Project 4: Predicting Default Risk

Load File

Building the Training Set

Removal

During the cleanup process, the following fields were removed for the following reasons:

- Guarantors: Low variability, due to only 43 'Yes' values and 457 'None' values.
- Duration-in-Current-address: Missing data, due to only 156 non-null values.
- Concurrent-Credits: Low variability, due to all 500 records having a value of 'Other Banks/Depts'.
- Occupation: Low variability, due to all 500 records having a value of 1.
- Telephone: Unrelated, as the telephone number does not impact the credit-worthiness of a lendee.
- Foreign-Worker: Low variability, due to only 19 2 values and 481 1 values. ## Imputation Because all of the null Age-years values are Creditworthy applications, and skew the variables credit-amount and no-credits-at-this-bank, they were imputed with the median Age.years value of 33.

Business and Data Understanding

A decision must be made on whether or not each of the 500 new loan applications can be approved. As a result, two main datasets must be gathered. The first must contain attributes that influence creditworthiness on all past applicants and the second must contain those same attributes for the new applicants. To determine the creditworthiness of each applicant from this data, binary classification models should be used, specifically logit regression, decision tree, random forest, and boosted models.

Train the Classification Models

Logit Regression Model

Significant Predictor Variables

Per the Stepwise Regression output below, the model with the lowest AIC Value, of 380.02, shows the most significant predictor variables to be credit amount, account balance, and payment status of previous credit.

```
## Start:  AIC=376.25
## credit.application.result ~ account.balance + duration.of.credit.month +
##     payment.status.of.previous.credit + purpose + credit.amount +
##     value.savings.stocks + length.of.current.employment + installment.per.cent +
##     most.valuable.available.asset + age.years + type.of.apartment +
##     no.of.credits.at.this.bank
##
##                                    Df Deviance    AIC
## - most.valuable.available.asset     3   331.64 371.64
## - type.of.apartment                 2   330.96 372.96
## - no.of.credits.at.this.bank        1   330.27 374.27
## - duration.of.credit.month          1   330.43 374.43
```

```
## - purpose                             3   334.66 374.66
## - installment.per.cent                3   335.66 375.66
## - value.savings.stocks                2   333.86 375.86
## <none>                                    330.25 376.25
## - age.years                           1   332.75 376.75
## - length.of.current.employment        2   335.38 377.38
## - credit.amount                       1   337.70 381.70
## - payment.status.of.previous.credit   2   344.26 386.26
## - account.balance                     1   346.05 390.05
##
## Step:  AIC=371.64
## credit.application.result ~ account.balance + duration.of.credit.month +
##     payment.status.of.previous.credit + purpose + credit.amount +
##     value.savings.stocks + length.of.current.employment + installment.per.cent +
##     age.years + type.of.apartment + no.of.credits.at.this.bank
##
##                                     Df Deviance    AIC
## - type.of.apartment                  2   333.30 369.30
## - no.of.credits.at.this.bank         1   331.68 369.68
## - duration.of.credit.month           1   331.84 369.84
## - purpose                            3   336.70 370.70
## - value.savings.stocks               2   335.11 371.11
## - installment.per.cent               3   337.62 371.62
## <none>                                   331.64 371.64
## - age.years                          1   334.31 372.31
## - length.of.current.employment       2   336.61 372.61
## + most.valuable.available.asset      3   330.25 376.25
## - credit.amount                      1   341.73 379.73
## - payment.status.of.previous.credit  2   346.86 382.86
## - account.balance                    1   347.99 385.99
##
## Step:  AIC=369.3
## credit.application.result ~ account.balance + duration.of.credit.month +
##     payment.status.of.previous.credit + purpose + credit.amount +
##     value.savings.stocks + length.of.current.employment + installment.per.cent +
##     age.years + no.of.credits.at.this.bank
##
##                                     Df Deviance    AIC
## - no.of.credits.at.this.bank         1   333.34 367.34
## - duration.of.credit.month           1   333.65 367.65
## - purpose                            3   338.84 368.84
## - age.years                          1   335.01 369.01
## - value.savings.stocks               2   337.21 369.21
## <none>                                   333.30 369.30
## - installment.per.cent               3   339.82 369.82
## - length.of.current.employment       2   338.24 370.24
## + type.of.apartment                  2   331.64 371.64
## + most.valuable.available.asset      3   330.96 372.96
## - credit.amount                      1   344.55 378.55
## - payment.status.of.previous.credit  2   348.42 380.42
## - account.balance                    1   349.14 383.14
##
## Step:  AIC=367.34
## credit.application.result ~ account.balance + duration.of.credit.month +
```

