Appendix

Zhiwen Wang

2025-12-11

Appendix

Data prepocessing

```
data_raw <- read.table("C:/Users/pbcs/Desktop/Slides-STAT 4052/Project/credit_default32.txt",header=T)
factor_var <- c("SEX","EDUCATION","MARRIAGE","PAY_0","PAY_2","PAY_3","PAY_4","PAY_5","PAY_6","Default")
data_raw[factor_var] <- lapply(data_raw[factor_var],factor)
data1 <- data_raw
data1 <- subset(data1,select=-ID)
num_var <- names(data1)[sapply(data1,is.numeric)]

for (a in num_var){
  data1[[a]][is.na(data1[[a]])] <- mean(data_raw[[a]],na.rm=T)
}

data1 <- data1[complete.cases(data1),]
```

Construct logistic regression model to predict Default

Construct model in different methods

```
# Construct a suitable logistic regression model to predict Default.
set.seed(123)

train_idx <- sample(1:nrow(data1),round(0.7*nrow(data1)))
train <- data1[train_idx,]
test <- data1[-train_idx,]

logit_full <- glm(Default ~ .,data=train,family="binomial")
summary(logit_full)
```

```
##
## Call:
## glm(formula = Default ~ ., family = "binomial", data = train)
##
## Coefficients: (7 not defined because of singularities)
##               Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept) -1.662e+01  9.502e+02  -0.017  0.98604
## LIMIT_BAL   -4.509e-07  6.785e-07  -0.665  0.50636
## SEX2         2.289e-02  1.285e-01   0.178  0.85859
## EDUCATION2   5.775e-02  1.469e-01   0.393  0.69412
## EDUCATION3   1.989e-01  1.949e-01   1.020  0.30763
## EDUCATION4  -1.461e+01  1.038e+03  -0.014  0.98877
## EDUCATION5  -1.129e-01  8.689e-01  -0.130  0.89662
## EDUCATION6  -1.429e+01  1.313e+03  -0.011  0.99132
## MARRIAGE1    1.510e+01  9.502e+02   0.016  0.98732
## MARRIAGE2    1.494e+01  9.502e+02   0.016  0.98745
## MARRIAGE3    1.505e+01  9.502e+02   0.016  0.98736
## AGE         -4.095e-03  8.496e-03  -0.482  0.62983
## PAY_0-1      1.137e+00  4.293e-01   2.648  0.00809 **
## PAY_00       3.600e-01  4.718e-01   0.763  0.44540
## PAY_01       1.492e+00  3.410e-01   4.376 1.21e-05 ***
## PAY_02       2.653e+00  4.306e-01   6.161 7.25e-10 ***
## PAY_03       1.686e+00  8.188e-01   2.059  0.03945 *
## PAY_04       1.772e+01  1.173e+03   0.015  0.98795
## PAY_07       1.021e+00  2.937e+03   0.000  0.99972
## PAY_08      -1.301e-02  2.037e+00  -0.006  0.99490
## PAY_2-1     -2.204e-01  3.976e-01  -0.554  0.57928
## PAY_20      -5.124e-02  5.123e-01  -0.100  0.92033
## PAY_21      -1.574e+01  1.656e+03  -0.010  0.99241
## PAY_22      -3.632e-01  4.373e-01  -0.831  0.40622
## PAY_23       3.968e-01  8.678e-01   0.457  0.64747
## PAY_25       1.311e+01  4.156e+03   0.003  0.99748
## PAY_26       3.110e+01  4.799e+03   0.006  0.99483
## PAY_27              NA         NA      NA       NA
## PAY_3-1     -2.453e-01  3.556e-01  -0.690  0.49033
## PAY_30      -1.031e-01  4.559e-01  -0.226  0.82101
## PAY_31              NA         NA      NA       NA
## PAY_32       4.988e-01  4.686e-01   1.065  0.28710
## PAY_33      -1.846e+00  1.644e+00  -1.123  0.26151
## PAY_34       2.197e-01  1.350e+00   0.163  0.87070
## PAY_35       1.541e+01  2.400e+03   0.006  0.99488
## PAY_36              NA         NA      NA       NA
## PAY_37       2.985e+01  3.393e+03   0.009  0.99298
## PAY_4-1     -2.007e-01  3.819e-01  -0.526  0.59923
## PAY_40      -2.107e-01  4.455e-01  -0.473  0.63627
## PAY_42      -5.638e-01  5.095e-01  -1.106  0.26855
## PAY_43       2.142e-01  8.965e-01   0.239  0.81113
## PAY_44      -3.071e+01  3.393e+03  -0.009  0.99278
## PAY_45              NA         NA      NA       NA
## PAY_46      -1.282e+01  4.156e+03  -0.003  0.99754
## PAY_47              NA         NA      NA       NA
## PAY_5-1      2.513e-01  3.866e-01   0.650  0.51576
## PAY_50       2.357e-01  4.560e-01   0.517  0.60522
## PAY_52       9.470e-01  5.065e-01   1.870  0.06153 .
## PAY_53       1.444e+00  1.109e+00   1.303  0.19268
## PAY_54      -9.623e-01  1.933e+00  -0.498  0.61855
## PAY_55       5.987e+01  4.799e+03   0.012  0.99005
## PAY_57              NA         NA      NA       NA
## PAY_6-1     -2.815e-01  3.051e-01  -0.923  0.35619
## PAY_60      -4.592e-01  3.328e-01  -1.380  0.16774
```

```
## PAY_62      -4.242e-01  3.986e-01  -1.064  0.28719
## PAY_63       1.556e+00  1.026e+00   1.517  0.12924
## PAY_64       1.485e+01  2.400e+03   0.006  0.99506
## PAY_66      -1.390e+01  2.400e+03  -0.006  0.99538
## PAY_67      -1.388e+01  3.794e+03  -0.004  0.99708
## PAY_68              NA         NA      NA       NA
## BILL_AMT1   -4.302e-06  4.787e-06  -0.899  0.36890
## BILL_AMT2    5.879e-06  5.956e-06   0.987  0.32360
## BILL_AMT3    5.975e-06  4.444e-06   1.344  0.17879
## BILL_AMT4   -4.879e-07  4.276e-06  -0.114  0.90914
## BILL_AMT5   -4.729e-06  6.226e-06  -0.760  0.44748
## BILL_AMT6   -2.097e-07  4.563e-06  -0.046  0.96335
## PAY_AMT1    -4.058e-05  1.494e-05  -2.717  0.00659 **
## PAY_AMT2    -7.353e-06  7.436e-06  -0.989  0.32274
## PAY_AMT3     5.729e-06  3.330e-06   1.721  0.08532 .
## PAY_AMT4     9.646e-07  7.051e-06   0.137  0.89119
## PAY_AMT5    -1.383e-05  9.847e-06  -1.404  0.16021
## PAY_AMT6     1.291e-06  3.991e-06   0.324  0.74631
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1998.3  on 1886  degrees of freedom
## Residual deviance: 1669.8  on 1822  degrees of freedom
## AIC: 1799.8
##
## Number of Fisher Scoring iterations: 15
```

```
# Backward stepwise
model_back <- step(logit_full,direction="backward")
```

```
## Start:  AIC=1799.83
## Default ~ LIMIT_BAL + SEX + EDUCATION + MARRIAGE + AGE + PAY_0 +
##     PAY_2 + PAY_3 + PAY_4 + PAY_5 + PAY_6 + BILL_AMT1 + BILL_AMT2 +
##     BILL_AMT3 + BILL_AMT4 + BILL_AMT5 + BILL_AMT6 + PAY_AMT1 +
##     PAY_AMT2 + PAY_AMT3 + PAY_AMT4 + PAY_AMT5 + PAY_AMT6
##
##             Df Deviance    AIC
## - PAY_2      5   1672.4 1792.4
## - EDUCATION  5   1673.0 1793.0
## - PAY_4      5   1673.2 1793.2
## - PAY_6      7   1677.3 1793.3
## - MARRIAGE   3   1673.3 1797.3
## - BILL_AMT6  1   1669.8 1797.8
## - BILL_AMT4  1   1669.8 1797.8
## - PAY_AMT4   1   1669.8 1797.8
## - SEX        1   1669.9 1797.9
## - PAY_AMT6   1   1669.9 1797.9
## - AGE        1   1670.1 1798.1
## - LIMIT_BAL  1   1670.3 1798.3
## - BILL_AMT5  1   1670.4 1798.4
## - PAY_5      6   1680.4 1798.4
## - BILL_AMT1  1   1670.7 1798.7
```

```
## - BILL_AMT2  1   1670.8 1798.8
## - PAY_AMT2   1   1671.0 1799.0
## - PAY_3      5   1679.1 1799.1
## - BILL_AMT3  1   1671.7 1799.7
## <none>           1669.8 1799.8
## - PAY_AMT5   1   1672.1 1800.1
## - PAY_AMT3   1   1672.5 1800.5
## - PAY_AMT1   1   1681.0 1809.0
## - PAY_0      7   1771.5 1887.5
##
## Step:  AIC=1792.38
## Default ~ LIMIT_BAL + SEX + EDUCATION + MARRIAGE + AGE + PAY_0 +
##     PAY_3 + PAY_4 + PAY_5 + PAY_6 + BILL_AMT1 + BILL_AMT2 + BILL_AMT3 +
##     BILL_AMT4 + BILL_AMT5 + BILL_AMT6 + PAY_AMT1 + PAY_AMT2 +
##     PAY_AMT3 + PAY_AMT4 + PAY_AMT5 + PAY_AMT6
##
##             Df Deviance    AIC
## - PAY_4      6   1677.5 1785.5
## - EDUCATION  5   1675.6 1785.6
## - PAY_6      7   1679.8 1785.8
## - MARRIAGE   3   1675.8 1789.8
## - BILL_AMT6  1   1672.4 1790.4
## - PAY_AMT4   1   1672.4 1790.4
## - BILL_AMT4  1   1672.4 1790.4
## - SEX        1   1672.4 1790.4
## - PAY_AMT6   1   1672.5 1790.5
## - AGE        1   1672.6 1790.6
## - LIMIT_BAL  1   1672.8 1790.8
## - BILL_AMT5  1   1672.9 1790.9
## - BILL_AMT1  1   1673.5 1791.5
## - PAY_AMT2   1   1673.6 1791.6
## - BILL_AMT2  1   1673.6 1791.6
## - PAY_5      6   1683.7 1791.7
## - BILL_AMT3  1   1674.3 1792.3
## <none>           1672.4 1792.4
## - PAY_3      7   1686.6 1792.6
## - PAY_AMT5   1   1674.6 1792.6
## - PAY_AMT3   1   1675.0 1793.0
## - PAY_AMT1   1   1684.1 1802.1
## - PAY_0      7   1798.6 1904.6
##
## Step:  AIC=1785.46
## Default ~ LIMIT_BAL + SEX + EDUCATION + MARRIAGE + AGE + PAY_0 +
##     PAY_3 + PAY_5 + PAY_6 + BILL_AMT1 + BILL_AMT2 + BILL_AMT3 +
##     BILL_AMT4 + BILL_AMT5 + BILL_AMT6 + PAY_AMT1 + PAY_AMT2 +
##     PAY_AMT3 + PAY_AMT4 + PAY_AMT5 + PAY_AMT6
##
##             Df Deviance    AIC
## - PAY_6      8   1685.1 1777.1
## - EDUCATION  5   1680.9 1778.9
## - PAY_5      6   1686.8 1782.8
## - MARRIAGE   3   1681.0 1783.0
## - BILL_AMT6  1   1677.5 1783.5
## - PAY_AMT4   1   1677.5 1783.5
```

```
## - SEX        1   1677.5 1783.5
## - BILL_AMT4  1   1677.5 1783.5
## - PAY_AMT6   1   1677.5 1783.5
## - PAY_3      7   1689.6 1783.6
## - AGE        1   1677.7 1783.7
## - LIMIT_BAL  1   1677.8 1783.8
## - BILL_AMT5  1   1677.9 1783.9
## - BILL_AMT1  1   1678.6 1784.6
## - BILL_AMT2  1   1678.8 1784.8
## - PAY_AMT2   1   1678.8 1784.8
## - BILL_AMT3  1   1679.3 1785.3
## <none>           1677.5 1785.5
## - PAY_AMT5   1   1679.6 1785.6
## - PAY_AMT3   1   1680.3 1786.3
## - PAY_AMT1   1   1688.9 1794.9
## - PAY_0      7   1801.8 1895.8
##
## Step:  AIC=1777.1
## Default ~ LIMIT_BAL + SEX + EDUCATION + MARRIAGE + AGE + PAY_0 +
##     PAY_3 + PAY_5 + BILL_AMT1 + BILL_AMT2 + BILL_AMT3 + BILL_AMT4 +
##     BILL_AMT5 + BILL_AMT6 + PAY_AMT1 + PAY_AMT2 + PAY_AMT3 +
##     PAY_AMT4 + PAY_AMT5 + PAY_AMT6
##
##             Df Deviance    AIC
## - EDUCATION  5   1688.6 1770.6
## - PAY_3      7   1696.2 1774.2
## - MARRIAGE   3   1688.7 1774.7
## - BILL_AMT4  1   1685.1 1775.1
## - SEX        1   1685.1 1775.1
## - BILL_AMT6  1   1685.1 1775.1
## - PAY_AMT4   1   1685.1 1775.1
## - PAY_AMT6   1   1685.2 1775.2
## - PAY_5      6   1695.3 1775.3
## - AGE        1   1685.3 1775.3
## - LIMIT_BAL  1   1685.5 1775.5
## - BILL_AMT5  1   1685.9 1775.9
## - BILL_AMT1  1   1686.2 1776.2
## - BILL_AMT2  1   1686.4 1776.4
## - PAY_AMT2   1   1686.5 1776.5
## - BILL_AMT3  1   1687.0 1777.0
## <none>           1685.1 1777.1
## - PAY_AMT5   1   1687.6 1777.6
## - PAY_AMT3   1   1688.3 1778.3
## - PAY_AMT1   1   1696.6 1786.6
## - PAY_0      7   1811.7 1889.7
##
## Step:  AIC=1770.58
## Default ~ LIMIT_BAL + SEX + MARRIAGE + AGE + PAY_0 + PAY_3 +
##     PAY_5 + BILL_AMT1 + BILL_AMT2 + BILL_AMT3 + BILL_AMT4 + BILL_AMT5 +
##     BILL_AMT6 + PAY_AMT1 + PAY_AMT2 + PAY_AMT3 + PAY_AMT4 + PAY_AMT5 +
##     PAY_AMT6
##
##             Df Deviance    AIC
## - PAY_3      7   1699.8 1767.8
```

```
## - MARRIAGE   3   1692.0 1768.0
## - BILL_AMT4  1   1688.6 1768.6
## - BILL_AMT6  1   1688.6 1768.6
## - PAY_AMT4   1   1688.6 1768.6
## - SEX        1   1688.6 1768.6
## - AGE        1   1688.6 1768.6
## - PAY_AMT6   1   1688.7 1768.7
## - PAY_5      6   1698.8 1768.8
## - LIMIT_BAL  1   1689.3 1769.3
## - BILL_AMT5  1   1689.3 1769.3
## - BILL_AMT1  1   1689.6 1769.6
## - PAY_AMT2   1   1689.8 1769.8
## - BILL_AMT2  1   1689.9 1769.9
## - BILL_AMT3  1   1690.2 1770.2
## <none>           1688.6 1770.6
## - PAY_AMT5   1   1691.2 1771.2
## - PAY_AMT3   1   1691.6 1771.6
## - PAY_AMT1   1   1700.8 1780.8
## - PAY_0      7   1816.2 1884.2
##
## Step:  AIC=1767.78
## Default ~ LIMIT_BAL + SEX + MARRIAGE + AGE + PAY_0 + PAY_5 +
##     BILL_AMT1 + BILL_AMT2 + BILL_AMT3 + BILL_AMT4 + BILL_AMT5 +
##     BILL_AMT6 + PAY_AMT1 + PAY_AMT2 + PAY_AMT3 + PAY_AMT4 + PAY_AMT5 +
##     PAY_AMT6
##
##             Df Deviance    AIC
## - MARRIAGE   3   1703.6 1765.6
## - BILL_AMT6  1   1699.8 1765.8
## - PAY_AMT4   1   1699.8 1765.8
## - BILL_AMT4  1   1699.8 1765.8
## - SEX        1   1699.8 1765.8
## - AGE        1   1699.9 1765.9
## - PAY_AMT6   1   1700.0 1766.0
## - LIMIT_BAL  1   1700.8 1766.8
## - BILL_AMT5  1   1700.8 1766.8
## - BILL_AMT1  1   1700.9 1766.9
## - BILL_AMT2  1   1701.0 1767.0
## - BILL_AMT3  1   1701.5 1767.5
## <none>           1699.8 1767.8
## - PAY_AMT2   1   1701.9 1767.9
## - PAY_AMT5   1   1702.6 1768.6
## - PAY_AMT3   1   1703.0 1769.0
## - PAY_5      7   1723.1 1777.1
## - PAY_AMT1   1   1712.1 1778.1
## - PAY_0      8   1850.5 1902.5
##
## Step:  AIC=1765.57
## Default ~ LIMIT_BAL + SEX + AGE + PAY_0 + PAY_5 + BILL_AMT1 +
##     BILL_AMT2 + BILL_AMT3 + BILL_AMT4 + BILL_AMT5 + BILL_AMT6 +
##     PAY_AMT1 + PAY_AMT2 + PAY_AMT3 + PAY_AMT4 + PAY_AMT5 + PAY_AMT6
##
##             Df Deviance    AIC
## - BILL_AMT6  1   1703.6 1763.6
```

