44
##  5 Used Car           40
##  6 Education          23
##  7 Repairs            12
```

```
##   8 Other                 6
##   9 Large Appliance       4
## 10 Retraining            2
```

Tujuan kredit yang paling banyak adalah Small Appliance

```
credit %>%
  count(Gender, name = "freq", sort = TRUE)
```

```
## # A tibble: 2 x 2
##   Gender  freq
##   <chr>  <int>
## 1 M        290
## 2 F        135
```

Jenis kelamin yang paling banyak mengajukan pinjaman adalah Laki-laki

```
table(credit$CreditRisk)
```

```
##
## High  Low
##  211  214
```

Frekewnsi resiko kredit High = 211 dan Low = 21

```
prop.table(table(credit$CreditRisk))
```

```
##
##      High       Low
## 0.4964706 0.5035294
```

Proporsi jumlah resiko kredit High sebesar 49,6% dan Low sebesar 50,4%

```
prop.table(table(credit$CreditRisk, credit$Gender), margin = 2)
```

```
##
##                  F         M
##   High 0.5777778 0.4586207
##   Low  0.4222222 0.5413793
```

Perbandingan tingkat resikonya lebih tinggi perempuan dibandingkan laki-laki yaitu 57,7% : 45,9%

```
prop.table(table(credit$CreditRisk, credit$LoanPurpose), margin = 2)
```

```
##
##         Business Education Furniture Large Appliance   New Car     Other
##   High 0.5227273 0.6086957 0.5058824       0.7500000 0.6250000 0.6666667
##   Low  0.4772727 0.3913043 0.4941176       0.2500000 0.3750000 0.3333333
##
##          Repairs Retraining Small Appliance  Used Car
##   High 0.3333333  0.5000000       0.4000000 0.3000000
##   Low  0.6666667  0.5000000       0.6000000 0.7000000
```

Resiko paling tinggi adalah jenis pangajuan kredit *Large Appliance* yaitu sebesar 75%

Klasifikasi dengan Random Forest

Panggil library

```
#package untuk praktisi data
library(caret)
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
#package untuk klasifikasi
library(randomForest)
```

```
## randomForest 4.6-14
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
#package untuk mengukur perfomansi model klasifikasi
library(e1071)
#package untuk menguji kehandalan dari model prediksi
library(ROCit)
```

Bagi partisi data

```
set.seed(100) #pengambilan data secara random
#untuk data training diambil 70%, sisanya untuk data testing
index_train <- createDataPartition(credit$CreditRisk,
                                   p = 0.7,list = FALSE)

data.train <- credit[index_train,]
data.test <- credit[-index_train,]
```

Melihat hasil pembagian data

```
dim(data.train)
```

```
## [1] 298  12
```

```
dim(data.test)
```

```
## [1] 127  12
```

Model klasifikasi dengan Random Forest

```
set.seed(100) #menentukan nilai acak dari data
forestKu <- randomForest(data=data.train,
               as.factor(CreditRisk)~.,
               ntree=500)
```

```
forestKu
```

```
##
## Call:
##  randomForest(formula = as.factor(CreditRisk) ~ ., data = data.train,      ntree = 500)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 34.9%
## Confusion matrix:
##      High Low class.error
## High   94  54   0.3648649
## Low    50 100   0.3333333
```

Bisa dilihat bahwa error rate dari model adalah 34,9%, dengan akurasi sebesar 65,1% (0,65)

Importance