```
##     payment.status.of.previous.credit + purpose + credit.amount +
##     value.savings.stocks + length.of.current.employment + installment.per.cent +
##     age.years
##
##                                    Df Deviance    AIC
## - duration.of.credit.month          1   333.67 365.67
## - purpose                           3   338.85 366.85
## - age.years                         1   335.05 367.05
## - value.savings.stocks              2   337.30 367.30
## <none>                                  333.34 367.34
## - installment.per.cent              3   339.86 367.86
## - length.of.current.employment      2   338.26 368.26
## + no.of.credits.at.this.bank        1   333.30 369.30
## + type.of.apartment                 2   331.68 369.68
## + most.valuable.available.asset     3   330.99 370.99
## - credit.amount                     1   344.55 376.55
## - payment.status.of.previous.credit 2   348.42 378.42
## - account.balance                   1   349.17 381.17
##
## Step:  AIC=365.67
## credit.application.result ~ account.balance + payment.status.of.previous.credit +
##     purpose + credit.amount + value.savings.stocks + length.of.current.employment +
##     installment.per.cent + age.years
##
##                                    Df Deviance    AIC
## - purpose                           3   339.08 365.08
## - age.years                         1   335.54 365.54
## - value.savings.stocks              2   337.64 365.64
## <none>                                  333.67 365.67
## - length.of.current.employment      2   338.47 366.47
## + duration.of.credit.month          1   333.34 367.34
## - installment.per.cent              3   341.48 367.48
## + no.of.credits.at.this.bank        1   333.65 367.65
## + type.of.apartment                 2   331.86 367.86
## + most.valuable.available.asset     3   331.16 369.16
## - payment.status.of.previous.credit 2   349.09 377.09
## - account.balance                   1   349.59 379.59
## - credit.amount                     1   354.88 384.88
##
## Step:  AIC=365.08
## credit.application.result ~ account.balance + payment.status.of.previous.credit +
##     credit.amount + value.savings.stocks + length.of.current.employment +
##     installment.per.cent + age.years
##
##                                    Df Deviance    AIC
## - age.years                         1   340.94 364.94
## <none>                                  339.08 365.08
## - value.savings.stocks              2   343.34 365.34
## + purpose                           3   333.67 365.67
## - length.of.current.employment      2   343.71 365.71
## + type.of.apartment                 2   336.81 366.81
## + duration.of.credit.month          1   338.85 366.85
## + no.of.credits.at.this.bank        1   339.08 367.08
## + most.valuable.available.asset     3   336.19 368.19
```

```
## - installment.per.cent               3   348.72 368.72
## - payment.status.of.previous.credit  2   353.91 375.91
## - account.balance                    1   356.28 380.28
## - credit.amount                      1   357.32 381.32
##
## Step:  AIC=364.94
## credit.application.result ~ account.balance + payment.status.of.previous.credit +
##     credit.amount + value.savings.stocks + length.of.current.employment +
##     installment.per.cent
##
##                                    Df Deviance    AIC
## <none>                                  340.94 364.94
## + age.years                         1   339.08 365.08
## - value.savings.stocks              2   345.13 365.13
## + purpose                           3   335.54 365.54
## + duration.of.credit.month          1   340.57 366.57
## - length.of.current.employment      2   346.61 366.61
## + no.of.credits.at.this.bank        1   340.94 366.94
## + type.of.apartment                 2   339.82 367.82
## - installment.per.cent              3   350.62 368.62
## + most.valuable.available.asset     3   338.76 368.76
## - payment.status.of.previous.credit  2   356.34 376.34
## - account.balance                   1   358.01 380.01
## - credit.amount                     1   358.02 380.02
```