```
## - PAY_AMT4   1   1703.6 1763.6
## - BILL_AMT4  1   1703.6 1763.6
## - SEX        1   1703.6 1763.6
## - AGE        1   1703.7 1763.7
## - PAY_AMT6   1   1703.7 1763.7
## - LIMIT_BAL  1   1704.3 1764.3
## - BILL_AMT5  1   1704.5 1764.5
## - BILL_AMT1  1   1704.7 1764.7
## - BILL_AMT2  1   1704.8 1764.8
## - BILL_AMT3  1   1705.2 1765.2
## <none>           1703.6 1765.6
## - PAY_AMT2   1   1705.8 1765.8
## - PAY_AMT5   1   1706.2 1766.2
## - PAY_AMT3   1   1706.9 1766.9
## - PAY_5      7   1726.5 1774.5
## - PAY_AMT1   1   1715.9 1775.9
## - PAY_0      8   1856.3 1902.3
##
## Step:  AIC=1763.57
## Default ~ LIMIT_BAL + SEX + AGE + PAY_0 + PAY_5 + BILL_AMT1 +
##     BILL_AMT2 + BILL_AMT3 + BILL_AMT4 + BILL_AMT5 + PAY_AMT1 +
##     PAY_AMT2 + PAY_AMT3 + PAY_AMT4 + PAY_AMT5 + PAY_AMT6
##
##             Df Deviance    AIC
## - PAY_AMT4   1   1703.6 1761.6
## - BILL_AMT4  1   1703.6 1761.6
## - SEX        1   1703.6 1761.6
## - AGE        1   1703.7 1761.7
## - PAY_AMT6   1   1703.8 1761.8
## - LIMIT_BAL  1   1704.3 1762.3
## - BILL_AMT1  1   1704.7 1762.7
## - BILL_AMT2  1   1704.8 1762.8
## - BILL_AMT3  1   1705.2 1763.2
## <none>           1703.6 1763.6
## - BILL_AMT5  1   1705.6 1763.6
## - PAY_AMT2   1   1705.8 1763.8
## - PAY_AMT3   1   1706.9 1764.9
## - PAY_AMT5   1   1707.5 1765.5
## - PAY_5      7   1726.5 1772.5
## - PAY_AMT1   1   1715.9 1773.9
## - PAY_0      8   1856.4 1900.4
##
## Step:  AIC=1761.57
## Default ~ LIMIT_BAL + SEX + AGE + PAY_0 + PAY_5 + BILL_AMT1 +
##     BILL_AMT2 + BILL_AMT3 + BILL_AMT4 + BILL_AMT5 + PAY_AMT1 +
##     PAY_AMT2 + PAY_AMT3 + PAY_AMT5 + PAY_AMT6
##
##             Df Deviance    AIC
## - BILL_AMT4  1   1703.6 1759.6
## - SEX        1   1703.6 1759.6
## - AGE        1   1703.7 1759.7
## - PAY_AMT6   1   1703.8 1759.8
## - LIMIT_BAL  1   1704.3 1760.3
## - BILL_AMT1  1   1704.7 1760.7
```

```
## - BILL_AMT2  1   1704.8 1760.8
## - BILL_AMT3  1   1705.2 1761.2
## <none>           1703.6 1761.6
## - PAY_AMT2   1   1705.8 1761.8
## - BILL_AMT5  1   1706.2 1762.2
## - PAY_AMT3   1   1706.9 1762.9
## - PAY_AMT5   1   1707.5 1763.5
## - PAY_5      7   1726.7 1770.7
## - PAY_AMT1   1   1716.1 1772.1
## - PAY_0      8   1856.7 1898.7
##
## Step:  AIC=1759.59
## Default ~ LIMIT_BAL + SEX + AGE + PAY_0 + PAY_5 + BILL_AMT1 +
##     BILL_AMT2 + BILL_AMT3 + BILL_AMT5 + PAY_AMT1 + PAY_AMT2 +
##     PAY_AMT3 + PAY_AMT5 + PAY_AMT6
##
##             Df Deviance    AIC
## - SEX        1   1703.7 1757.7
## - AGE        1   1703.7 1757.7
## - PAY_AMT6   1   1703.8 1757.8
## - LIMIT_BAL  1   1704.4 1758.4
## - BILL_AMT1  1   1704.7 1758.7
## - BILL_AMT2  1   1704.8 1758.8
## <none>           1703.6 1759.6
## - BILL_AMT3  1   1705.7 1759.7
## - PAY_AMT2   1   1705.9 1759.9
## - PAY_AMT3   1   1707.3 1761.3
## - PAY_AMT5   1   1707.5 1761.5
## - BILL_AMT5  1   1707.6 1761.6
## - PAY_5      7   1726.7 1768.7
## - PAY_AMT1   1   1716.1 1770.1
## - PAY_0      8   1856.9 1896.9
##
## Step:  AIC=1757.65
## Default ~ LIMIT_BAL + AGE + PAY_0 + PAY_5 + BILL_AMT1 + BILL_AMT2 +
##     BILL_AMT3 + BILL_AMT5 + PAY_AMT1 + PAY_AMT2 + PAY_AMT3 +
##     PAY_AMT5 + PAY_AMT6
##
##             Df Deviance    AIC
## - AGE        1   1703.7 1755.7
## - PAY_AMT6   1   1703.8 1755.8
## - LIMIT_BAL  1   1704.4 1756.4
## - BILL_AMT1  1   1704.8 1756.8
## - BILL_AMT2  1   1704.9 1756.9
## <none>           1703.7 1757.7
## - BILL_AMT3  1   1705.8 1757.8
## - PAY_AMT2   1   1706.0 1758.0
## - PAY_AMT3   1   1707.4 1759.4
## - PAY_AMT5   1   1707.5 1759.5
## - BILL_AMT5  1   1707.7 1759.7
## - PAY_5      7   1726.8 1766.8
## - PAY_AMT1   1   1716.2 1768.2
## - PAY_0      8   1857.0 1895.0
##
```

```
## Step:  AIC=1755.71
## Default ~ LIMIT_BAL + PAY_0 + PAY_5 + BILL_AMT1 + BILL_AMT2 +
##     BILL_AMT3 + BILL_AMT5 + PAY_AMT1 + PAY_AMT2 + PAY_AMT3 +
##     PAY_AMT5 + PAY_AMT6
##
##             Df Deviance    AIC
## - PAY_AMT6   1   1703.9 1753.9
## - LIMIT_BAL  1   1704.4 1754.4
## - BILL_AMT1  1   1704.8 1754.8
## - BILL_AMT2  1   1704.9 1754.9
## <none>           1703.7 1755.7
## - BILL_AMT3  1   1705.9 1755.9
## - PAY_AMT2   1   1706.0 1756.0
## - PAY_AMT3   1   1707.5 1757.5
## - PAY_AMT5   1   1707.6 1757.6
## - BILL_AMT5  1   1707.8 1757.8
## - PAY_5      7   1727.0 1765.0
## - PAY_AMT1   1   1716.2 1766.2
## - PAY_0      8   1857.2 1893.2
##
## Step:  AIC=1753.92
## Default ~ LIMIT_BAL + PAY_0 + PAY_5 + BILL_AMT1 + BILL_AMT2 +
##     BILL_AMT3 + BILL_AMT5 + PAY_AMT1 + PAY_AMT2 + PAY_AMT3 +
##     PAY_AMT5
##
##             Df Deviance    AIC
## - LIMIT_BAL  1   1704.6 1752.6
## - BILL_AMT1  1   1705.0 1753.0
## - BILL_AMT2  1   1705.1 1753.1
## <none>           1703.9 1753.9
## - PAY_AMT2   1   1706.1 1754.1
## - BILL_AMT3  1   1706.1 1754.1
## - PAY_AMT3   1   1707.7 1755.7
## - PAY_AMT5   1   1707.8 1755.8
## - BILL_AMT5  1   1708.0 1756.0
## - PAY_5      7   1727.2 1763.2
## - PAY_AMT1   1   1716.3 1764.3
## - PAY_0      8   1857.5 1891.5
##
## Step:  AIC=1752.6
## Default ~ PAY_0 + PAY_5 + BILL_AMT1 + BILL_AMT2 + BILL_AMT3 +
##     BILL_AMT5 + PAY_AMT1 + PAY_AMT2 + PAY_AMT3 + PAY_AMT5
##
##             Df Deviance    AIC
## - BILL_AMT1  1   1705.8 1751.8
## - BILL_AMT2  1   1705.9 1751.9
## <none>           1704.6 1752.6
## - BILL_AMT3  1   1706.8 1752.8
## - PAY_AMT2   1   1707.0 1753.0
## - PAY_AMT3   1   1708.3 1754.3
## - PAY_AMT5   1   1708.8 1754.8
## - BILL_AMT5  1   1709.2 1755.2
## - PAY_5      7   1728.8 1762.8
## - PAY_AMT1   1   1717.9 1763.9
```

```
## - PAY_0      8   1859.2 1891.2
##
## Step:  AIC=1751.85
## Default ~ PAY_0 + PAY_5 + BILL_AMT2 + BILL_AMT3 + BILL_AMT5 +
##     PAY_AMT1 + PAY_AMT2 + PAY_AMT3 + PAY_AMT5
##
##             Df Deviance    AIC
## - BILL_AMT2  1   1706.0 1750.0
## <none>           1705.8 1751.8
## - BILL_AMT3  1   1708.2 1752.2
## - PAY_AMT2   1   1708.6 1752.6
## - PAY_AMT3   1   1709.6 1753.6
## - PAY_AMT5   1   1710.3 1754.3
## - BILL_AMT5  1   1710.5 1754.5
## - PAY_AMT1   1   1718.0 1762.0
## - PAY_5      7   1730.1 1762.1
## - PAY_0      8   1862.9 1892.9
##
## Step:  AIC=1749.99
## Default ~ PAY_0 + PAY_5 + BILL_AMT3 + BILL_AMT5 + PAY_AMT1 +
##     PAY_AMT2 + PAY_AMT3 + PAY_AMT5
##
##             Df Deviance    AIC
## <none>           1706.0 1750.0
## - PAY_AMT2   1   1710.2 1752.2
## - PAY_AMT5   1   1710.4 1752.4
## - PAY_AMT3   1   1710.5 1752.5
## - BILL_AMT5  1   1710.6 1752.6
## - BILL_AMT3  1   1713.2 1755.2
## - PAY_5      7   1730.1 1760.1
## - PAY_AMT1   1   1718.2 1760.2
## - PAY_0      8   1863.8 1891.8
```

```
summary(model_back)
```

```
##
## Call:
## glm(formula = Default ~ PAY_0 + PAY_5 + BILL_AMT3 + BILL_AMT5 +
##     PAY_AMT1 + PAY_AMT2 + PAY_AMT3 + PAY_AMT5, family = "binomial",
##     data = train)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.830e+00  2.731e-01  -6.700 2.09e-11 ***
## PAY_0-1      8.019e-01  3.293e-01   2.435  0.01489 *
## PAY_00       2.314e-01  3.335e-01   0.694  0.48779
## PAY_01       1.346e+00  3.168e-01   4.250 2.14e-05 ***
## PAY_02       2.444e+00  3.604e-01   6.782 1.18e-11 ***
## PAY_03       1.417e+00  7.448e-01   1.903  0.05708 .
## PAY_04       1.770e+01  6.684e+02   0.026  0.97888
## PAY_07       1.617e+01  1.029e+03   0.016  0.98746
## PAY_08       7.519e-01  1.575e+00   0.477  0.63317
## PAY_5-1     -2.135e-01  2.246e-01  -0.950  0.34191
## PAY_50      -2.468e-01  2.178e-01  -1.134  0.25700
```

```
## PAY_52       5.485e-01  2.583e-01   2.123  0.03376 *
## PAY_53       9.497e-01  7.795e-01   1.218  0.22310
## PAY_54      -5.852e-02  1.046e+00  -0.056  0.95539
## PAY_55       1.553e+01  1.455e+03   0.011  0.99149
## PAY_57       1.562e+01  6.839e+02   0.023  0.98178
## BILL_AMT3    7.089e-06  2.561e-06   2.768  0.00564 **
## BILL_AMT5   -6.124e-06  2.842e-06  -2.155  0.03116 *
## PAY_AMT1    -3.607e-05  1.284e-05  -2.809  0.00497 **
## PAY_AMT2    -1.265e-05  7.685e-06  -1.647  0.09962 .
## PAY_AMT3     7.109e-06  2.795e-06   2.544  0.01097 *
## PAY_AMT5    -1.512e-05  8.578e-06  -1.762  0.07801 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1998.3  on 1886  degrees of freedom
## Residual deviance: 1706.0  on 1865  degrees of freedom
## AIC: 1750
##
## Number of Fisher Scoring iterations: 14
```

```
# Forward stepwise (Failure)
logit_null <- glm(Default ~ 1,data=train,family="binomial")
step(logit_null,scope=list(lower=~1,upper=~.),direction="forward")
```

```
## Start:  AIC=2000.28
## Default ~ 1

##
## Call:  glm(formula = Default ~ 1, family = "binomial", data = train)
##
## Coefficients:
## (Intercept)
##      -1.254
##
## Degrees of Freedom: 1886 Total (i.e. Null);  1886 Residual
## Null Deviance:       1998
## Residual Deviance: 1998  AIC: 2000
```

```
# LASSO logistic regression
library(glmnet)
```

```
## Warning: package 'glmnet' was built under R version 4.4.3

## Loading required package: Matrix

## Warning: package 'Matrix' was built under R version 4.4.3

## Loaded glmnet 4.1-10
```

```r
x_train <- model.matrix(Default ~ .,train)[,-1]
y <- as.numeric(train$Default)-1

set.seed(123)
cvfit <- cv.glmnet(x_train,y,family="binomial",alpha=1)

best_lambda <- cvfit$lambda.min
model_lasso <- glmnet(x_train,y,family="binomial",lambda=best_lambda)
coef(model_lasso)
```

```
## 73 x 1 sparse Matrix of class "dgCMatrix"
##                          s0
## (Intercept) -1.447443e+00
## LIMIT_BAL   -5.243064e-08
## SEX2         .
## EDUCATION2   .
## EDUCATION3   6.259483e-02
## EDUCATION4   .
## EDUCATION5   .
## EDUCATION6   .
## MARRIAGE1    4.664055e-03
## MARRIAGE2    .
## MARRIAGE3    .
## AGE          .
## PAY_0-1      .
## PAY_00      -3.077519e-01
## PAY_01       6.121978e-01
## PAY_02       1.630258e+00
## PAY_03       7.967036e-02
## PAY_04       1.482796e+00
## PAY_07       .
## PAY_08       .
## PAY_2-1      .
## PAY_20       .
## PAY_21       .
## PAY_22       .
## PAY_23       7.204375e-01
## PAY_24       .
## PAY_25       .
## PAY_26       5.558626e-01
## PAY_27       .
## PAY_3-1      .
## PAY_30       .
## PAY_31       .
## PAY_32       3.640934e-01
## PAY_33       .
## PAY_34       .
## PAY_35       1.461396e+00
## PAY_36       .
## PAY_37       1.256564e+00
## PAY_4-1      .
## PAY_40       .
## PAY_42       .
```

```
## PAY_43       6.724284e-02
## PAY_44       .
## PAY_45       .
## PAY_46       .
## PAY_47       .
## PAY_5-1      .
## PAY_50       .
## PAY_52       4.247674e-01
## PAY_53       4.136615e-01
## PAY_54       .
## PAY_55       .
## PAY_57       3.097277e-02
## PAY_6-1      .
## PAY_60      -5.007639e-02
## PAY_62       .
## PAY_63       4.382988e-01
## PAY_64       .
## PAY_66       .
## PAY_67       .
## PAY_68       .
## BILL_AMT1    .
## BILL_AMT2    .
## BILL_AMT3    .
## BILL_AMT4    .
## BILL_AMT5    .
## BILL_AMT6    .
## PAY_AMT1    -1.206597e-05
## PAY_AMT2    -1.881032e-06
## PAY_AMT3     .
## PAY_AMT4     .
## PAY_AMT5    -3.144705e-06
## PAY_AMT6     .
```