```
importance(forestKu)
```

```
##                MeanDecreaseGini
## LoanPurpose           13.339594
## Checking              13.362254
## Savings               22.900031
## MonthsCustomer        23.435885
## MonthsEmployed        22.615667
## Gender                 3.467050
## MaritalStatus          5.152297
## Age                   23.989816
## Housing                5.695325
## Years                  7.402706
## Job                    5.518078
```

Bisa dilihat bahwa variabel yang sangat penting yaitu variabel *age*

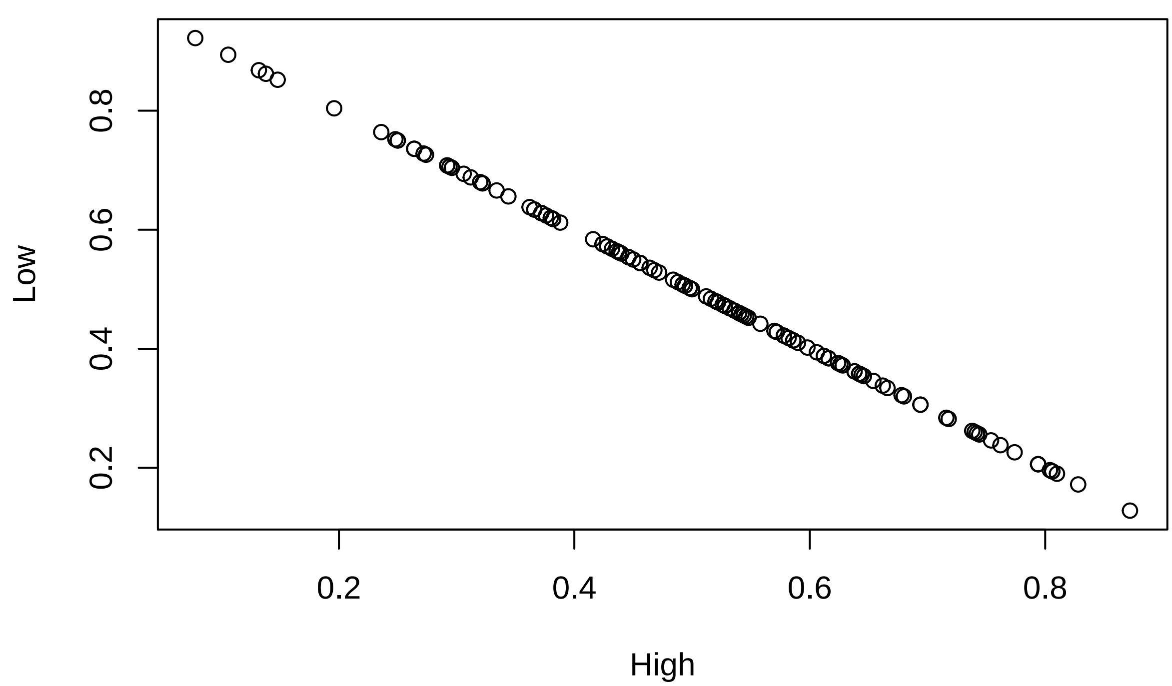
Mengukur kinerja prediksi

```r
hasilPrediksi <- predict(forestKu, data.test, type="prob")
head(hasilPrediksi, n=10)
```

```
##     High   Low
## 1  0.716 0.284
## 2  0.248 0.752
## 3  0.522 0.478
## 4  0.740 0.260
## 5  0.306 0.694
## 6  0.498 0.502
## 7  0.646 0.354
## 8  0.578 0.422
## 9  0.272 0.728
## 10 0.572 0.428
```

Menampilkan plot hasil prediksi