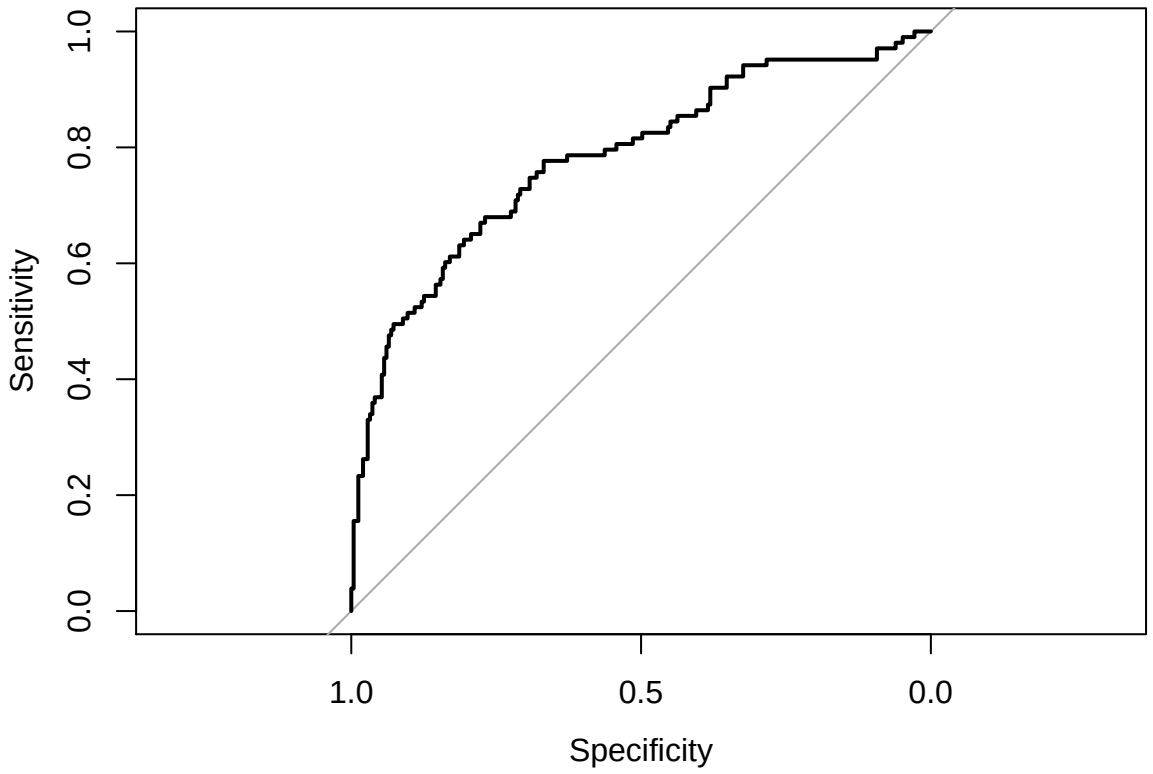
ROC Curve

Per the ROC Curve below, the area under the curve is 0.782.

```
## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```




```
## Area under the curve: 0.782
```

Confusion Matrix

Per the confusion matrix below, the logit regression model produced an overall accuracy of 77%, with a 79% chance of predicting non-creditworthy applicants and a 63% chance of predicting creditworthy applicants.