Compare different models

```
## Backward model
pred_back_prob <- predict(model_back,newdata=test,type="response")

# Convert to class
pred_back_class <- ifelse(pred_back_prob>0.5,1,0)

# Truth
truth <- as.numeric(test$Default)-1

acc_back <- mean(pred_back_class == truth)
acc_back
```

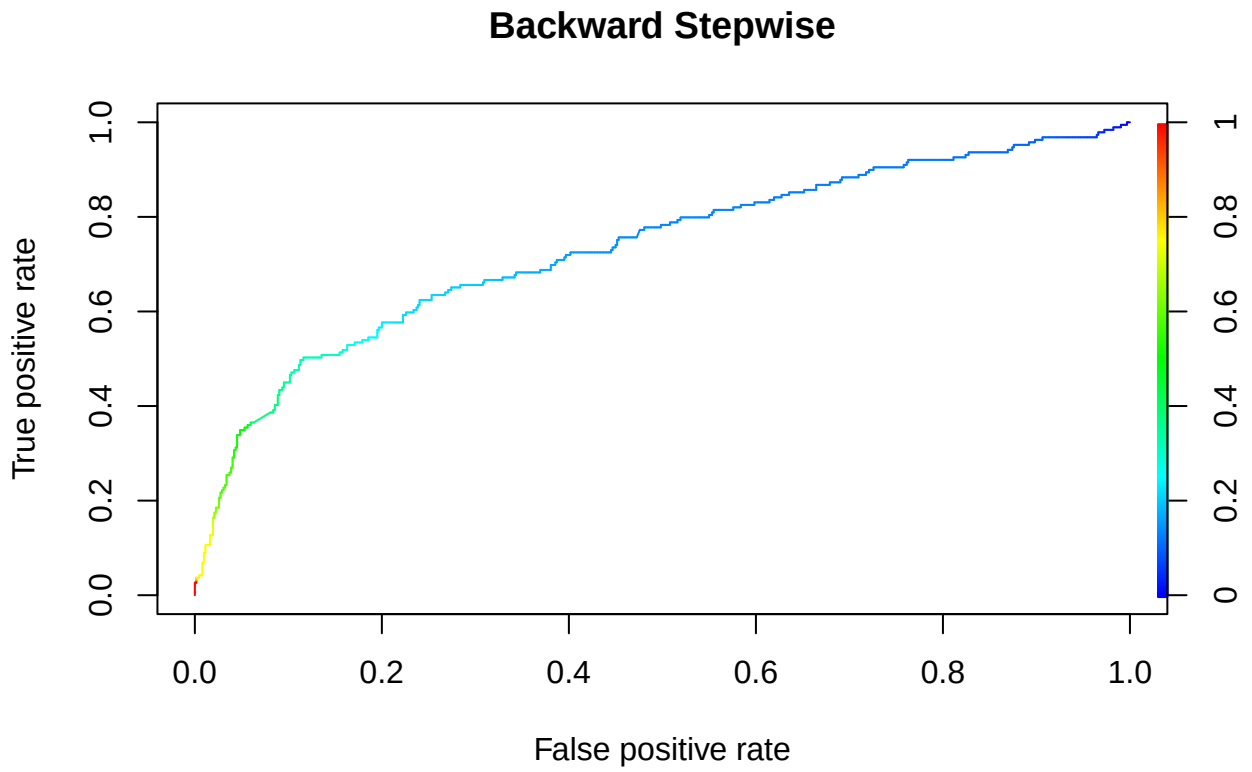
```
## [1] 0.8096415
```

```
library(ROCR)
```

```
## Warning: package 'ROCR' was built under R version 4.4.3
```

```r
pred_back <- prediction(pred_back_prob,truth)
perf_back <- performance(pred_back,"tpr","fpr")

plot(perf_back,colorize=T,main="Backward Stepwise")
```




```r
AUC_back <- performance(pred_back,"auc")@y.values[[1]]
AUC_back
```

```
## [1] 0.732881
```

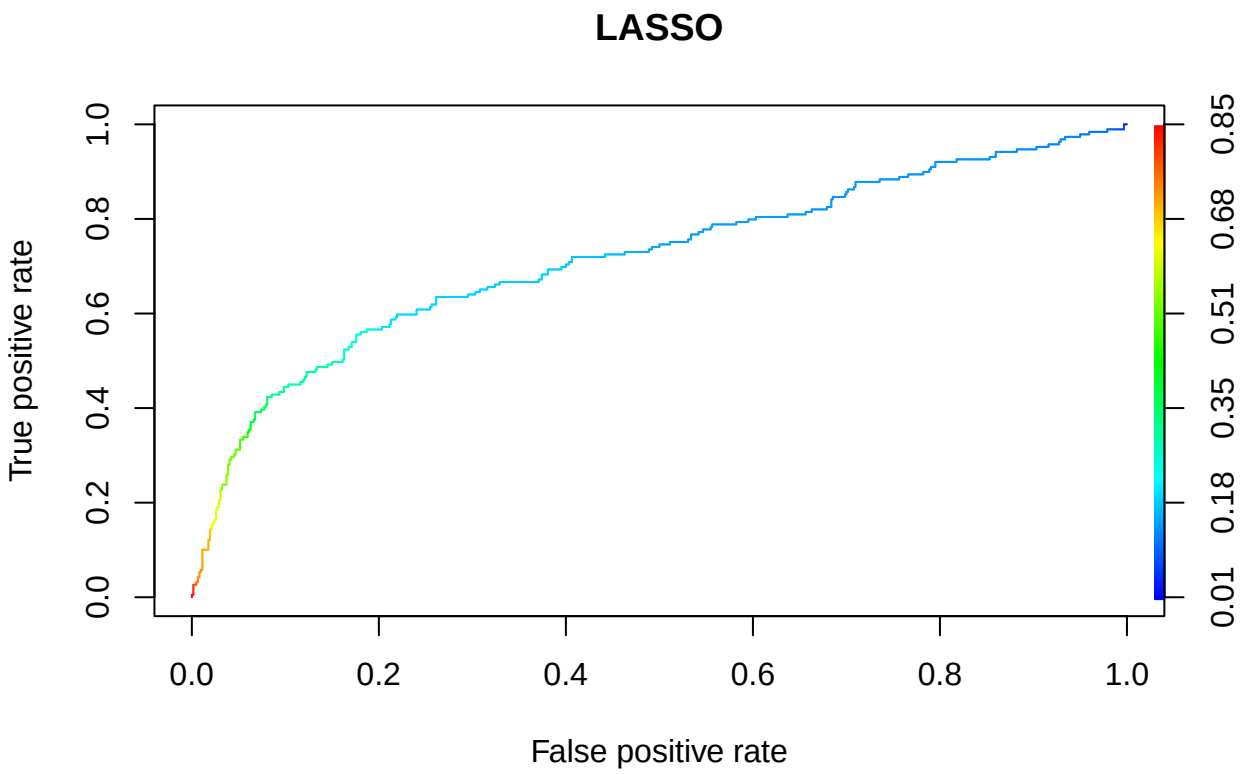
```r
## LASSO model
x_test <- model.matrix(Default ~ .,test)[,-1]
pred_lasso_prob <- predict(model_lasso,x_test,type="response")

pred_lasso_class <- ifelse(pred_lasso_prob>0.5,1,0)
acc_lasso <- mean(pred_lasso_class == truth)
acc_lasso
```

```
## [1] 0.8009889
```

```r
pred_lasso <- prediction(pred_lasso_prob,truth)
perf_lasso <- performance(pred_lasso,"tpr","fpr")

plot(perf_lasso,colorize=T,main="LASSO")
```




```
AUC_lasso <- performance(pred_lasso,"auc")@y.values[[1]]
AUC_lasso
```

```
## [1] 0.7171958
```

Use KNN to predict Default (procedure discussed in Handout 8)

This section is developed with reference to ChatGPT and the materials from Lab 8.

```
library(class)
```

```
## Warning: package 'class' was built under R version 4.4.3
```

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.4.3
```

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

```
## Warning: package 'ggplot2' was built under R version 4.4.3
```

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

```
## Warning: package 'lattice' was built under R version 4.4.3
```

```
library(fastDummies)
```

```
## Warning: package 'fastDummies' was built under R version 4.4.3
```

```
set.seed(123)
train_knn <- train
test_knn <- test

train_knn$Default <- as.numeric(train_knn$Default) - 1
test_knn$Default <- as.numeric(test_knn$Default) - 1

# Create dummy variables
cat_vars <- names(train_knn)[sapply(train_knn,is.factor)]
train_dummy <- dummy_cols(train_knn,remove_selected_columns = TRUE,remove_first_dummy = TRUE)
test_dummy <- dummy_cols(test_knn,remove_selected_columns = TRUE,remove_first_dummy = TRUE)

train_y <- factor(train_dummy$Default,levels = c(0,1))
test_y <- factor(test_dummy$Default,levels = c(0,1))

train_dummy$Default <- NULL
test_dummy$Default  <- NULL

num_vars <- names(train_knn)[sapply(train_knn,is.numeric) & names(train_knn) != "Default"]

preProcValues <- preProcess(train_knn[,num_vars],method = c("center", "scale"))
train_dummy_scaled <- train_dummy
test_dummy_scaled <- test_dummy

train_dummy_scaled[,num_vars] <- predict(preProcValues,train_knn[,num_vars])
test_dummy_scaled[,num_vars] <- predict(preProcValues,test_knn[,num_vars])

k_values <- seq(1,50,by=2)
acc_list <- c()

for (k in k_values){
  pred_k <- knn(train_dummy_scaled,test_dummy_scaled,cl = train_y,k = k)
  acc_list <- c(acc_list,mean(pred_k == test_y))
}

best_k <- k_values[which.max(acc_list)]
best_k
```

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

```
pred_knn <- knn(train_dummy_scaled,test_dummy_scaled,cl = train_y,k = best_k)

# Accuracy
acc_knn <- mean(pred_knn == test_y)
acc_knn
```

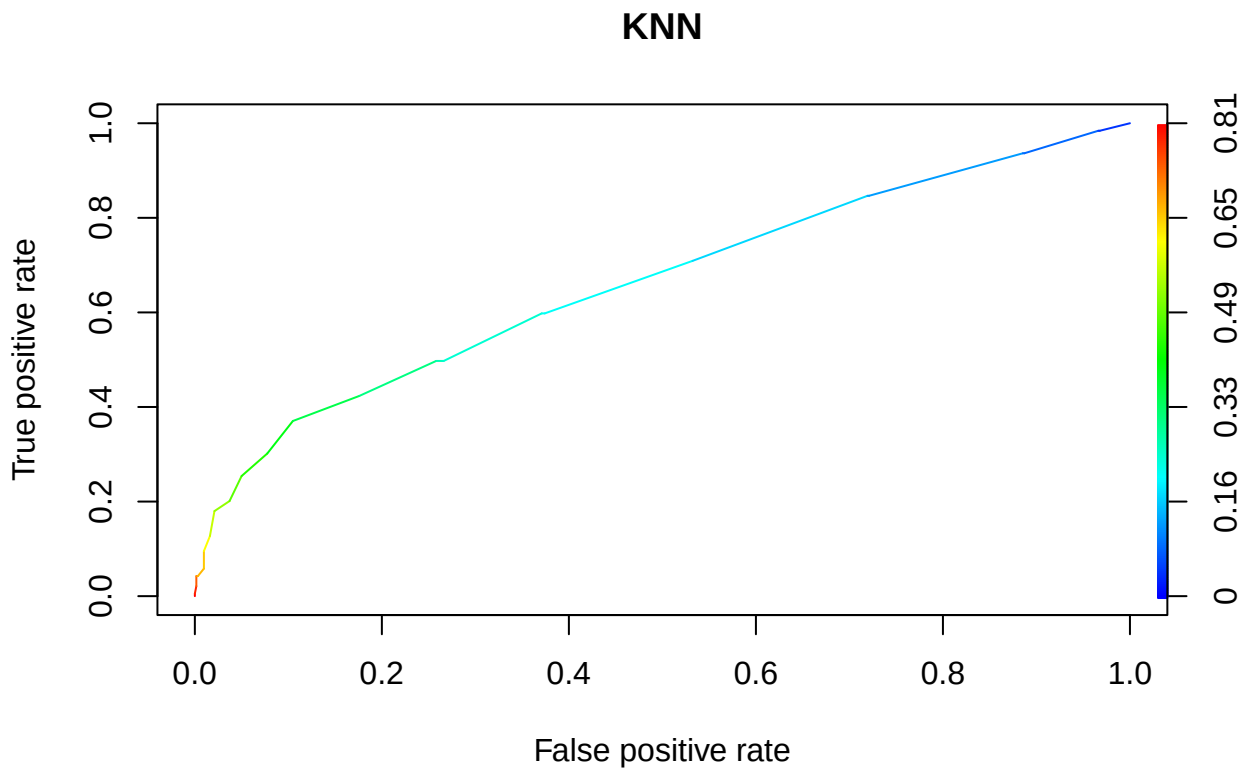
```
## [1] 0.7923362
```

```r
knn_prob <- attributes(knn(train_dummy_scaled, test_dummy_scaled,
                           cl = train_y, k = best_k, prob = TRUE))$prob
knn_prob <- ifelse(pred_knn == "1", knn_prob, 1 - knn_prob)
knn_prob <- as.numeric(knn_prob)

truth_knn <- as.numeric(test_y) - 1
pred_knn_ <- prediction(knn_prob,truth_knn)
AUC_knn <- performance(pred_knn_,"auc")@y.values[[1]]
AUC_knn
```

```
## [1] 0.6619261
```

```r
# ROC curve
perf_knn <- performance(pred_knn_,"tpr","fpr")
plot(perf_knn,colorize = TRUE,main = "KNN")
```



Use ensemble methods to predict Default

```r
# Random forest
library(randomForest)
```

```
## Warning: package 'randomForest' was built under R version 4.4.3
```

```
## randomForest 4.7-1.2

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

##
## Attaching package: 'randomForest'

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

```
set.seed(111)
m1 <- randomForest(Default~.,data=train,mtry=floor(sqrt(ncol(train)-1)),importance=T)
m1
```

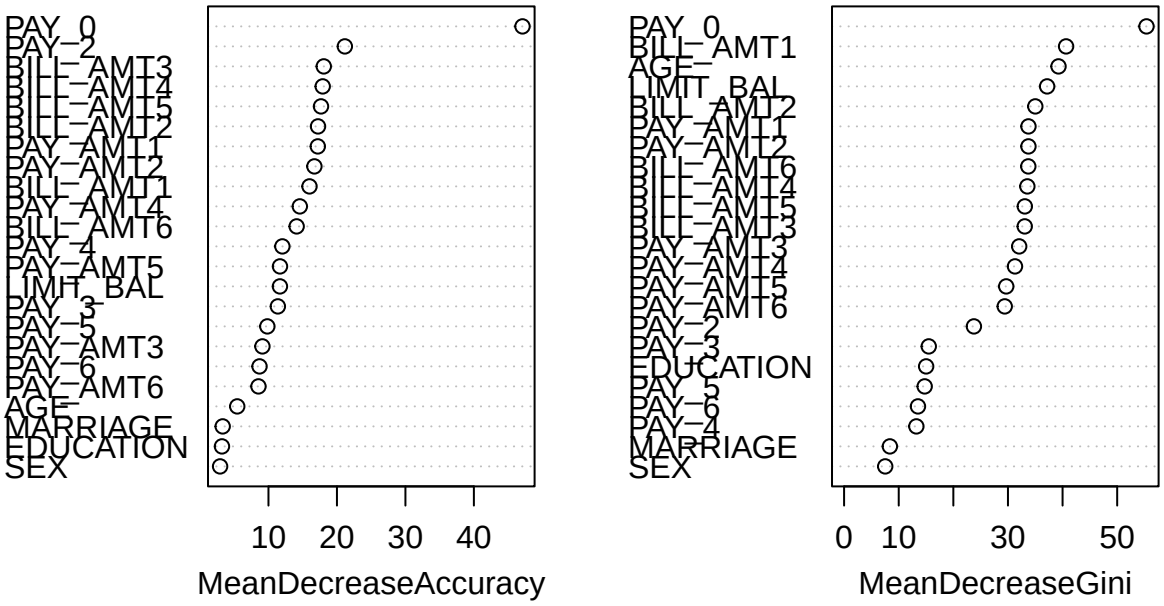
```
##
## Call:
##  randomForest(formula = Default ~ ., data = train, mtry = floor(sqrt(ncol(train) -      1)), importa
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 4
##
##         OOB estimate of  error rate: 20.56%
## Confusion matrix:
##      0   1 class.error
## 0 1388  80  0.05449591
## 1  308 111  0.73508353
```

```
## Predict on test set
pred_rf <- predict(m1,newdata=test,type="class")
mean(pred_rf!=test$Default)
```

```
## [1] 0.2076638
```