```r
plot(hasilPrediksi )
```



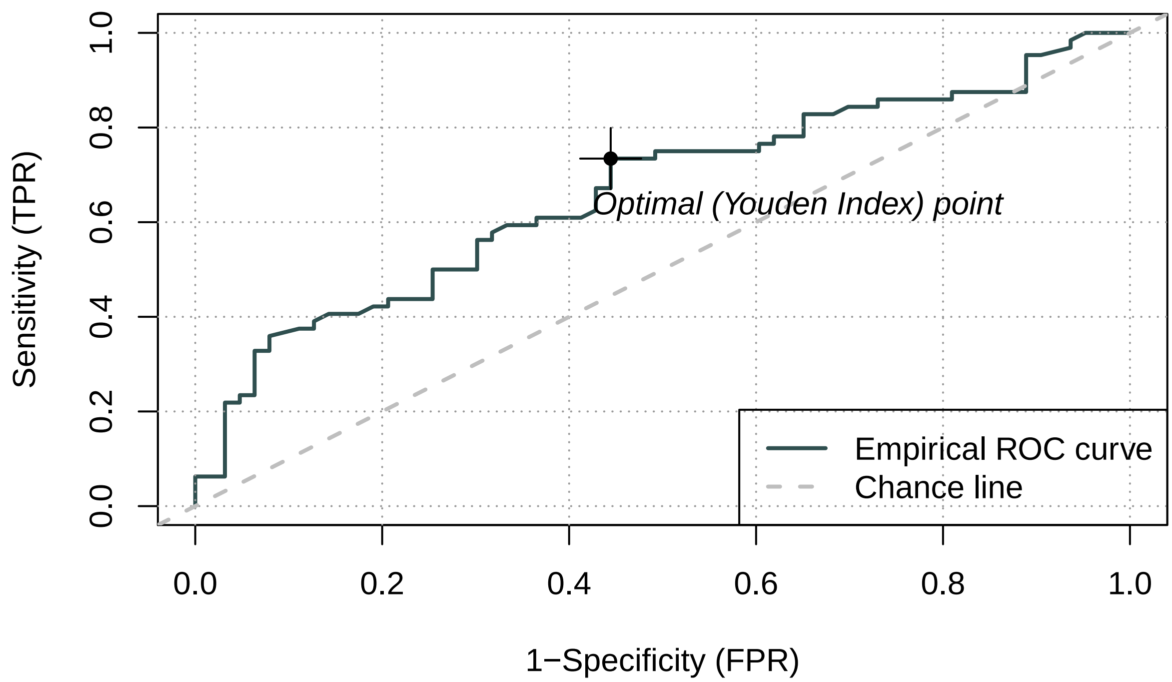
```
prediksi.status.f <- ifelse(hasilPrediksi[,2] > 0.5, "Low", "High")

#menghitung ukuran kinerja prediksi
confusionMatrix(as.factor(prediksi.status.f), as.factor(data.test$CreditRisk))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction High Low
##       High   40  26
##       Low    23  38
##
##                Accuracy : 0.6142
##                  95% CI : (0.5237, 0.6992)
##     No Information Rate : 0.5039
##     P-Value [Acc > NIR] : 0.008094
##
##                   Kappa : 0.2286
##
##  Mcnemar's Test P-Value : 0.775097
##
##             Sensitivity : 0.6349
##             Specificity : 0.5938
##          Pos Pred Value : 0.6061
##          Neg Pred Value : 0.6230
##              Prevalence : 0.4961
##          Detection Rate : 0.3150
##    Detection Prevalence : 0.5197
##       Balanced Accuracy : 0.6143
##
##        'Positive' Class : High
##
```

Berdasarkan hasil diatas bisa lihat bahwa nilai akurasi sebesar 61,4%, Sensitivity (High) 63,5% dan Specificity (Low) 59,3%

Hitung Nilai Performance dari Prediksi

```
ngitungROCf <- rocit(score=hasilPrediksi[,2],class=data.test$CreditRisk)
plot(ngitungROCf)
```




```
AUCf <- ngitungROCf$AUC
AUCf
```

```
## [1] 0.6639385
```

Nilai AUC nya adalah 0,66, jadi bisa disimpulkan bahwa klasifikasi yang dihasilkan termasuk pada *poor classification*

O P T I O N A L # Klasifikasi dengan Decision Tree ## Panggil package

```
library(rpart)
library(rpart.plot)
```

Model Klasifikasi dengan Decision Tree