```
## $positive
## [1] "0"
##
## $table
##           Reference
## Prediction   0   1
##          0 104   7
##          1  27  12
##
## $overall
##       Accuracy          Kappa  AccuracyLower  AccuracyUpper   AccuracyNull
##    0.773333333    0.293433084    0.697922567    0.837630728    0.873333333
## AccuracyPValue  McnemarPValue
##    0.999767869    0.001120135
##
## $byClass
##          Sensitivity          Specificity       Pos Pred Value
##            0.7938931            0.6315789            0.9369369
##       Neg Pred Value            Precision               Recall
##            0.3076923            0.9369369            0.7938931
##                   F1           Prevalence       Detection Rate
##            0.8595041            0.8733333            0.6933333
```

```
## Detection Prevalence    Balanced Accuracy
##            0.7400000            0.7127360
##
## $mode
## [1] "sens_spec"
##
## $dots
## list()
##
## attr(,"class")
## [1] "confusionMatrix"
```

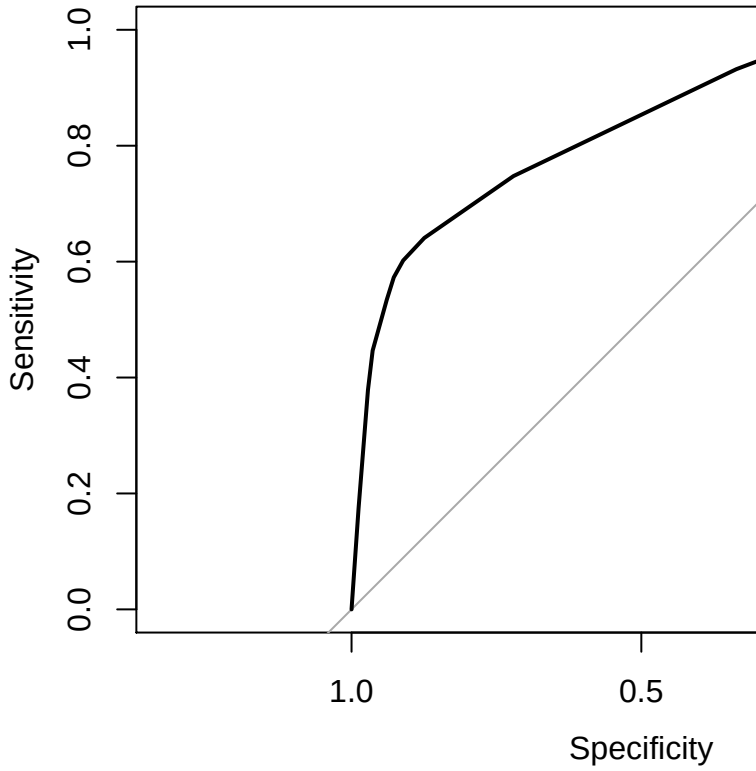
Decision Tree Model

Significant Predictor Variables

Per the Decision Tree variable importance output below, the most significant predictor variables are account balance, payment status of previous credit, and duration of credit month.

```
##                   account.balance payment.status.of.previous.credit
##                         6.8373810                         5.3370172
##          duration.of.credit.month                     credit.amount
##                         3.8895600                         3.6455751
##              installment.per.cent     most.valuable.available.asset
##                         2.9120954                         2.5000243
##                         age.years      length.of.current.employment
##                         1.6926355                         1.3656572
##              value.savings.stocks                           purpose
##                         1.2997728                         0.7367691
##        no.of.credits.at.this.bank                 type.of.apartment
##                         0.6511791                         0.2593040
```

ROC Curve



Per the ROC Curve below, the area under the curve is 0.8174.

```
## Area under the curve: 0.8174
```

Confusion Matrix

Per the confusion matrix below, the decision tree model produced an overall accuracy of 71%, with a 79% chance of predicting non-creditworthy applicants and a 44% chance of predicting creditworthy applicants.