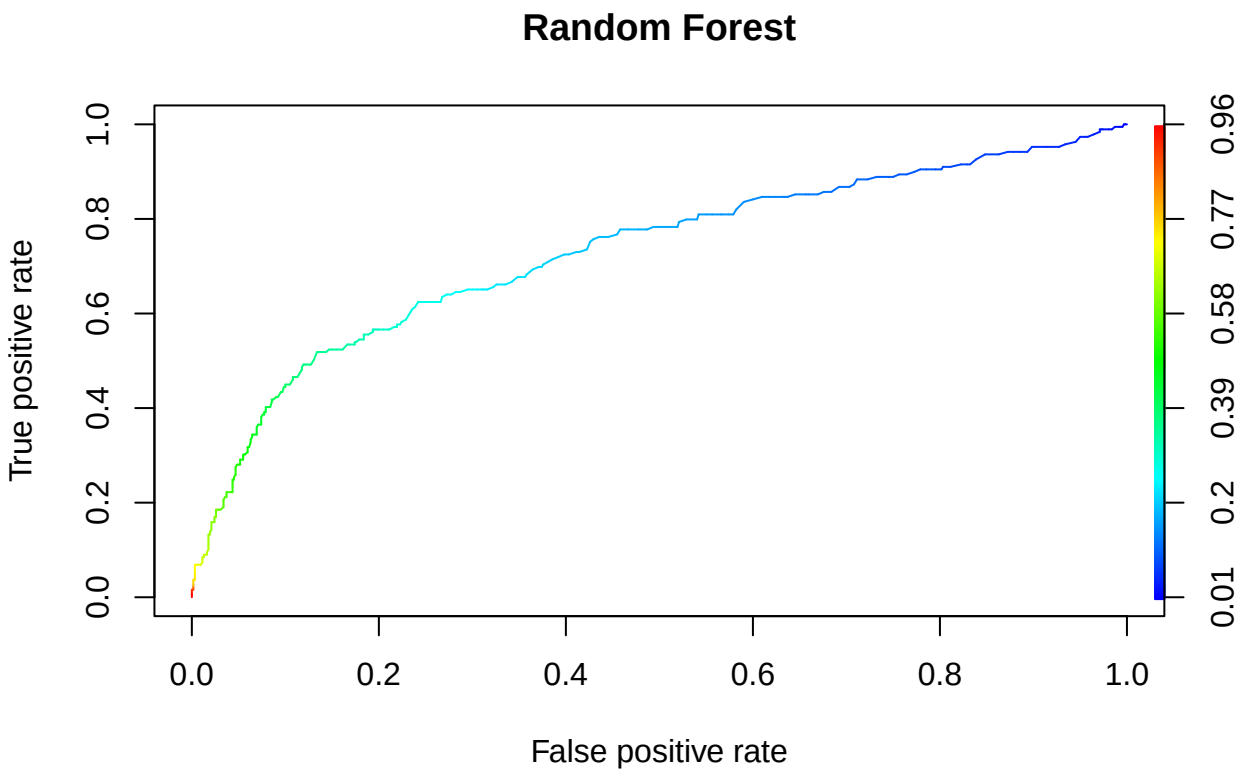
```
varImpPlot(m1)
```

m1




```
pred_rf_prob <- predict(m1,newdata=test,type="prob")[,2]
pred_rf <- prediction(pred_rf_prob,truth)
perf_rf <- performance(pred_rf,"tpr","fpr")

plot(perf_rf,colorize=T,main="Random Forest")
```




```
AUC_rf <- performance(pred_rf,"auc")@y.values[[1]]
AUC_rf
```

```
## [1] 0.7285501
```

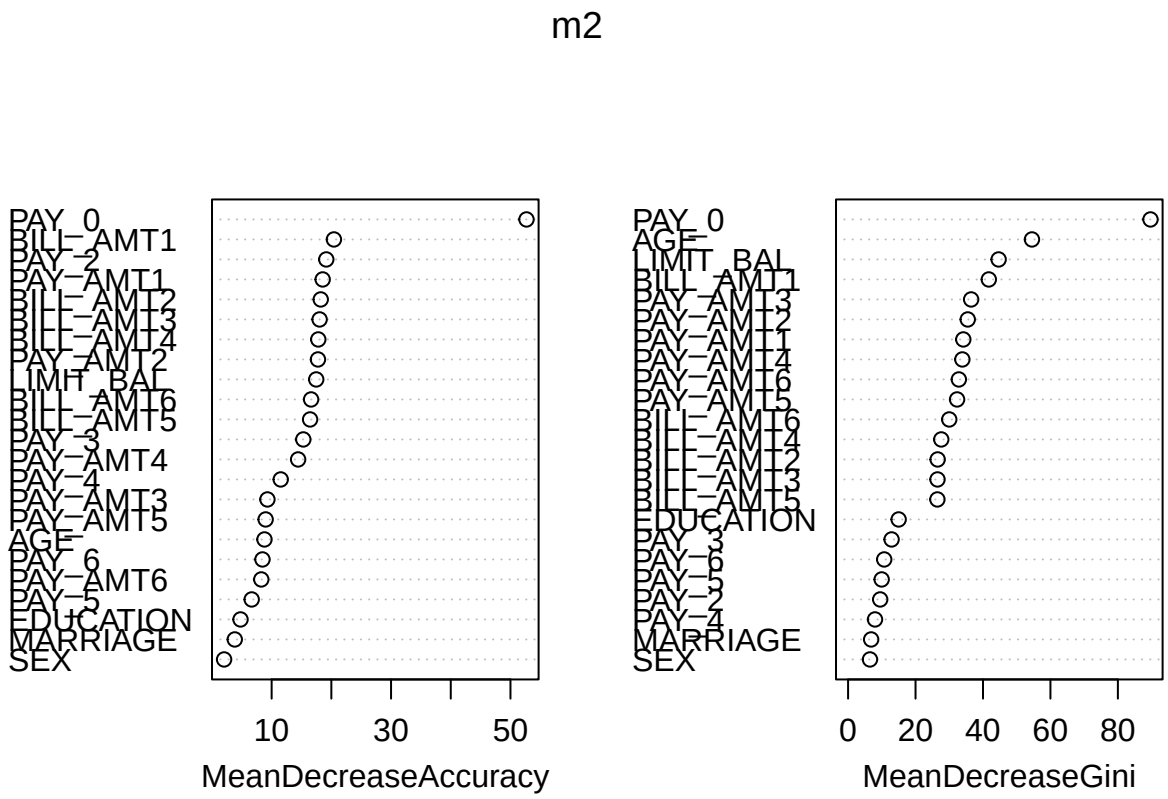
```
# Bagging
set.seed(222)
m2 <- randomForest(Default ~ .,data=train,mtry=ncol(train)-1,importance=T)
m2
```

```
##
## Call:
##  randomForest(formula = Default ~ ., data = train, mtry = ncol(train) -      1, importance = T)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 23
##
##         OOB estimate of  error rate: 20.4%
## Confusion matrix:
##      0   1 class.error
## 0 1380  88   0.0599455
## 1  297 122   0.7088305
```

```r
## Predict on test set
pred_bg <- predict(m2,newdata=test,type="class")
mean(pred_bg!=test$Default)
```

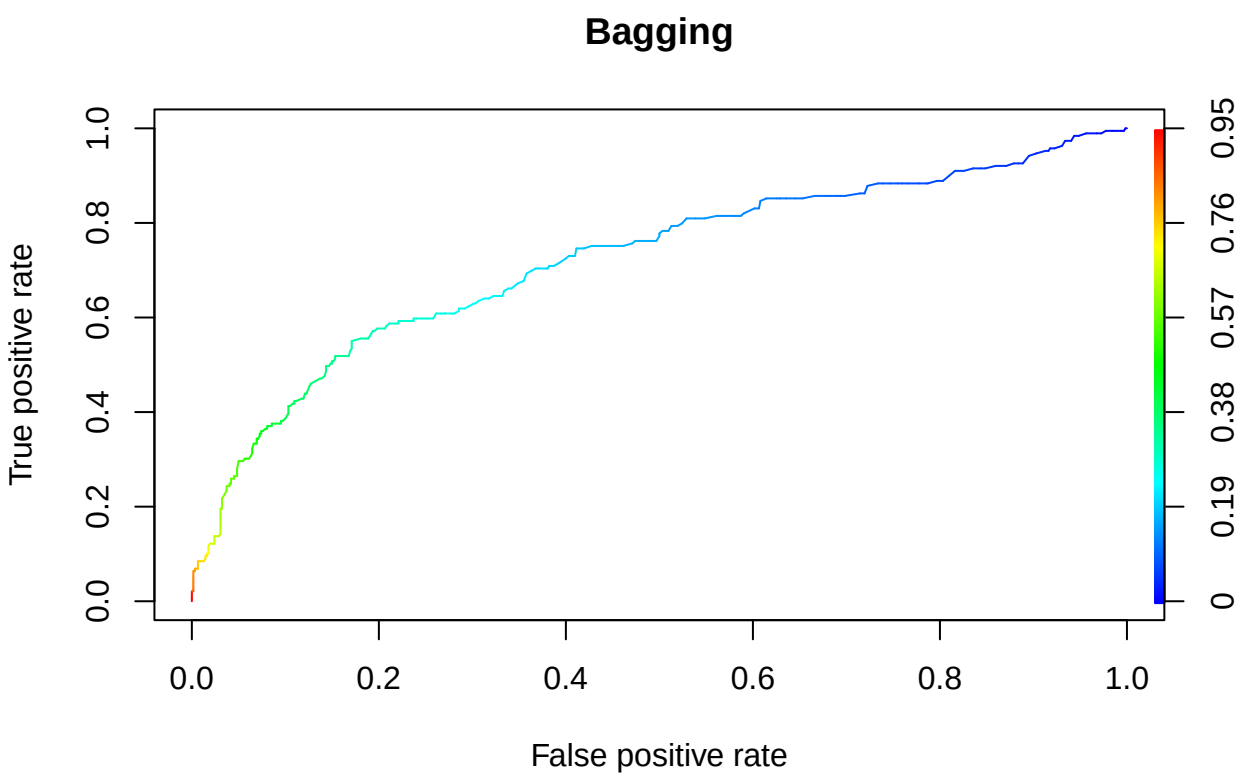
```
## [1] 0.2088999
```

```r
varImpPlot(m2)
```




```r
pred_bg_prob <- predict(m2,newdata=test,type="prob")[,2]
pred_bg <- prediction(pred_bg_prob,truth)
perf_bg <- performance(pred_bg,"tpr","fpr")

plot(perf_bg,colorize=T,main="Bagging")
```




```
AUC_bg <- performance(pred_bg,"auc")@y.values[[1]]
AUC_bg
```

```
## [1] 0.7206563
```

```
# Boosting
library(gbm)
```

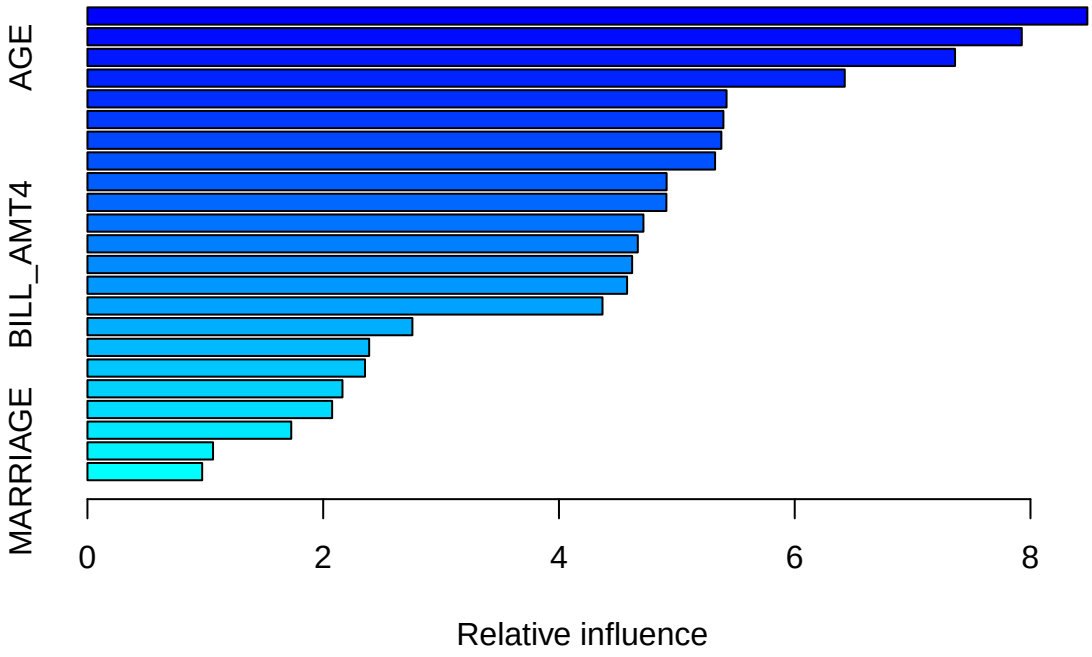
```
## Warning: package 'gbm' was built under R version 4.4.3
```

```
## Loaded gbm 2.2.2
```

```
## This version of gbm is no longer under development. Consider transitioning to gbm3, https://github.c
```

```
set.seed(333)
train_bt <- train
test_bt <- test
train_bt$Default <- as.numeric(train_bt$Default)-1
test_bt$Default <- as.numeric(test_bt$Default)-1

m3 <- gbm(Default ~ .,data=train_bt,distribution="bernoulli",n.trees=5000,interaction.depth=3,shrinkage=
summary(m3)
```



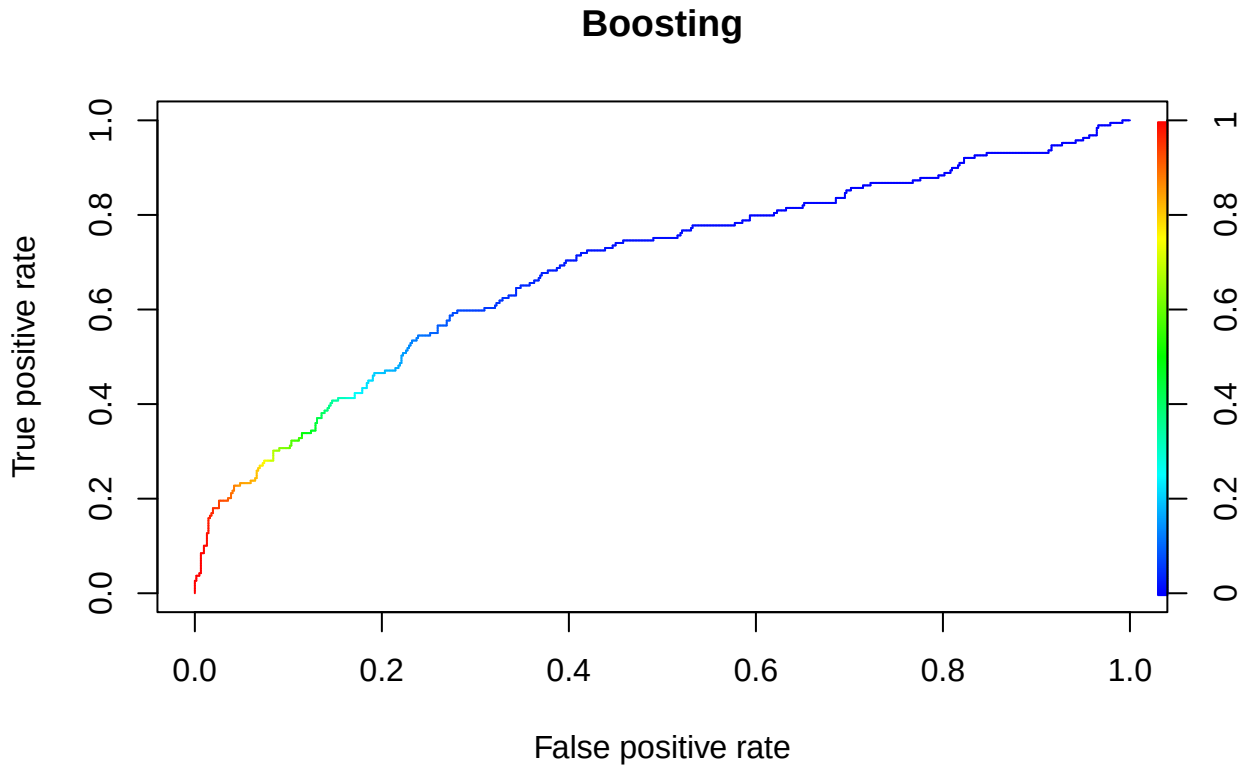

```
##                 var   rel.inf
## PAY_0         PAY_0 8.4825776
## BILL_AMT1 BILL_AMT1 7.9263687
## AGE             AGE 7.3601254
## LIMIT_BAL LIMIT_BAL 6.4244735
## PAY_AMT2   PAY_AMT2 5.4217857
## PAY_AMT1   PAY_AMT1 5.3952001
## PAY_AMT5   PAY_AMT5 5.3776112
## BILL_AMT6 BILL_AMT6 5.3245070
## PAY_AMT3   PAY_AMT3 4.9132171
## PAY_AMT4   PAY_AMT4 4.9113548
## BILL_AMT3 BILL_AMT3 4.7167525
## BILL_AMT2 BILL_AMT2 4.6688981
## BILL_AMT4 BILL_AMT4 4.6207570
## PAY_AMT6   PAY_AMT6 4.5779840
## BILL_AMT5 BILL_AMT5 4.3695503
## PAY_3         PAY_3 2.7567291
## PAY_2         PAY_2 2.3900478
## EDUCATION EDUCATION 2.3548279
## PAY_6         PAY_6 2.1636061
## PAY_5         PAY_5 2.0757142
## PAY_4         PAY_4 1.7290783
## SEX             SEX 1.0655085
## MARRIAGE   MARRIAGE 0.9733251
```

```r
## Predict on test set
pred_bt <- predict(m3,test_bt,type="response",n.trees=5000)
pred_bt_class <- ifelse(pred_bt>0.5,1,0)
pred_bt_class <- factor(pred_bt_class,levels=c(0,1))
test_bt$Default <- factor(test_bt$Default,levels=c(0,1))
mean(pred_bt_class!=test_bt$Default)
```

```
## [1] 0.2435105
```

```r
pred_bt_ <- prediction(pred_bt,truth)
perf_bt <- performance(pred_bt_,"tpr","fpr")

plot(perf_bt,colorize=T,main="Boosting")
```




```r
AUC_bt <- performance(pred_bt_,"auc")@y.values[[1]]
AUC_bt
```

```
## [1] 0.6879928
```