```
pohonKu <- rpart(data=data.train,
                 CreditRisk~.,
                 control = rpart.control(cp=0, minsplit=100))
```

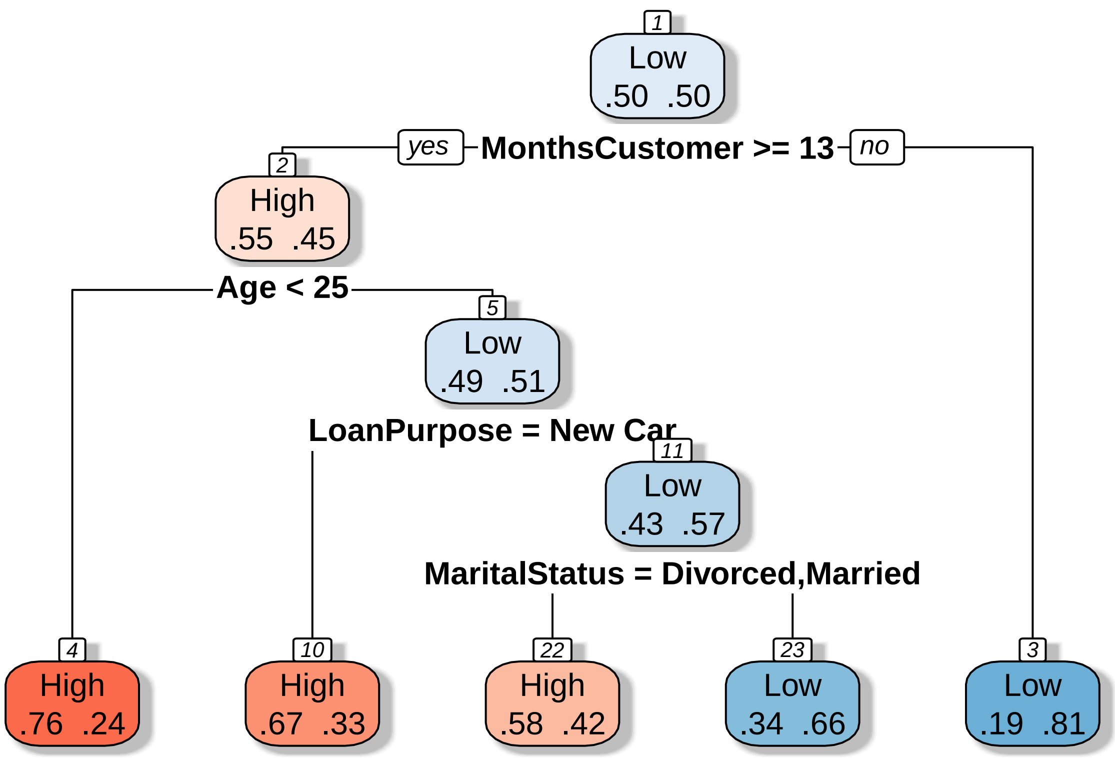
```
pohonKu
```

```
## n= 298
```

```
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##  1) root 298 148 Low (0.4966443 0.5033557)
##    2) MonthsCustomer>=12.5 251 112 High (0.5537849 0.4462151)
##      4) Age< 24.5 58  14 High (0.7586207 0.2413793) *
##      5) Age>=24.5 193  95 Low (0.4922280 0.5077720)
##       10) LoanPurpose=New Car 49  16 High (0.6734694 0.3265306) *
##       11) LoanPurpose=Business,Education,Furniture,Large Appliance,Other,Repairs,Retraining,Small App
##         22) MaritalStatus=Divorced,Married 55  23 High (0.5818182 0.4181818) *
##         23) MaritalStatus=Single 89  30 Low (0.3370787 0.6629213) *
##    3) MonthsCustomer< 12.5 47   9 Low (0.1914894 0.8085106) *
```

Menampilkan pohon klasifikasi

```
rpart.plot(pohonKu, extra=4,box.palette="RdBu", shadow.col="gray", nn=TRUE)
```



Mengukur kinerja prediksi