```
## $positive
## [1] "0"
##
## $table
##           Reference
## Prediction  0  1
##          0 92 19
##          1 24 15
##
## $overall
##       Accuracy          Kappa  AccuracyLower  AccuracyUpper   AccuracyNull
##      0.7133333      0.2227043      0.6338911      0.7841387      0.7733333
## AccuracyPValue  McnemarPValue
##      0.9652772      0.5418656
##
## $byClass
##          Sensitivity          Specificity       Pos Pred Value
##            0.7931034            0.4411765            0.8288288
##       Neg Pred Value            Precision               Recall
##            0.3846154            0.8288288            0.7931034
```

```
##                   F1           Prevalence       Detection Rate
##            0.8105727            0.7733333            0.6133333
## Detection Prevalence    Balanced Accuracy
##            0.7400000            0.6171400
##
## $mode
## [1] "sens_spec"
##
## $dots
## list()
##
## attr(,"class")
## [1] "confusionMatrix"
```

Random Forest Model

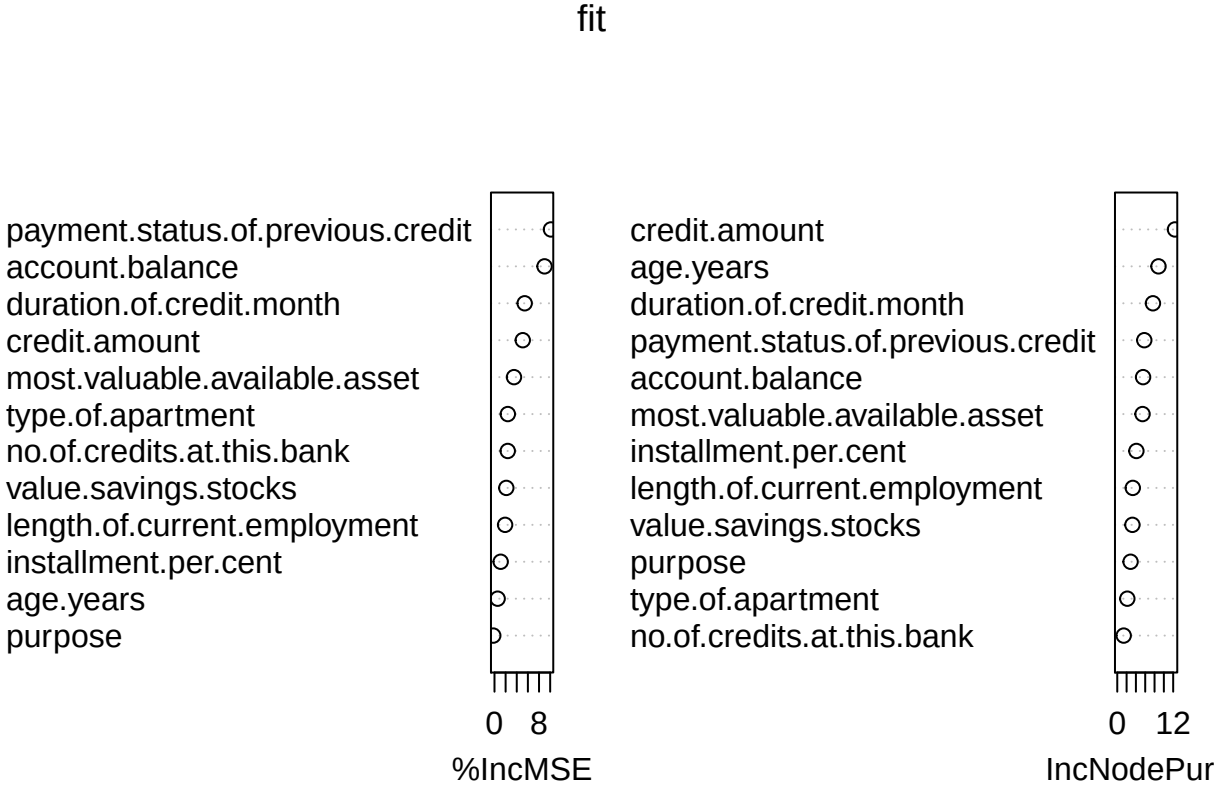
Significant Predictor Variables

Per the Random Forest Model variable importance output below, the most significant predictor variables are payment status of previous credit, account balance, and most valuable available asset.

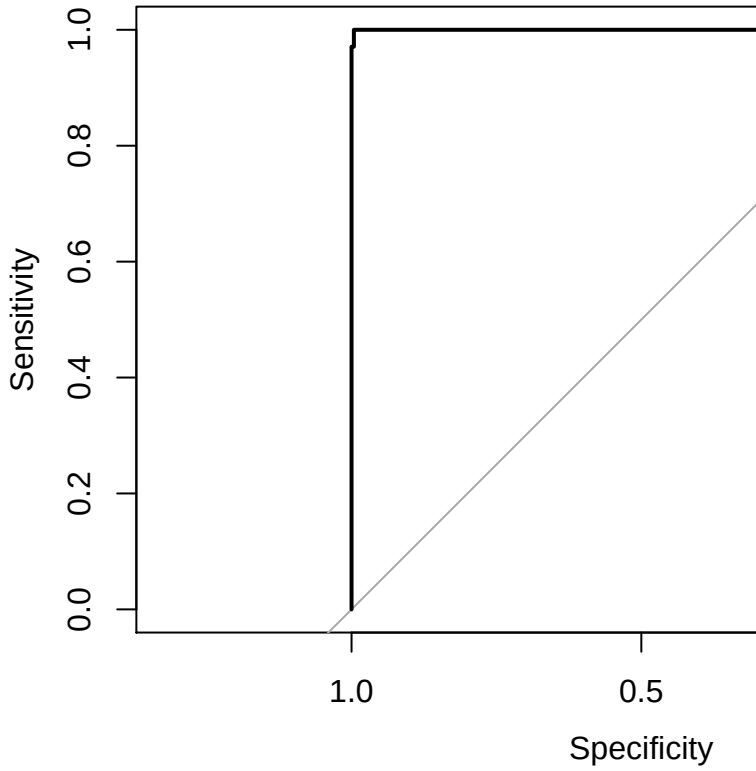
```
## randomForest 4.6-14
```

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