Impute the missings by iterative regression

This section is developed with reference to ChatGPT and the materials from Lab 12.

```r
data_iter <- data_raw
data_iter <- subset(data_iter,select=-ID)

# First impute variables
var_num <- names(data_iter)[sapply(data_iter,is.numeric)]
var_cat <- names(data_iter)[sapply(data_iter,is.factor)]

for (v in var_num){
  data_iter[[v]][is.na(data_iter[[v]])] <- mean(data_iter[[v]],na.rm=T)
}

f1 <- function(x)names(which.max(table(x)))
for (v in var_cat){
  data_iter[[v]][is.na(data_iter[[v]])] <- f1(data_iter[[v]])
}

library(nnet)
```

```
## Warning: package 'nnet' was built under R version 4.4.3
```

```r
n_iter <- 10

for (i in 1:n_iter){
  for (v in var_num){
    missing_index <- is.na(data_iter[[v]])
    if(any(missing_index)){
      f2 <- as.formula(paste(v,"~."))
      model_num <- lm(f2,data=data_iter,subset=!missing_index)
      pred <- predict(model_num,data_iter[missing_index,])
      data_iter[[v]][missing_index] <- pred
    }
  }

  for(v in var_cat){
    missing_index <- is.na(data_iter[[v]])
    if(!any(missing_index)) next
    levels_count <- length(levels(data_iter[[v]]))

    # 2-level
    if(levels_count == 2){
      f3 <- as.formula(paste(v,"~."))
      model_bin <- glm(f3,data=data_iter,
                       family=binomial,
                       subset=!missing_index)
      p <- predict(model_bin,data_iter[missing_index,],type="response")
      pred_class <- ifelse(p > 0.5,
                           levels(data_iter[[v]])[2],
                           levels(data_iter[[v]])[1])
      data_iter[[v]][missing_index] <- pred_class
    }

    # multinomial
    else {
```

```
      f4 <- as.formula(paste(v,"~."))
      model_mul <- multinom(f4,data=data_iter,subset=!missing_index,trace=FALSE)
      pred_class <- predict(model_mul,data_iter[missing_index,])
      data_iter[[v]][missing_index] <- pred_class
    }
  }
}

summary(data_iter)
```

```
##    LIMIT_BAL       SEX      EDUCATION MARRIAGE      AGE            PAY_0
##  Min.   :  10000   1:1245   1:1122    0:   7   Min.   :21.00   0      :1416
##  1st Qu.:  50000   2:1720   2:1332    1:1189   1st Qu.:28.00   -1     : 623
##  Median : 150000            3: 489    2:1732   Median :34.00   1      : 403
##  Mean   : 163030            4:   7    3:  37   Mean   :35.34   2      : 266
##  3rd Qu.: 230000            5:  10             3rd Qu.:41.00   -2     : 221
##  Max.   :1000000            6:   5             Max.   :75.00   3      :  20
##                                                                (Other):  16
##      PAY_2          PAY_3          PAY_4          PAY_5          PAY_6
##  0      :1533   0      :1514   0      :1632   0      :1615   0      :1525
##  -1     : 635   -1     : 648   -1     : 603   -1     : 605   -1     : 655
##  2      : 391   2      : 387   -2     : 423   -2     : 436   -2     : 465
##  -2     : 359   -2     : 376   2      : 263   2      : 275   2      : 284
##  3      :  32   3      :  10   3      :  22   3      :  14   3      :  23
##  7      :   6   4      :   9   5      :   8   4      :  12   6      :   5
##  (Other):   9   (Other):  21   (Other):  14   (Other):   8   (Other):   8
##    BILL_AMT1        BILL_AMT2        BILL_AMT3        BILL_AMT4
##  Min.   :-14386   Min.   :-24704   Min.   :-15000   Min.   :-15000
##  1st Qu.:  3097   1st Qu.:  2894   1st Qu.:  2400   1st Qu.:  1770
##  Median : 21148   Median : 20313   Median : 19476   Median : 17647
##  Mean   : 49850   Mean   : 47836   Mean   : 44432   Mean   : 40224
##  3rd Qu.: 61189   3rd Qu.: 59011   3rd Qu.: 53978   3rd Qu.: 48443
##  Max.   :964511   Max.   :983931   Max.   :548020   Max.   :891586
##
##    BILL_AMT5        BILL_AMT6          PAY_AMT1         PAY_AMT2
##  Min.   :-28335   Min.   :-339603   Min.   :     0   Min.   :     0
##  1st Qu.:  1483   1st Qu.:    990   1st Qu.:   958   1st Qu.:   630
##  Median : 17314   Median :  15571   Median :  2081   Median :  2000
##  Mean   : 38808   Mean   :  37582   Mean   :  5397   Mean   :  5275
##  3rd Qu.: 47751   3rd Qu.:  46769   3rd Qu.:  5000   3rd Qu.:  4882
##  Max.   :927171   Max.   : 961664   Max.   :368199   Max.   :344261
##
##     PAY_AMT3         PAY_AMT4         PAY_AMT5         PAY_AMT6      Default
##  Min.   :     0   Min.   :     0   Min.   :     0   Min.   :     0   0:2311
##  1st Qu.:   215   1st Qu.:   218   1st Qu.:   264   1st Qu.:     0   1: 654
##  Median :  1355   Median :  1444   Median :  1500   Median :  1294
##  Mean   :  4711   Mean   :  4616   Mean   :  4818   Mean   :  5039
##  3rd Qu.:  3912   3rd Qu.:  4000   3rd Qu.:  4000   3rd Qu.:  3997
##  Max.   :896040   Max.   :205000   Max.   :332000   Max.   :528666
##
```

```r
sum(is.na(data_iter))
```

```
## [1] 0
```

