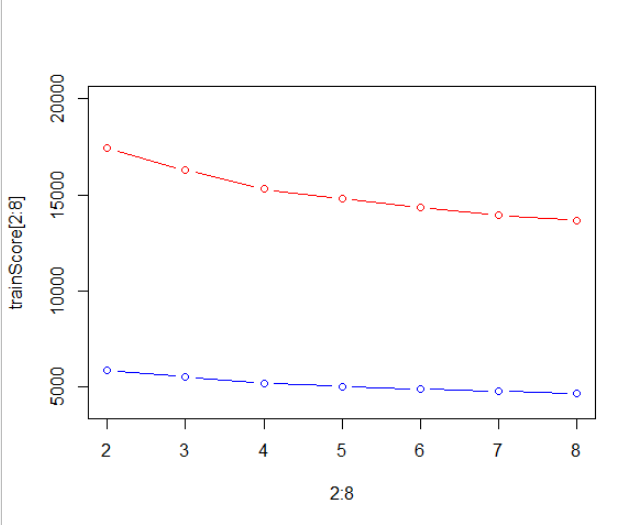
**1.1**



Optimal tree with 8 leaves.

Bayes variance: the more leaves, the higher variance but smaller bias.

> print(lT)

node), split, n, deviance, yval, (yprob)

\* denotes terminal node

1) root 19536 21610.00 <=50K ( 0.758344 0.241656 )

2) C8: Not-in-family, Other-relative, Own-child, Unmarried 10666 5224.00 <=50K ( 0.933340 0.066660 )

4) C9 < 7073.5 10479 4190.00 <=50K ( 0.949518 0.050482 )

8) C4: 10th, 11th, 12th, 1st-4th, 5th-6th, 7th-8th, 9th, Assoc-acdm, Assoc-voc, HS-grad, Preschool, Some-college 8431 2031.00 <=50K ( 0.974024 0.025976 ) \*

9) C4: Bachelors, Doctorate, Masters, Prof-school 2048 1741.00 <=50K ( 0.848633 0.151367 ) \*

5) C9 > 7073.5 187 46.08 >50K ( 0.026738 0.973262 ) \*

3) C8: Husband, Wife 8870 12210.00 <=50K ( 0.547914 0.452086 )

6) C7: ?, Adm-clerical, Craft-repair, Farming-fishing, Handlers-cleaners, Machine-op-inspct, Other-service, Priv-house-serv, Protective-serv, Transport-moving 4947 5988.00 <=50K ( 0.706489 0.293511 )

12) C9 < 5095.5 4736 5458.00 <=50K ( 0.736909 0.263091 ) \*

13) C9 > 5095.5 211 47.30 >50K ( 0.023697 0.976303 ) \*

7) C7: Armed-Forces, Exec-managerial, Prof-specialty, Sales, Tech-support 3923 5070.00 >50K ( 0.347948 0.652052 )

14) C9 < 5095.5 3412 4592.00 >50K ( 0.399766 0.600234 )

28) C4: 10th, 11th, 12th, 1st-4th, 5th-6th, 7th-8th, 9th, Assoc-acdm, Assoc-voc, HS-grad, Some-college 1614 2224.00 <=50K ( 0.545229 0.454771 ) \*

29) C4: Bachelors, Doctorate, Masters, Prof-school 1798 2094.00 >50K ( 0.269188 0.730812 ) \*

15) C9 > 5095.5 511 14.47 >50K ( 0.001957 0.998043 ) \*

Different interpretations can be done.

1. People who luckily cell their assets have high income
2. Married people with good education have high income
3. Married with Armed-Forces, Exec-managerial, Prof-specialty, Sales, Tech-support high income

**1.2**

> rbind(pis,acc,F1)

[,1] [,2] [,3] [,4] [,5] [,6] [,7]

pis 0.1000000 0.2000000 0.3000000 0.4000000 0.5000000 0.6000000 0.7000000

acc 0.6549977 0.7299248 0.8410871 0.8410871 0.8481499 0.8481499 0.8481499

F1 0.5712650 0.6126404 0.6580773 0.6580773 0.6013704 0.6013704 0.6013704

[,8] [,9]

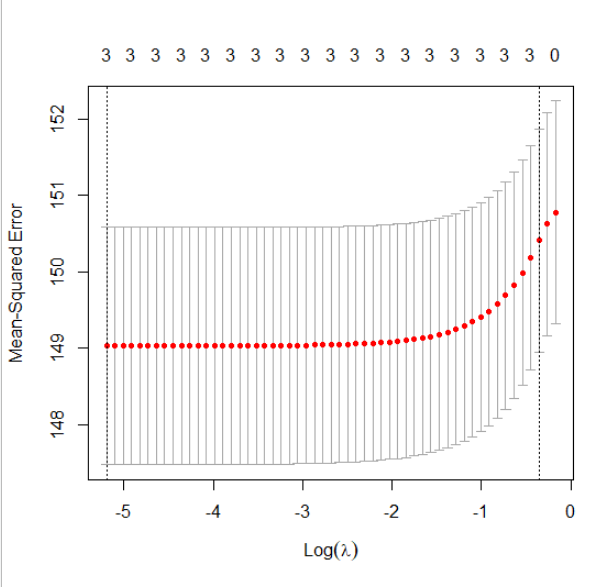
pis 0.8000000 0.9000000

acc 0.8063872 0.8063872

F1 0.3288984 0.3288984

Optimal threshold is 0.3 or 0.4, F1 should be prioritized since classes are quite imbalanced.

**1.3**

**¨**

model$lambda.min

[1] 0.005552479

> coef(model, s="lambda.min")

4 x 1 sparse Matrix of class "dgCMatrix"

1

(Intercept) 38.000806508

C1 0.052593047

C9 0.000107178

C10 0.001783425

3 variables selected, equation y=28+0.052C1+0.0001C9+0.0018C10+epsilon

The plot does not suggest that 0 features better than 3 (intervals overlap)