```
## Warning in randomForest.default(m, y, ...): The response has five or fewer
## unique values. Are you sure you want to do regression?
```



ROC Curve



Per the ROC Curve below, the area under the curve is 0.9998.

```
## Area under the curve: 0.9999
```

Confusion Matrix

Per the confusion matrix below, the random forest model produced an overall accuracy of 83%, with an 85% chance of predicting non-creditworthy applicants and a 72% chance of predicting creditworthy applicants.

```
## $positive
## [1] "0"
##
## $table
##           Reference
## Prediction   0   1
##          0 103   8
##          1  18  21
##
## $overall
##       Accuracy          Kappa  AccuracyLower  AccuracyUpper   AccuracyNull
##     0.82666667     0.50869237     0.75643325     0.88353783     0.80666667
## AccuracyPValue  McnemarPValue
##     0.30819045     0.07755617
##
## $byClass
##          Sensitivity          Specificity       Pos Pred Value
##            0.8512397            0.7241379            0.9279279
##       Neg Pred Value            Precision               Recall
##            0.5384615            0.9279279            0.8512397
```

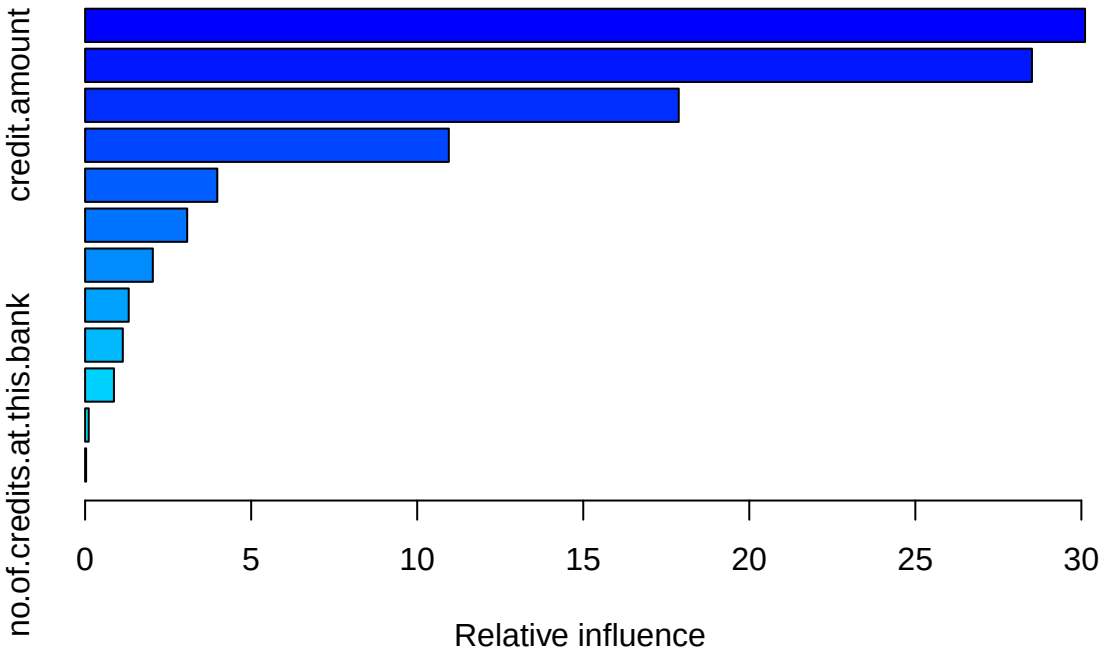
```
##                   F1           Prevalence       Detection Rate
##            0.8879310            0.8066667            0.6866667
## Detection Prevalence    Balanced Accuracy
##            0.7400000            0.7876888
##
## $mode
## [1] "sens_spec"
##
## $dots
## list()
##
## attr(,"class")
## [1] "confusionMatrix"
```

Boosted Model

Significant Predictor Variables

Per the Boosted Model variable importance output below, the most significant predictor variables are account balanace, payment status of previous credit, and credit amount.

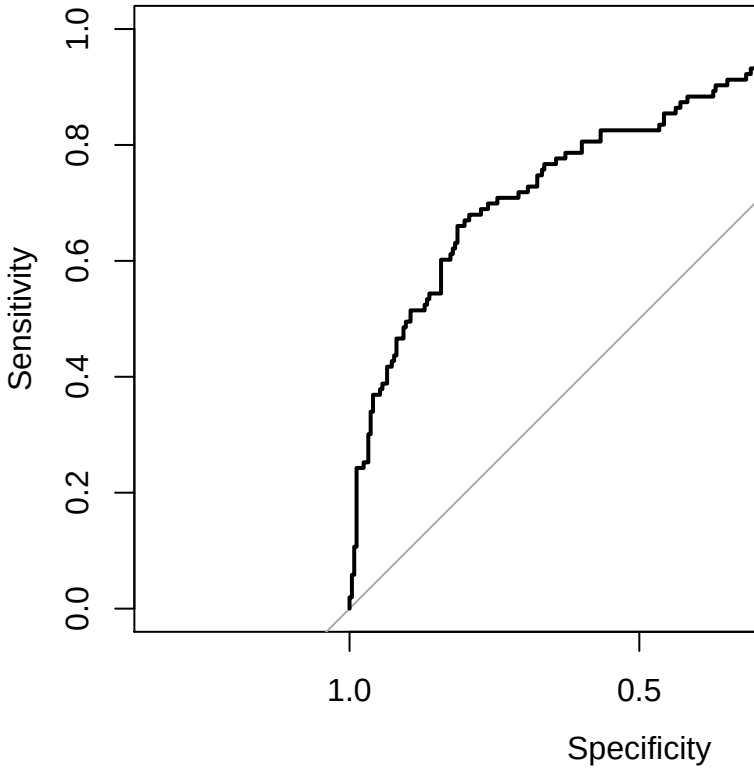
```
## Loading required package: survival
## Loading required package: lattice
## Loading required package: splines
## Loading required package: parallel
## Loaded gbm 2.1.3
## Distribution not specified, assuming bernoulli ...
```