Repeat the analyses on the newly imputed dataset

Logistic regression

```r
set.seed(1234)

train_idx1 <- sample(1:nrow(data_iter),round(0.7*nrow(data_iter)))
train1 <- data_iter[train_idx1,]
test1 <- data_iter[-train_idx1,]

logit_full1 <- glm(Default ~ .,data=train1,family="binomial")
summary(logit_full1)
```

```
##
## Call:
## glm(formula = Default ~ ., family = "binomial", data = train1)
##
## Coefficients: (7 not defined because of singularities)
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.591e+01  8.866e+02  -0.018  0.98569
## LIMIT_BAL   -8.750e-07  6.585e-07  -1.329  0.18389
## SEX2        -7.721e-02  1.237e-01  -0.624  0.53242
## EDUCATION2  -1.146e-01  1.414e-01  -0.811  0.41758
## EDUCATION3   5.316e-02  1.914e-01   0.278  0.78118
## EDUCATION4  -1.476e+01  8.809e+02  -0.017  0.98663
## EDUCATION5  -8.976e-01  1.148e+00  -0.782  0.43410
## EDUCATION6  -1.382e+01  1.170e+03  -0.012  0.99058
## MARRIAGE1    1.492e+01  8.866e+02   0.017  0.98658
## MARRIAGE2    1.459e+01  8.866e+02   0.016  0.98687
## MARRIAGE3    1.390e+01  8.866e+02   0.016  0.98749
## AGE         -6.555e-03  7.582e-03  -0.865  0.38731
## PAY_0-1      1.175e+00  4.294e-01   2.736  0.00623 **
## PAY_00      -3.010e-02  4.662e-01  -0.065  0.94852
## PAY_01       1.189e+00  3.339e-01   3.561  0.00037 ***
## PAY_02       2.179e+00  4.208e-01   5.178 2.24e-07 ***
## PAY_03       1.470e+00  6.766e-01   2.173  0.02976 *
## PAY_04       2.690e+00  1.322e+00   2.034  0.04196 *
## PAY_07       4.326e-01  2.935e+03   0.000  0.99988
## PAY_08      -3.171e+01  4.475e+03  -0.007  0.99435
## PAY_2-1     -7.074e-01  4.087e-01  -1.731  0.08349 .
## PAY_20      -2.646e-01  5.093e-01  -0.520  0.60340
## PAY_21      -1.541e+01  2.400e+03  -0.006  0.99487
## PAY_22      -5.733e-01  4.304e-01  -1.332  0.18282
## PAY_23      -7.325e-01  7.479e-01  -0.979  0.32739
## PAY_24       1.584e+01  1.394e+03   0.011  0.99094
## PAY_25      -4.906e+01  3.777e+03  -0.013  0.98964
## PAY_26       3.255e+01  3.393e+03   0.010  0.99235
```

```
## PAY_27              NA         NA      NA       NA
## PAY_3-1     -1.390e-01  3.734e-01  -0.372  0.70963
## PAY_30       8.238e-02  4.475e-01   0.184  0.85395
## PAY_31              NA         NA      NA       NA
## PAY_32       7.788e-01  4.526e-01   1.720  0.08535 .
## PAY_33      -7.120e-01  1.478e+00  -0.482  0.63004
## PAY_34       6.778e-01  1.443e+00   0.470  0.63866
## PAY_35      -4.783e+01  3.777e+03  -0.013  0.98990
## PAY_36              NA         NA      NA       NA
## PAY_37       3.243e+01  2.917e+03   0.011  0.99113
## PAY_4-1      3.468e-02  3.958e-01   0.088  0.93019
## PAY_40       1.945e-01  4.455e-01   0.437  0.66239
## PAY_42       2.032e-01  4.928e-01   0.412  0.68010
## PAY_43       4.902e-01  9.635e-01   0.509  0.61087
## PAY_44       3.214e+01  2.917e+03   0.011  0.99121
## PAY_45       3.159e+01  4.475e+03   0.007  0.99437
## PAY_46      -1.392e+01  3.777e+03  -0.004  0.99706
## PAY_47              NA         NA      NA       NA
## PAY_5-1      3.116e-01  3.926e-01   0.794  0.42734
## PAY_50       2.192e-01  4.468e-01   0.490  0.62379
## PAY_52       1.059e+00  4.964e-01   2.134  0.03286 *
## PAY_53       8.582e-01  1.155e+00   0.743  0.45762
## PAY_54       2.420e-01  1.919e+00   0.126  0.89967
## PAY_55              NA         NA      NA       NA
## PAY_57              NA         NA      NA       NA
## PAY_6-1     -4.439e-01  3.145e-01  -1.411  0.15813
## PAY_60      -6.013e-01  3.422e-01  -1.757  0.07885 .
## PAY_62      -6.991e-01  4.097e-01  -1.706  0.08794 .
## PAY_63      -2.776e-01  9.775e-01  -0.284  0.77640
## PAY_64       1.413e+01  1.696e+03   0.008  0.99335
## PAY_66      -1.587e+01  1.658e+03  -0.010  0.99237
## PAY_67      -1.574e+01  3.229e+03  -0.005  0.99611
## PAY_68              NA         NA      NA       NA
## BILL_AMT1   -6.240e-07  3.723e-06  -0.168  0.86687
## BILL_AMT2   -7.746e-07  5.347e-06  -0.145  0.88481
## BILL_AMT3    9.999e-06  4.890e-06   2.045  0.04087 *
## BILL_AMT4   -8.793e-07  4.325e-06  -0.203  0.83889
## BILL_AMT5    5.062e-06  6.106e-06   0.829  0.40708
## BILL_AMT6   -9.931e-06  4.451e-06  -2.231  0.02566 *
## PAY_AMT1    -4.561e-05  1.562e-05  -2.920  0.00350 **
## PAY_AMT2    -1.591e-05  8.232e-06  -1.932  0.05334 .
## PAY_AMT3     5.747e-06  4.403e-06   1.305  0.19179
## PAY_AMT4    -3.059e-06  7.028e-06  -0.435  0.66340
## PAY_AMT5     2.809e-06  6.573e-06   0.427  0.66912
## PAY_AMT6    -5.235e-06  4.622e-06  -1.133  0.25731
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2188.4  on 2075  degrees of freedom
## Residual deviance: 1787.4  on 2010  degrees of freedom
## AIC: 1919.4
##
```

```
## Number of Fisher Scoring iterations: 15
```

```
# Backward stepwise
model_back1 <- step(logit_full1,direction="backward")
```

```
## Start:  AIC=1919.37
## Default ~ LIMIT_BAL + SEX + EDUCATION + MARRIAGE + AGE + PAY_0 +
##     PAY_2 + PAY_3 + PAY_4 + PAY_5 + PAY_6 + BILL_AMT1 + BILL_AMT2 +
##     BILL_AMT3 + BILL_AMT4 + BILL_AMT5 + BILL_AMT6 + PAY_AMT1 +
##     PAY_AMT2 + PAY_AMT3 + PAY_AMT4 + PAY_AMT5 + PAY_AMT6
##
##             Df Deviance    AIC
## - PAY_4      5   1787.9 1909.9
## - PAY_6      7   1792.3 1910.3
## - EDUCATION  5   1791.8 1913.8
## - PAY_2      6   1797.3 1917.3
## - BILL_AMT2  1   1787.4 1917.4
## - BILL_AMT1  1   1787.4 1917.4
## - BILL_AMT4  1   1787.4 1917.4
## - PAY_AMT5   1   1787.5 1917.5
## - PAY_AMT4   1   1787.6 1917.6
## - SEX        1   1787.8 1917.8
## - BILL_AMT5  1   1788.1 1918.1
## - AGE        1   1788.1 1918.1
## - PAY_AMT3   1   1788.6 1918.6
## - PAY_AMT6   1   1788.7 1918.7
## - LIMIT_BAL  1   1789.2 1919.2
## - PAY_5      5   1797.2 1919.2
## <none>           1787.4 1919.4
## - BILL_AMT3  1   1791.8 1921.8
## - PAY_3      5   1800.2 1922.2
## - PAY_AMT2   1   1792.3 1922.3
## - BILL_AMT6  1   1792.3 1922.3
## - MARRIAGE   3   1796.9 1922.9
## - PAY_AMT1   1   1801.3 1931.3
## - PAY_0      7   1874.2 1992.2
##
## Step:  AIC=1909.89
## Default ~ LIMIT_BAL + SEX + EDUCATION + MARRIAGE + AGE + PAY_0 +
##     PAY_2 + PAY_3 + PAY_5 + PAY_6 + BILL_AMT1 + BILL_AMT2 + BILL_AMT3 +
##     BILL_AMT4 + BILL_AMT5 + BILL_AMT6 + PAY_AMT1 + PAY_AMT2 +
##     PAY_AMT3 + PAY_AMT4 + PAY_AMT5 + PAY_AMT6
##
##             Df Deviance    AIC
## - PAY_6      8   1792.9 1898.9
## - EDUCATION  5   1792.3 1904.3
## - BILL_AMT2  1   1787.9 1907.9
## - BILL_AMT1  1   1787.9 1907.9
## - BILL_AMT4  1   1787.9 1907.9
## - PAY_AMT5   1   1788.1 1908.1
## - PAY_AMT4   1   1788.1 1908.1
## - PAY_2      7   1800.3 1908.3
## - SEX        1   1788.3 1908.3
## - AGE        1   1788.6 1908.6
```

```
## - BILL_AMT5  1   1788.6 1908.6
## - PAY_AMT3   1   1788.9 1908.9
## - PAY_AMT6   1   1789.2 1909.2
## - LIMIT_BAL  1   1789.8 1909.8
## <none>           1787.9 1909.9
## - PAY_5      6   1801.9 1911.9
## - BILL_AMT3  1   1792.4 1912.4
## - PAY_AMT2   1   1792.6 1912.6
## - BILL_AMT6  1   1792.8 1912.8
## - MARRIAGE   3   1797.3 1913.3
## - PAY_3      6   1806.6 1916.6
## - PAY_AMT1   1   1802.1 1922.1
## - PAY_0      7   1874.8 1982.8
##
## Step:  AIC=1898.91
## Default ~ LIMIT_BAL + SEX + EDUCATION + MARRIAGE + AGE + PAY_0 +
##     PAY_2 + PAY_3 + PAY_5 + BILL_AMT1 + BILL_AMT2 + BILL_AMT3 +
##     BILL_AMT4 + BILL_AMT5 + BILL_AMT6 + PAY_AMT1 + PAY_AMT2 +
##     PAY_AMT3 + PAY_AMT4 + PAY_AMT5 + PAY_AMT6
##
##             Df Deviance    AIC
## - EDUCATION  5   1797.6 1893.6
## - BILL_AMT4  1   1792.9 1896.9
## - BILL_AMT1  1   1792.9 1896.9
## - BILL_AMT2  1   1792.9 1896.9
## - PAY_AMT5   1   1793.0 1897.0
## - PAY_AMT4   1   1793.2 1897.2
## - BILL_AMT5  1   1793.4 1897.4
## - SEX        1   1793.4 1897.4
## - PAY_2      7   1805.5 1897.5
## - AGE        1   1793.7 1897.7
## - PAY_AMT3   1   1794.0 1898.0
## - PAY_AMT6   1   1794.4 1898.4
## - LIMIT_BAL  1   1794.6 1898.6
## <none>           1792.9 1898.9
## - BILL_AMT3  1   1797.4 1901.4
## - PAY_AMT2   1   1797.7 1901.7
## - PAY_5      6   1808.0 1902.0
## - BILL_AMT6  1   1798.2 1902.2
## - MARRIAGE   3   1802.9 1902.9
## - PAY_3      6   1811.7 1905.7
## - PAY_AMT1   1   1807.4 1911.4
## - PAY_0      7   1880.2 1972.2
##
## Step:  AIC=1893.62
## Default ~ LIMIT_BAL + SEX + MARRIAGE + AGE + PAY_0 + PAY_2 +
##     PAY_3 + PAY_5 + BILL_AMT1 + BILL_AMT2 + BILL_AMT3 + BILL_AMT4 +
##     BILL_AMT5 + BILL_AMT6 + PAY_AMT1 + PAY_AMT2 + PAY_AMT3 +
##     PAY_AMT4 + PAY_AMT5 + PAY_AMT6
##
##             Df Deviance    AIC
## - BILL_AMT4  1   1797.6 1891.6
## - BILL_AMT2  1   1797.6 1891.6
## - BILL_AMT1  1   1797.6 1891.6
```

```
## - PAY_AMT5   1   1797.8 1891.8
## - PAY_AMT4   1   1797.9 1891.9
## - SEX        1   1798.1 1892.1
## - AGE        1   1798.2 1892.2
## - BILL_AMT5  1   1798.3 1892.3
## - PAY_2      7   1810.6 1892.6
## - PAY_AMT3   1   1798.6 1892.6
## - PAY_AMT6   1   1799.3 1893.3
## - LIMIT_BAL  1   1799.4 1893.4
## <none>           1797.6 1893.6
## - BILL_AMT3  1   1801.7 1895.7
## - PAY_AMT2   1   1802.2 1896.2
## - MARRIAGE   3   1807.1 1897.1
## - PAY_5      6   1813.1 1897.1
## - BILL_AMT6  1   1803.3 1897.3
## - PAY_3      6   1816.9 1900.9
## - PAY_AMT1   1   1812.4 1906.4
## - PAY_0      7   1885.3 1967.3
##
## Step:  AIC=1891.62
## Default ~ LIMIT_BAL + SEX + MARRIAGE + AGE + PAY_0 + PAY_2 +
##     PAY_3 + PAY_5 + BILL_AMT1 + BILL_AMT2 + BILL_AMT3 + BILL_AMT5 +
##     BILL_AMT6 + PAY_AMT1 + PAY_AMT2 + PAY_AMT3 + PAY_AMT4 + PAY_AMT5 +
##     PAY_AMT6
##
##             Df Deviance    AIC
## - BILL_AMT2  1   1797.6 1889.6
## - BILL_AMT1  1   1797.6 1889.6
## - PAY_AMT5   1   1797.8 1889.8
## - PAY_AMT4   1   1797.9 1889.9
## - SEX        1   1798.1 1890.1
## - AGE        1   1798.2 1890.2
## - BILL_AMT5  1   1798.4 1890.4
## - PAY_2      7   1810.6 1890.6
## - PAY_AMT3   1   1798.7 1890.7
## - PAY_AMT6   1   1799.3 1891.3
## - LIMIT_BAL  1   1799.4 1891.4
## <none>           1797.6 1891.6
## - PAY_AMT2   1   1802.2 1894.2
## - BILL_AMT3  1   1802.2 1894.2
## - MARRIAGE   3   1807.1 1895.1
## - PAY_5      6   1813.2 1895.2
## - BILL_AMT6  1   1803.3 1895.3
## - PAY_3      6   1817.0 1899.0
## - PAY_AMT1   1   1812.4 1904.4
## - PAY_0      7   1885.3 1965.3
##
## Step:  AIC=1889.63
## Default ~ LIMIT_BAL + SEX + MARRIAGE + AGE + PAY_0 + PAY_2 +
##     PAY_3 + PAY_5 + BILL_AMT1 + BILL_AMT3 + BILL_AMT5 + BILL_AMT6 +
##     PAY_AMT1 + PAY_AMT2 + PAY_AMT3 + PAY_AMT4 + PAY_AMT5 + PAY_AMT6
##
##             Df Deviance    AIC
## - BILL_AMT1  1   1797.7 1887.7
```

```
## - PAY_AMT5   1   1797.8 1887.8
## - PAY_AMT4   1   1797.9 1887.9
## - SEX        1   1798.1 1888.1
## - AGE        1   1798.2 1888.2
## - BILL_AMT5  1   1798.4 1888.4
## - PAY_2      7   1810.6 1888.6
## - PAY_AMT3   1   1798.7 1888.7
## - PAY_AMT6   1   1799.3 1889.3
## - LIMIT_BAL  1   1799.4 1889.4
## <none>           1797.6 1889.6
## - PAY_AMT2   1   1802.5 1892.5
## - MARRIAGE   3   1807.1 1893.1
## - PAY_5      6   1813.2 1893.2
## - BILL_AMT6  1   1803.3 1893.3
## - BILL_AMT3  1   1804.2 1894.2
## - PAY_3      6   1817.0 1897.0
## - PAY_AMT1   1   1814.8 1904.8
## - PAY_0      7   1885.7 1963.7
##
## Step:  AIC=1887.71
## Default ~ LIMIT_BAL + SEX + MARRIAGE + AGE + PAY_0 + PAY_2 +
##     PAY_3 + PAY_5 + BILL_AMT3 + BILL_AMT5 + BILL_AMT6 + PAY_AMT1 +
##     PAY_AMT2 + PAY_AMT3 + PAY_AMT4 + PAY_AMT5 + PAY_AMT6
##
##             Df Deviance    AIC
## - PAY_AMT5   1   1797.8 1885.8
## - PAY_AMT4   1   1798.0 1886.0
## - SEX        1   1798.2 1886.2
## - AGE        1   1798.2 1886.2
## - BILL_AMT5  1   1798.5 1886.5
## - PAY_AMT3   1   1798.7 1886.7
## - PAY_2      7   1810.8 1886.8
## - PAY_AMT6   1   1799.3 1887.3
## - LIMIT_BAL  1   1799.6 1887.6
## <none>           1797.7 1887.7
## - PAY_AMT2   1   1802.6 1890.6
## - MARRIAGE   3   1807.2 1891.2
## - BILL_AMT6  1   1803.3 1891.3
## - PAY_5      6   1813.4 1891.4
## - BILL_AMT3  1   1806.8 1894.8
## - PAY_3      6   1817.3 1895.3
## - PAY_AMT1   1   1814.8 1902.8
## - PAY_0      7   1885.7 1961.7
##
## Step:  AIC=1885.85
## Default ~ LIMIT_BAL + SEX + MARRIAGE + AGE + PAY_0 + PAY_2 +
##     PAY_3 + PAY_5 + BILL_AMT3 + BILL_AMT5 + BILL_AMT6 + PAY_AMT1 +
##     PAY_AMT2 + PAY_AMT3 + PAY_AMT4 + PAY_AMT6
##
##             Df Deviance    AIC
## - PAY_AMT4   1   1798.1 1884.1
## - SEX        1   1798.3 1884.3
## - AGE        1   1798.4 1884.4
## - BILL_AMT5  1   1798.5 1884.5
```

```
## - PAY_2      7   1810.8 1884.8
## - PAY_AMT3   1   1798.9 1884.9
## - PAY_AMT6   1   1799.4 1885.4
## - LIMIT_BAL  1   1799.7 1885.7
## <none>           1797.8 1885.8
## - PAY_AMT2   1   1802.7 1888.7
## - MARRIAGE   3   1807.4 1889.4
## - PAY_5      6   1813.5 1889.5
## - BILL_AMT6  1   1804.1 1890.1
## - BILL_AMT3  1   1807.3 1893.3
## - PAY_3      6   1817.5 1893.5
## - PAY_AMT1   1   1814.8 1900.8
## - PAY_0      7   1886.0 1960.0
##
## Step:  AIC=1884.11
## Default ~ LIMIT_BAL + SEX + MARRIAGE + AGE + PAY_0 + PAY_2 +
##     PAY_3 + PAY_5 + BILL_AMT3 + BILL_AMT5 + BILL_AMT6 + PAY_AMT1 +
##     PAY_AMT2 + PAY_AMT3 + PAY_AMT6
##
##             Df Deviance    AIC
## - BILL_AMT5  1   1798.5 1882.5
## - SEX        1   1798.6 1882.6
## - AGE        1   1798.7 1882.7
## - PAY_2      7   1811.2 1883.2
## - PAY_AMT3   1   1799.3 1883.3
## - PAY_AMT6   1   1799.5 1883.5
## - LIMIT_BAL  1   1800.0 1884.0
## <none>           1798.1 1884.1
## - PAY_AMT2   1   1803.5 1887.5
## - MARRIAGE   3   1807.8 1887.8
## - BILL_AMT6  1   1804.1 1888.1
## - PAY_5      6   1814.2 1888.2
## - PAY_3      6   1817.7 1891.7
## - BILL_AMT3  1   1810.4 1894.4
## - PAY_AMT1   1   1815.4 1899.4
## - PAY_0      7   1886.6 1958.6
##
## Step:  AIC=1882.51
## Default ~ LIMIT_BAL + SEX + MARRIAGE + AGE + PAY_0 + PAY_2 +
##     PAY_3 + PAY_5 + BILL_AMT3 + BILL_AMT6 + PAY_AMT1 + PAY_AMT2 +
##     PAY_AMT3 + PAY_AMT6
##
##             Df Deviance    AIC
## - SEX        1   1799.0 1881.0
## - AGE        1   1799.1 1881.1
## - PAY_2      7   1811.5 1881.5
## - PAY_AMT6   1   1799.6 1881.6
## - PAY_AMT3   1   1800.0 1882.0
## - LIMIT_BAL  1   1800.3 1882.3
## <none>           1798.5 1882.5
## - MARRIAGE   3   1808.3 1886.3
## - PAY_5      6   1814.7 1886.7
## - PAY_AMT2   1   1804.9 1886.9
## - PAY_3      6   1818.0 1890.0
```

```
## - BILL_AMT6  1   1808.3 1890.3
## - PAY_AMT1   1   1815.8 1897.8
## - BILL_AMT3  1   1816.7 1898.7
## - PAY_0      7   1887.1 1957.1
##
## Step:  AIC=1881.02
## Default ~ LIMIT_BAL + MARRIAGE + AGE + PAY_0 + PAY_2 + PAY_3 +
##     PAY_5 + BILL_AMT3 + BILL_AMT6 + PAY_AMT1 + PAY_AMT2 + PAY_AMT3 +
##     PAY_AMT6
##
##             Df Deviance    AIC
## - AGE        1   1799.5 1879.5
## - PAY_2      7   1812.0 1880.0
## - PAY_AMT6   1   1800.0 1880.0
## - PAY_AMT3   1   1800.5 1880.5
## - LIMIT_BAL  1   1800.8 1880.8
## <none>           1799.0 1881.0
## - MARRIAGE   3   1808.7 1884.7
## - PAY_5      6   1815.1 1885.1
## - PAY_AMT2   1   1805.4 1885.4
## - PAY_3      6   1818.5 1888.5
## - BILL_AMT6  1   1808.9 1888.9
## - PAY_AMT1   1   1816.4 1896.4
## - BILL_AMT3  1   1817.3 1897.3
## - PAY_0      7   1887.3 1955.3
##
## Step:  AIC=1879.46
## Default ~ LIMIT_BAL + MARRIAGE + PAY_0 + PAY_2 + PAY_3 + PAY_5 +
##     BILL_AMT3 + BILL_AMT6 + PAY_AMT1 + PAY_AMT2 + PAY_AMT3 +
##     PAY_AMT6
##
##             Df Deviance    AIC
## - PAY_2      7   1812.3 1878.3
## - PAY_AMT6   1   1800.5 1878.5
## - PAY_AMT3   1   1801.0 1879.0
## - LIMIT_BAL  1   1801.4 1879.4
## <none>           1799.5 1879.5
## - MARRIAGE   3   1808.7 1882.7
## - PAY_5      6   1815.5 1883.5
## - PAY_AMT2   1   1805.9 1883.9
## - PAY_3      6   1819.0 1887.0
## - BILL_AMT6  1   1809.4 1887.4
## - PAY_AMT1   1   1816.7 1894.7
## - BILL_AMT3  1   1817.7 1895.7
## - PAY_0      7   1887.6 1953.6
##
## Step:  AIC=1878.33
## Default ~ LIMIT_BAL + MARRIAGE + PAY_0 + PAY_3 + PAY_5 + BILL_AMT3 +
##     BILL_AMT6 + PAY_AMT1 + PAY_AMT2 + PAY_AMT3 + PAY_AMT6
##
##             Df Deviance    AIC
## - PAY_AMT6   1   1813.3 1877.3
## - PAY_AMT3   1   1813.8 1877.8
## - LIMIT_BAL  1   1814.1 1878.1
```

```
## <none>           1812.3 1878.3
## - PAY_3      7   1828.9 1880.9
## - MARRIAGE   3   1822.0 1882.0
## - PAY_AMT2   1   1818.9 1882.9
## - PAY_5      6   1830.6 1884.6
## - BILL_AMT6  1   1822.5 1886.5
## - PAY_AMT1   1   1830.0 1894.0
## - BILL_AMT3  1   1830.8 1894.8
## - PAY_0      7   1926.9 1978.9
##
## Step:  AIC=1877.34
## Default ~ LIMIT_BAL + MARRIAGE + PAY_0 + PAY_3 + PAY_5 + BILL_AMT3 +
##     BILL_AMT6 + PAY_AMT1 + PAY_AMT2 + PAY_AMT3
##
##             Df Deviance    AIC
## - PAY_AMT3   1   1814.6 1876.6
## - LIMIT_BAL  1   1815.3 1877.3
## <none>           1813.3 1877.3
## - PAY_3      7   1829.9 1879.9
## - MARRIAGE   3   1823.0 1881.0
## - PAY_AMT2   1   1819.7 1881.7
## - PAY_5      6   1831.3 1883.3
## - BILL_AMT6  1   1822.5 1884.5
## - BILL_AMT3  1   1830.8 1892.8
## - PAY_AMT1   1   1832.0 1894.0
## - PAY_0      7   1927.6 1977.6
##
## Step:  AIC=1876.63
## Default ~ LIMIT_BAL + MARRIAGE + PAY_0 + PAY_3 + PAY_5 + BILL_AMT3 +
##     BILL_AMT6 + PAY_AMT1 + PAY_AMT2
##
##             Df Deviance    AIC
## - LIMIT_BAL  1   1816.4 1876.4
## <none>           1814.6 1876.6
## - PAY_3      7   1831.0 1879.0
## - MARRIAGE   3   1824.5 1880.5
## - PAY_AMT2   1   1820.7 1880.7
## - PAY_5      6   1832.6 1882.6
## - BILL_AMT6  1   1822.7 1882.7
## - BILL_AMT3  1   1831.0 1891.0
## - PAY_AMT1   1   1832.2 1892.2
## - PAY_0      7   1928.1 1976.1
##
## Step:  AIC=1876.39
## Default ~ MARRIAGE + PAY_0 + PAY_3 + PAY_5 + BILL_AMT3 + BILL_AMT6 +
##     PAY_AMT1 + PAY_AMT2
##
##             Df Deviance    AIC
## <none>           1816.4 1876.4
## - MARRIAGE   3   1825.5 1879.5
## - PAY_3      7   1834.0 1880.0
## - PAY_AMT2   1   1822.8 1880.8
## - PAY_5      6   1835.0 1883.0
## - BILL_AMT6  1   1825.3 1883.3
```

```
## - BILL_AMT3  1   1831.6 1889.6
## - PAY_AMT1   1   1836.5 1894.5
## - PAY_0      7   1930.0 1976.0
```

```
summary(model_back1)
```

```
##
## Call:
## glm(formula = Default ~ MARRIAGE + PAY_0 + PAY_3 + PAY_5 + BILL_AMT3 +
##     BILL_AMT6 + PAY_AMT1 + PAY_AMT2, family = "binomial", data = train1)
##
## Coefficients: (2 not defined because of singularities)
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.555e+01  5.350e+02  -0.029 0.976814
## MARRIAGE1    1.400e+01  5.350e+02   0.026 0.979121
## MARRIAGE2    1.374e+01  5.350e+02   0.026 0.979517
## MARRIAGE3    1.304e+01  5.350e+02   0.024 0.980564
## PAY_0-1      7.810e-01  3.445e-01   2.267 0.023370 *
## PAY_00      -5.872e-02  3.508e-01  -0.167 0.867071
## PAY_01       1.032e+00  3.206e-01   3.220 0.001283 **
## PAY_02       1.924e+00  3.703e-01   5.196 2.04e-07 ***
## PAY_03       1.101e+00  6.425e-01   1.714 0.086509 .
## PAY_04       2.525e+00  1.154e+00   2.188 0.028672 *
## PAY_07       1.646e+01  1.027e+03   0.016 0.987218
## PAY_08       8.519e-01  1.759e+00   0.484 0.628214
## PAY_3-1     -4.095e-01  2.747e-01  -1.490 0.136129
## PAY_30      -1.020e-01  2.980e-01  -0.342 0.732070
## PAY_31      -1.455e+01  1.455e+03  -0.010 0.992025
## PAY_32       5.044e-01  3.017e-01   1.672 0.094539 .
## PAY_33      -2.221e-01  9.652e-01  -0.230 0.818010
## PAY_34       1.566e-01  9.789e-01   0.160 0.872890
## PAY_35      -3.615e-01  1.480e+00  -0.244 0.807036
## PAY_36              NA         NA      NA       NA
## PAY_37       1.601e+01  6.189e+02   0.026 0.979368
## PAY_5-1     -5.038e-02  2.429e-01  -0.207 0.835691
## PAY_50      -7.100e-02  2.609e-01  -0.272 0.785521
## PAY_52       7.739e-01  3.051e-01   2.537 0.011189 *
## PAY_53       5.759e-01  8.762e-01   0.657 0.511000
## PAY_54      -7.705e-01  1.356e+00  -0.568 0.570030
## PAY_55       1.562e+01  1.455e+03   0.011 0.991438
## PAY_57              NA         NA      NA       NA
## BILL_AMT3    8.380e-06  2.111e-06   3.970 7.18e-05 ***
## BILL_AMT6   -6.638e-06  2.264e-06  -2.932 0.003365 **
## PAY_AMT1    -4.661e-05  1.408e-05  -3.309 0.000936 ***
## PAY_AMT2    -1.550e-05  7.399e-06  -2.095 0.036185 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2188.4  on 2075  degrees of freedom
## Residual deviance: 1816.4  on 2046  degrees of freedom
## AIC: 1876.4
##
```

```
## Number of Fisher Scoring iterations: 14
```

```
pred_back_prob1 <- predict(model_back1,newdata=test1,type="response")

# Convert to class
pred_back_class1 <- ifelse(pred_back_prob1>0.5,1,0)

# Truth
truth1 <- as.numeric(test1$Default)-1

acc_back1 <- mean(pred_back_class1 == truth1)
acc_back1
```

```
## [1] 0.7975253
```

```
library(ROCR)
pred_back1 <- prediction(pred_back_prob1,truth1)
AUC_back1 <- performance(pred_back1,"auc")@y.values[[1]]
AUC_back1
```

```
## [1] 0.7142433
```

```
# Forward stepwise (Failure)
logit_null1 <- glm(Default ~ 1,data=train1,family="binomial")
step(logit_null1,scope=list(lower=~1,upper=~.),direction="forward")
```

```
## Start:  AIC=2190.43
## Default ~ 1

##
## Call:  glm(formula = Default ~ 1, family = "binomial", data = train1)
##
## Coefficients:
## (Intercept)
##      -1.265
##
## Degrees of Freedom: 2075 Total (i.e. Null);  2075 Residual
## Null Deviance:       2188
## Residual Deviance: 2188  AIC: 2190
```

```
# LASSO logistic regression
library(glmnet)
x_train1 <- model.matrix(Default ~ .,train1)[,-1]
y1 <- as.numeric(train1$Default)-1

set.seed(123)
cvfit1 <- cv.glmnet(x_train1,y1,family="binomial",alpha=1)

best_lambda1 <- cvfit1$lambda.min
model_lasso1 <- glmnet(x_train1,y1,family="binomial",lambda=best_lambda1)
```