```
prediksiTree <- predict(pohonKu, data.test)
head(prediksiTree, n=10)
```

```
##          High       Low
```

```
## 1  0.3370787 0.6629213
## 2  0.1914894 0.8085106
## 3  0.5818182 0.4181818
## 4  0.6734694 0.3265306
## 5  0.1914894 0.8085106
## 6  0.3370787 0.6629213
## 7  0.3370787 0.6629213
## 8  0.5818182 0.4181818
## 9  0.3370787 0.6629213
## 10 0.3370787 0.6629213
```

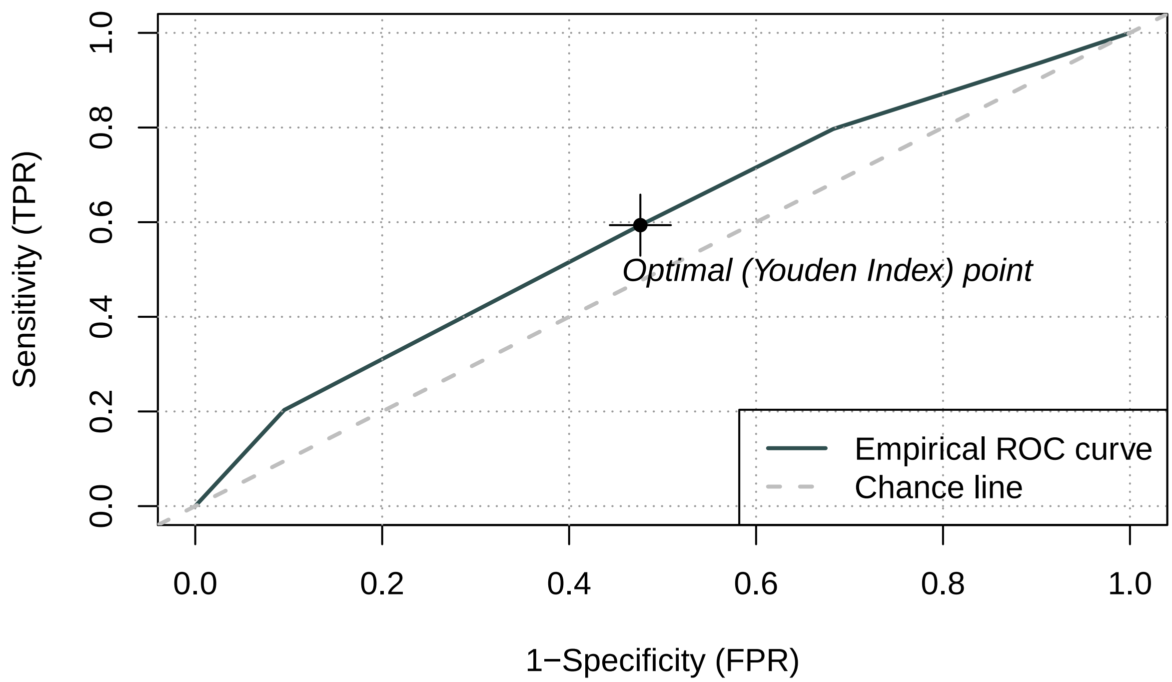
```
prediksi.status.t <- ifelse(prediksiTree[,2] > 0.5, "Low", "High")

#menghitung ukuran kinerja prediksi
confusionMatrix(as.factor(prediksi.status.t), as.factor(data.test$CreditRisk))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction High Low
##       High   33  26
##       Low    30  38
##
##                Accuracy : 0.5591
##                  95% CI : (0.4683, 0.647)
##     No Information Rate : 0.5039
##     P-Value [Acc > NIR] : 0.1243
##
##                   Kappa : 0.1176
##
##  Mcnemar's Test P-Value : 0.6885
##
##             Sensitivity : 0.5238
##             Specificity : 0.5938
##          Pos Pred Value : 0.5593
##          Neg Pred Value : 0.5588
##              Prevalence : 0.4961
##          Detection Rate : 0.2598
##    Detection Prevalence : 0.4646
##       Balanced Accuracy : 0.5588
##
##        'Positive' Class : High
##
```

Berdasarkan hasil diatas bisa lihat bahwa nilai akurasi sebesar 55,9%, Sensitivity (High) 52,4% dan Specificity (Low) 59,4%

Hitung Nilai Performance dari Prediksi

```
ngitungROCt <- rocit(score=prediksiTree[,2],class=data.test$CreditRisk)
plot(ngitungROCt)
```




```
AUCt <- ngitungROCt$AUC
AUCt
```

```
## [1] 0.5899058
```

Nilai AUC nya adalah 0,59, jadi bisa disimpulkan bahwa klasifikasi yang dihasilkan termasuk pada *poor classification*