```
##                                                  var
## account.balance                      account.balance
```

```
## payment.status.of.previous.credit payment.status.of.previous.credit
## credit.amount                                         credit.amount
## duration.of.credit.month                   duration.of.credit.month
## value.savings.stocks                           value.savings.stocks
## age.years                                                 age.years
## most.valuable.available.asset         most.valuable.available.asset
## length.of.current.employment           length.of.current.employment
## installment.per.cent                           installment.per.cent
## purpose                                                     purpose
## type.of.apartment                             type.of.apartment
## no.of.credits.at.this.bank               no.of.credits.at.this.bank
##                                    rel.inf
## account.balance                  30.11044529
## payment.status.of.previous.credit 28.51184276
## credit.amount                    17.87559744
## duration.of.credit.month         10.95258203
## value.savings.stocks              3.98071737
## age.years                         3.07359689
## most.valuable.available.asset     2.03928527
## length.of.current.employment      1.31383062
## installment.per.cent              1.13545488
## purpose                           0.86779011
## type.of.apartment                 0.10874228
## no.of.credits.at.this.bank        0.03011508
```

ROC Curve



Per the ROC Curve below, the area under the curve is 0.7831.

```
## Area under the curve: 0.7824
```

Confusion Matrix

Per the confusion matrix below, the boosted model produced an overall accuracy of 75%, with a 76% chance of predicting non-creditworthy applicants and a 55% chance of predicting creditworthy applicants.

```
## $positive
## [1] "0"
##
## $table
##           Reference
## Prediction   0   1
##          0 106   5
##          1  33   6
##
## $overall
##       Accuracy          Kappa  AccuracyLower  AccuracyUpper   AccuracyNull
##   7.466667e-01   1.418248e-01   6.692562e-01   8.140696e-01   9.266667e-01
## AccuracyPValue  McnemarPValue
##   1.000000e+00   1.186911e-05
##
## $byClass
##          Sensitivity          Specificity       Pos Pred Value
##            0.7625899            0.5454545            0.9549550
##       Neg Pred Value            Precision               Recall
##            0.1538462            0.9549550            0.7625899
##                   F1           Prevalence       Detection Rate
##            0.8480000            0.9266667            0.7066667
## Detection Prevalence    Balanced Accuracy
##            0.7400000            0.6540222
##
## $mode
## [1] "sens_spec"
##
## $dots
## list()
##
## attr(,"class")
## [1] "confusionMatrix"
```

Conclusions

The relative importance of accurately predicting creditworthy and non-creditworthy applicants should be inquired from the bank to determine whether the model should be weighted toward creditworthy or non-creditworthy applicants. Assuming these predictions are of equal importance, the Random Forest Model was used because it is the most accurate model with an accuracy of 83%. The ROC Curve further cements this choice, because the area under the Random Forest Model Curve is the largest at 0.9998. Though the Random Forest Model is the most accurate, the model is biased toward non-creditworthy applicants with an accuracy of 85% (412 applicants not approved) and away from creditworthy applicants with an accuracy of 72% (88 applicants approved).