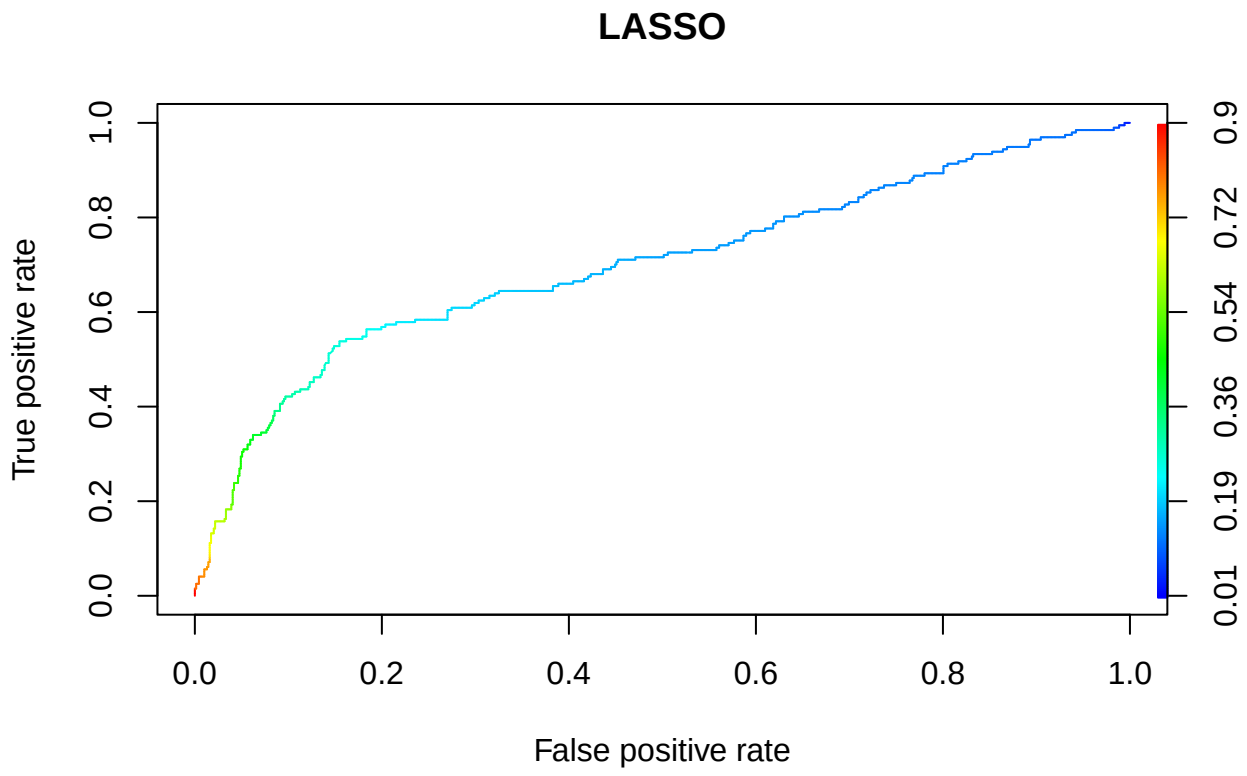
```
x_test1 <- model.matrix(Default ~ .,test1)[,-1]
pred_lasso_prob1 <- predict(model_lasso1,x_test1,type="response")

pred_lasso_class1 <- ifelse(pred_lasso_prob1>0.5,1,0)
acc_lasso1 <- mean(pred_lasso_class1 == truth1)
acc_lasso1
```

```
## [1] 0.7975253
```

```
pred_lasso1 <- prediction(pred_lasso_prob1,truth1)
perf_lasso1 <- performance(pred_lasso1,"tpr","fpr")

plot(perf_lasso1,colorize=T,main="LASSO")
```




```
AUC_lasso1 <- performance(pred_lasso1,"auc")@y.values[[1]]
AUC_lasso1
```

```
## [1] 0.700526
```

KNN

```r
library(class)
library(caret)
library(fastDummies)
set.seed(321)
train_knn1 <- train1
test_knn1 <- test1

train_knn1$Default <- as.numeric(train_knn1$Default) - 1
test_knn1$Default <- as.numeric(test_knn1$Default) - 1

# Create dummy variables
cat_vars1 <- names(train_knn1)[sapply(train_knn1,is.factor)]
train_dummy1 <- dummy_cols(train_knn1,remove_selected_columns = TRUE,remove_first_dummy = TRUE)
test_dummy1 <- dummy_cols(test_knn1,remove_selected_columns = TRUE,remove_first_dummy = TRUE)

train_y1 <- factor(train_dummy1$Default,levels = c(0,1))
test_y1 <- factor(test_dummy1$Default,levels = c(0,1))

train_dummy1$Default <- NULL
test_dummy1$Default  <- NULL

num_vars1 <- names(train_knn1)[sapply(train_knn1,is.numeric) & names(train_knn1) != "Default"]

preProcValues1 <- preProcess(train_knn1[,num_vars1],method = c("center", "scale"))
train_dummy_scaled1 <- train_dummy1
test_dummy_scaled1 <- test_dummy1

train_dummy_scaled1[,num_vars1] <- predict(preProcValues1,train_knn1[,num_vars1])
test_dummy_scaled1[,num_vars1] <- predict(preProcValues1,test_knn1[,num_vars1])

k_values1 <- seq(1,50,by=2)
acc_list1 <- c()

for (k in k_values1){
  pred_k1 <- knn(train_dummy_scaled1,test_dummy_scaled1,cl = train_y1,k = k)
  acc_list1 <- c(acc_list1,mean(pred_k1 == test_y1))
}

best_k1 <- k_values1[which.max(acc_list1)]
best_k1
```

```
## [1] 31
```

```r
pred_knn1 <- knn(train_dummy_scaled1,test_dummy_scaled1,cl = train_y1,k = best_k1)

# Accuracy
acc_knn1 <- mean(pred_knn1 == test_y1)
acc_knn1
```

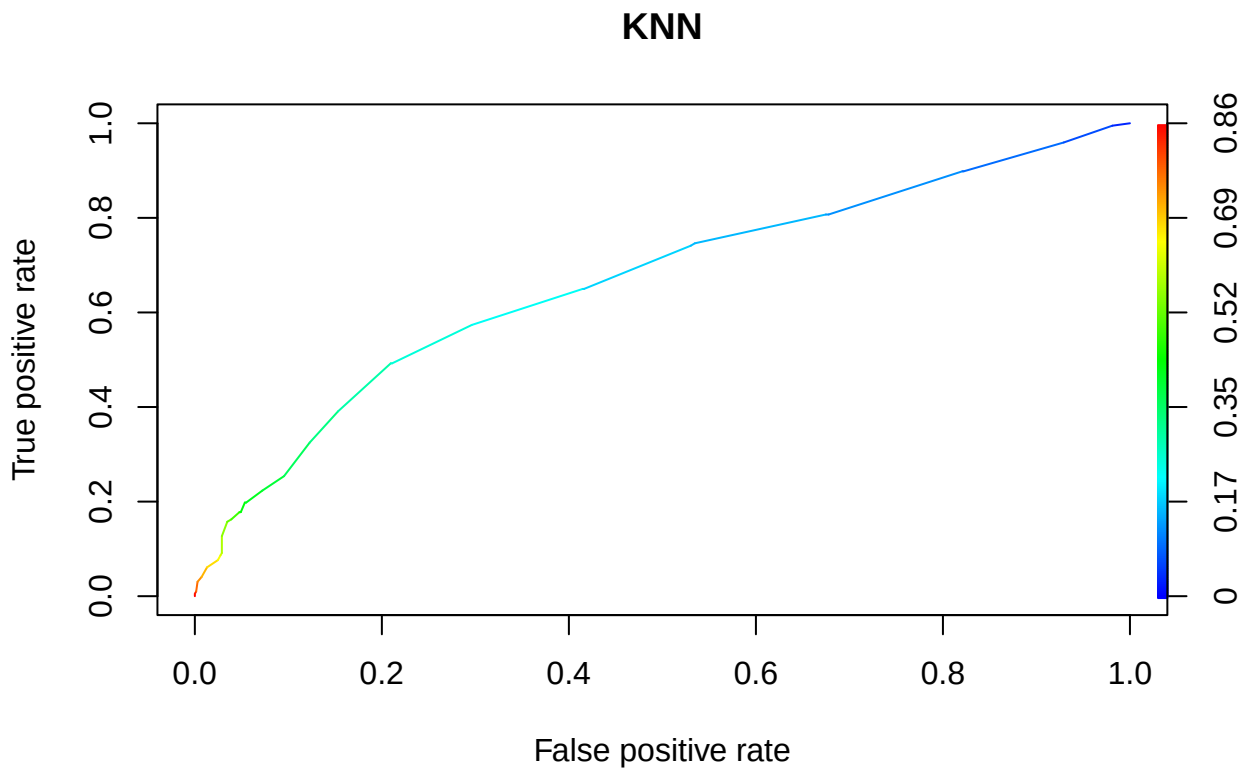
```
## [1] 0.7862767
```

```
knn_prob1 <- attributes(knn(train_dummy_scaled1, test_dummy_scaled1,
                            cl = train_y1, k = best_k1, prob = TRUE))$prob
knn_prob1 <- ifelse(pred_knn1 == "1", knn_prob1, 1 - knn_prob1)
knn_prob1 <- as.numeric(knn_prob1)

truth_knn1 <- as.numeric(test_y1) - 1
pred_knn_1 <- prediction(knn_prob1,truth_knn1)
AUC_knn1 <- performance(pred_knn_1,"auc")@y.values[[1]]
AUC_knn1
```

```
## [1] 0.6644538
```

```
# ROC curve
perf_knn1 <- performance(pred_knn_1,"tpr","fpr")
plot(perf_knn1,colorize = TRUE,main = "KNN")
```



Ensemble methods

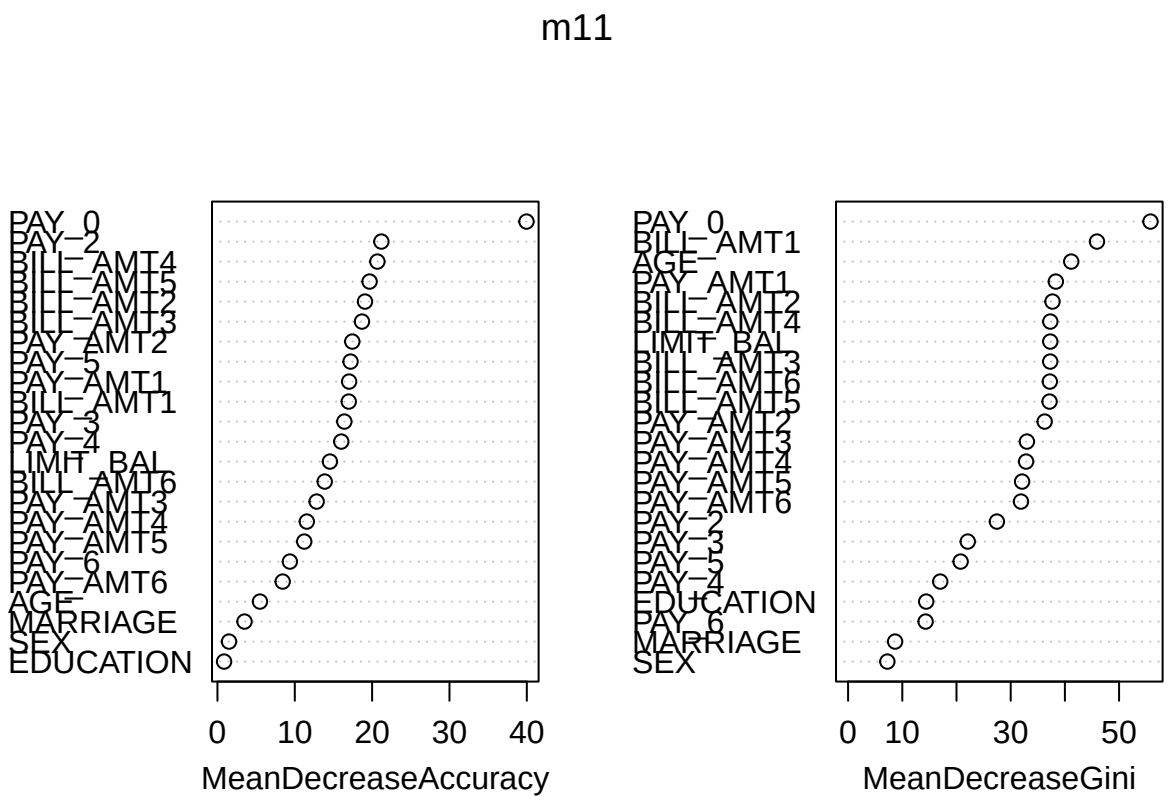
```
# Random forest
library(randomForest)
set.seed(1111)
m11 <- randomForest(Default~.,data=train1,mtry=floor(sqrt(ncol(train1)-1)),importance=T)
m11
```

```
##
## Call:
##  randomForest(formula = Default ~ ., data = train1, mtry = floor(sqrt(ncol(train1) -      1)), impor
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 4
##
##         OOB estimate of  error rate: 19.8%
## Confusion matrix:
##      0   1 class.error
## 0 1528  91  0.05620754
## 1  320 137  0.70021882
```

```
## Predict on test set
pred_rf1 <- predict(m11,newdata=test1,type="class")
mean(pred_rf1!=test1$Default)
```

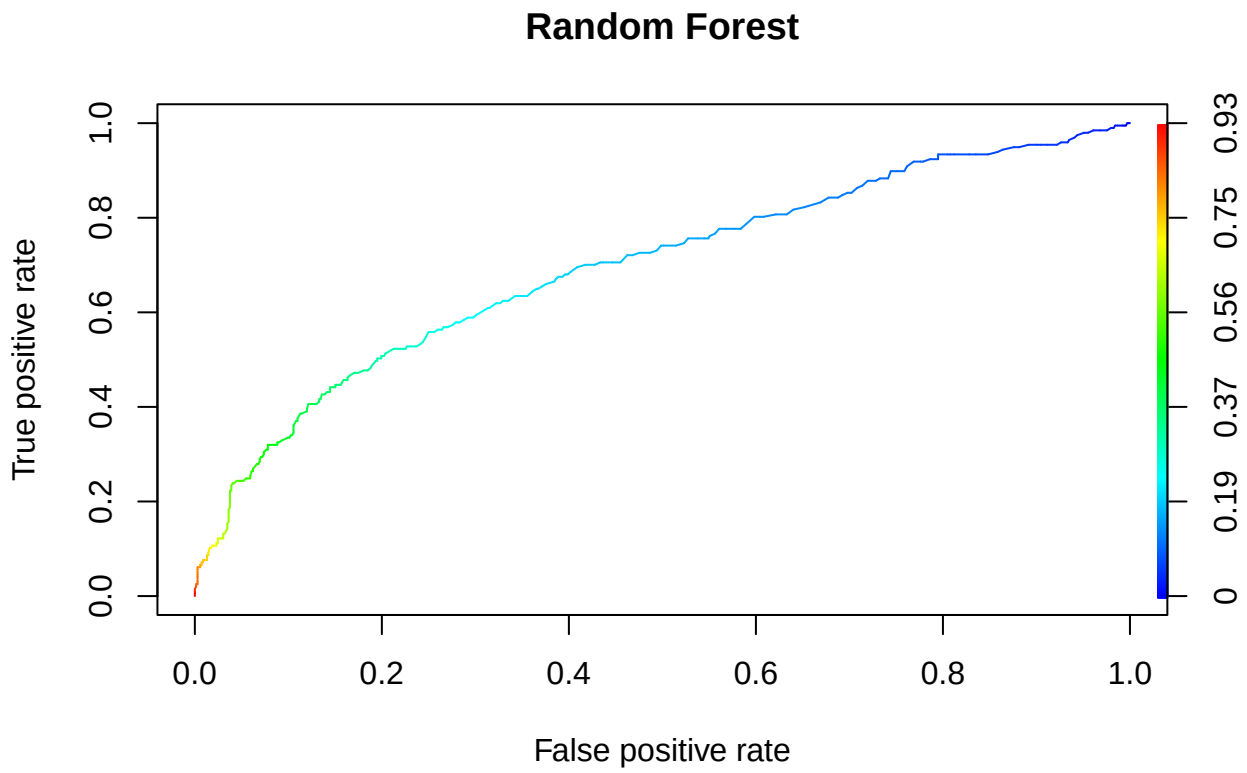
```
## [1] 0.2103487
```

```
varImpPlot(m11)
```




```
pred_rf_prob1 <- predict(m11,newdata=test1,type="prob")[,2]
pred_rf1 <- prediction(pred_rf_prob1,truth1)
perf_rf1 <- performance(pred_rf1,"tpr","fpr")
```

```
plot(perf_rf1,colorize=T,main="Random Forest")
```




```
AUC_rf1 <- performance(pred_rf1,"auc")@y.values[[1]]
AUC_rf1
```

```
## [1] 0.6949583
```

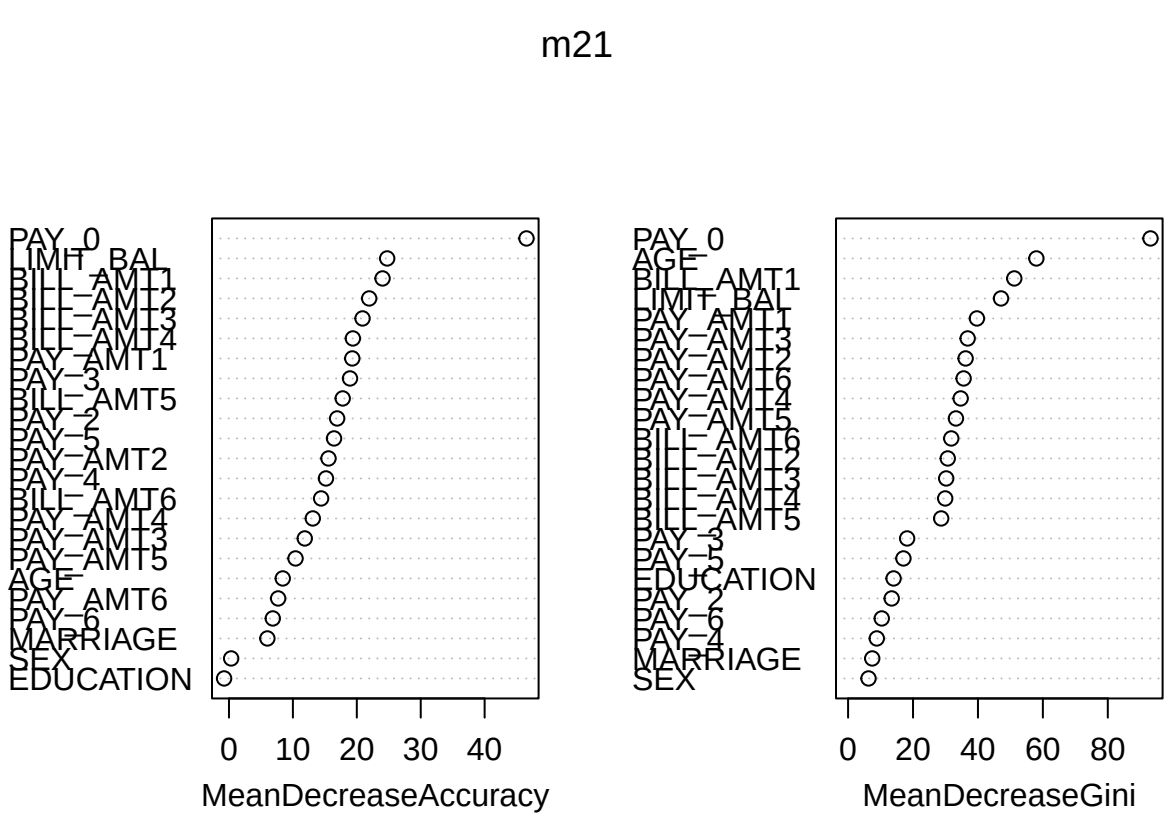
```
# Bagging
set.seed(2222)
m21 <- randomForest(Default ~ .,data=train1,mtry=ncol(train1)-1,importance=T)
m21
```

```
##
## Call:
##  randomForest(formula = Default ~ ., data = train1, mtry = ncol(train1) -      1, importance = T)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 23
##
##         OOB estimate of  error rate: 19.75%
## Confusion matrix:
##      0   1 class.error
## 0 1518 101  0.06238419
## 1  309 148  0.67614880
```

```
## Predict on test set
pred_bg1 <- predict(m21,newdata=test1,type="class")
mean(pred_bg1!=test1$Default)
```

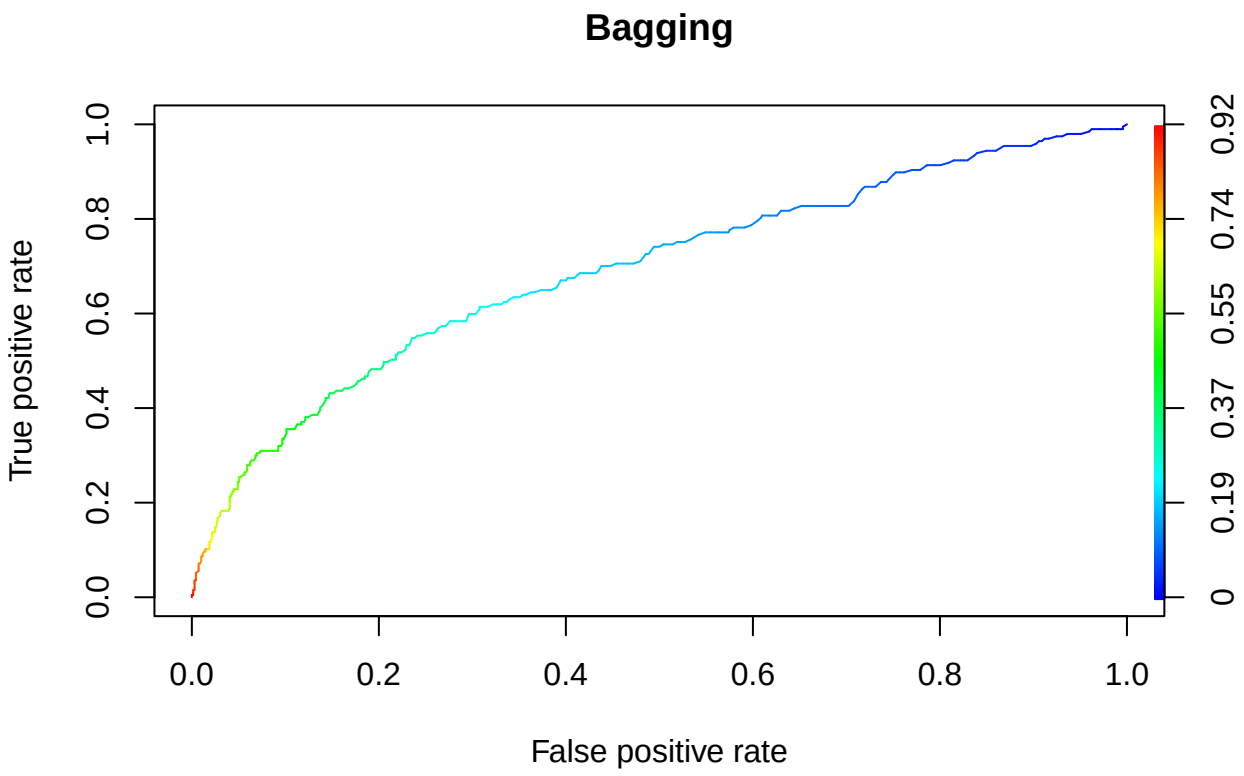
```
## [1] 0.208099
```

```
varImpPlot(m21)
```




```
pred_bg_prob1 <- predict(m21,newdata=test1,type="prob")[,2]
pred_bg1 <- prediction(pred_bg_prob1,truth1)
perf_bg1 <- performance(pred_bg1,"tpr","fpr")

plot(perf_bg1,colorize=T,main="Bagging")
```




```
AUC_bg1 <- performance(pred_bg1,"auc")@y.values[[1]]
AUC_bg1
```

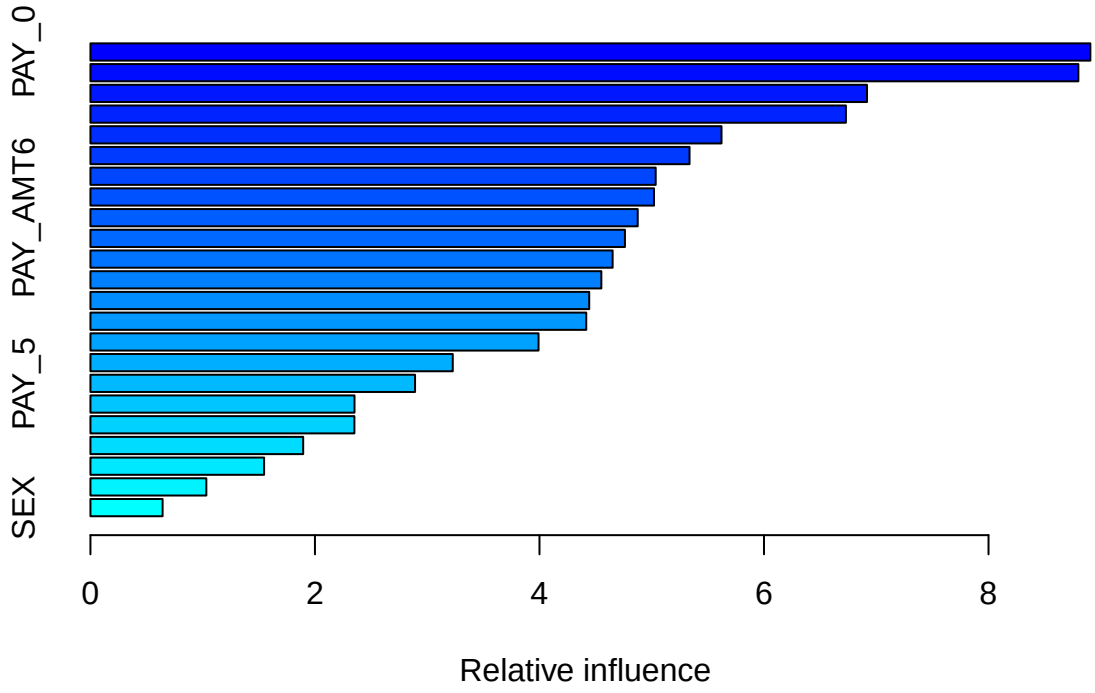
```
## [1] 0.6912136
```

```
# Boosting
library(gbm)
set.seed(3333)
train_bt1 <- train1
test_bt1 <- test1
train_bt1$Default <- as.numeric(train_bt1$Default)-1
test_bt1$Default <- as.numeric(test_bt1$Default)-1

m31 <- gbm(Default ~ .,data=train_bt1,distribution="bernoulli",n.trees=5000,interaction.depth=3,shrinka
m31
```

```
## gbm(formula = Default ~ ., distribution = "bernoulli", data = train_bt1,
##     n.trees = 5000, interaction.depth = 3, shrinkage = 0.1)
## A gradient boosted model with bernoulli loss function.
## 5000 iterations were performed.
## There were 23 predictors of which 23 had non-zero influence.
```

```
summary(m31)
```



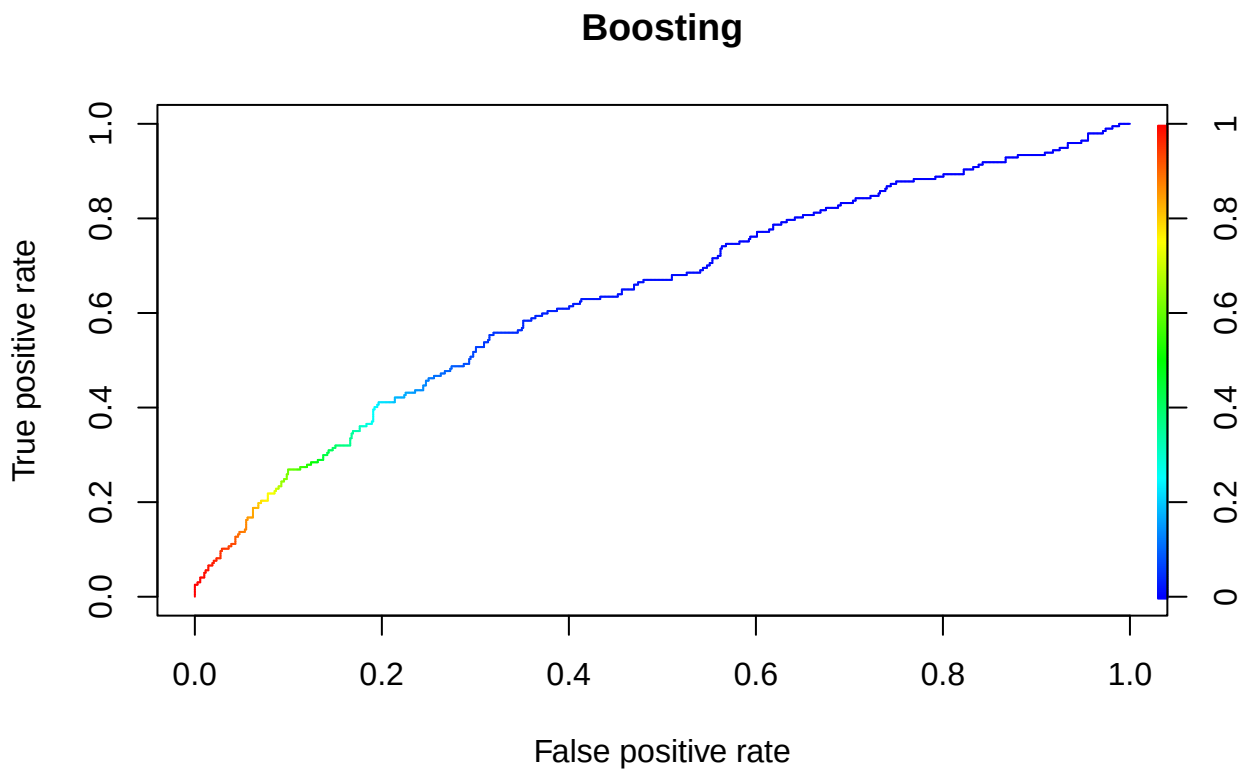

```
##                 var   rel.inf
## PAY_0         PAY_0 8.9073099
## BILL_AMT1 BILL_AMT1 8.8004332
## AGE             AGE 6.9176497
## LIMIT_BAL LIMIT_BAL 6.7297344
## PAY_AMT1   PAY_AMT1 5.6216232
## PAY_AMT2   PAY_AMT2 5.3368345
## PAY_AMT5   PAY_AMT5 5.0352465
## PAY_AMT3   PAY_AMT3 5.0203871
## PAY_AMT6   PAY_AMT6 4.8752476
## BILL_AMT2 BILL_AMT2 4.7614512
## BILL_AMT4 BILL_AMT4 4.6515340
## BILL_AMT6 BILL_AMT6 4.5510577
## PAY_AMT4   PAY_AMT4 4.4431537
## BILL_AMT3 BILL_AMT3 4.4168187
## BILL_AMT5 BILL_AMT5 3.9918212
## PAY_3         PAY_3 3.2275410
## PAY_5         PAY_5 2.8917564
## PAY_2         PAY_2 2.3522245
## PAY_4         PAY_4 2.3511270
## EDUCATION EDUCATION 1.8951651
## PAY_6         PAY_6 1.5468807
## MARRIAGE   MARRIAGE 1.0322893
## SEX             SEX 0.6427135
```

```
## Predict on test set
pred_bt1 <- predict(m31,test_bt1,type="response",n.trees=5000)
pred_bt_class1 <- ifelse(pred_bt1>0.5,1,0)
pred_bt_class1 <- factor(pred_bt_class1,levels=c(0,1))
test_bt1$Default <- factor(test_bt1$Default,levels=c(0,1))
mean(pred_bt_class1!=test_bt1$Default)
```

```
## [1] 0.2564679
```

```
pred_bt_1 <- prediction(pred_bt1,truth1)
perf_bt1 <- performance(pred_bt_1,"tpr","fpr")

plot(perf_bt1,colorize=T,main="Boosting")
```




```
AUC_bt1 <- performance(pred_bt_1,"auc")@y.values[[1]]
AUC_bt1
```

```
## [1] 0.6396306
```