Homework #3: Crime Prediction

Data 621 Business Analytics and Data Mining

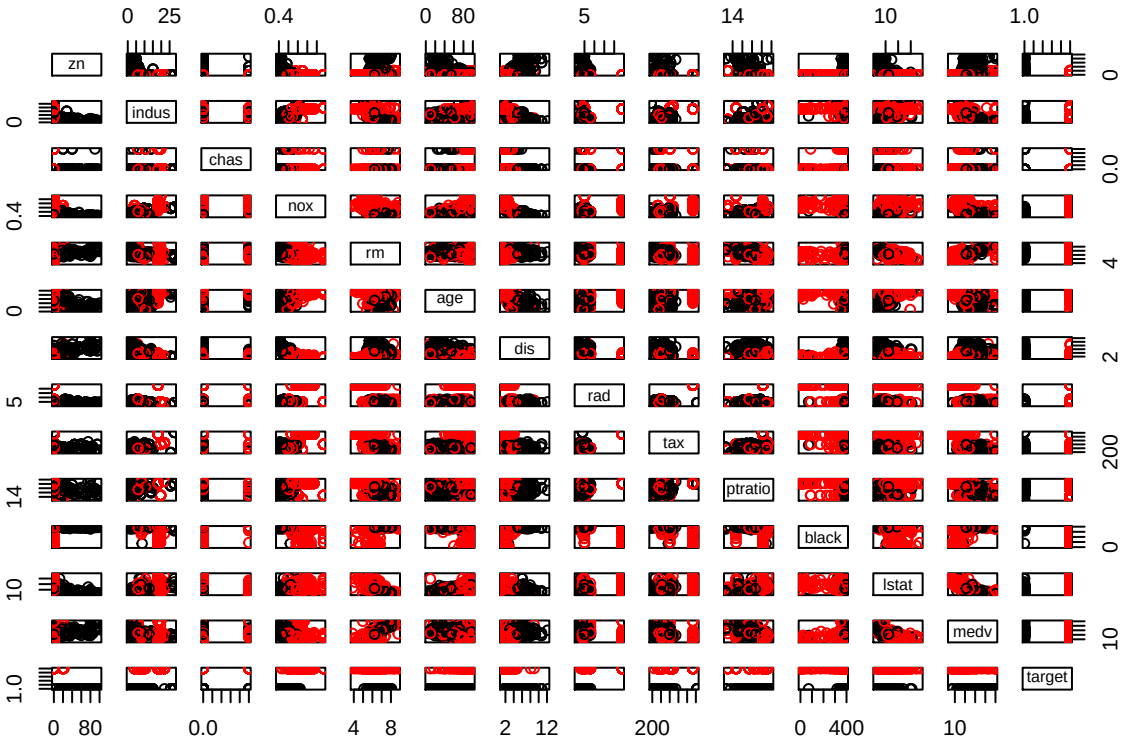
Aadi Kalloo, Nathan Lim, Asher Meyers, Daniel Smilowitz, Logan Thomson

Due July 3, 2016

Contents

Data Exploration

```
pairs(train_df, col= train_df$target)
```

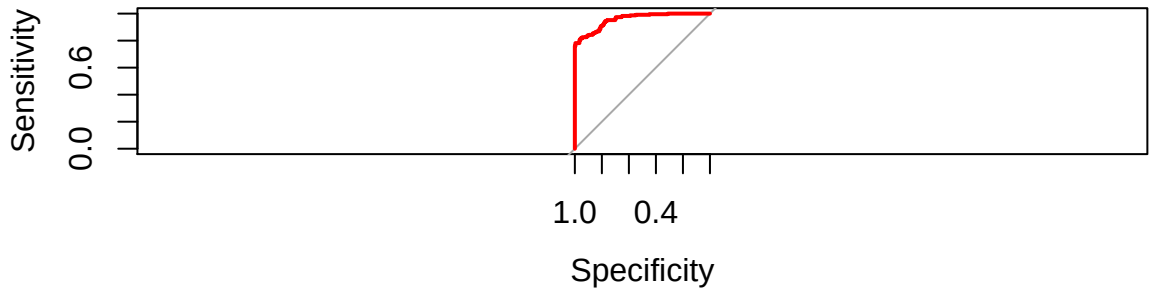


Data Preparation

Model Creation

Model 1

```
Call:
glm(formula = target ~ nox + age + rad + medv, family = binomial,
    data = train_df)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.76145  -0.33936  -0.06729   0.01665   2.69085

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -17.626271   2.167700  -8.131 4.25e-16 ***
nox          23.623709   3.935564   6.003 1.94e-09 ***
age           0.018240   0.009172   1.989   0.0467 *
rad           0.452771   0.109259   4.144 3.41e-05 ***
medv          0.044807   0.023194   1.932   0.0534 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 645.88  on 465  degrees of freedom
Residual deviance: 232.80  on 461  degrees of freedom
AIC: 242.8

Number of Fisher Scoring iterations: 8
```

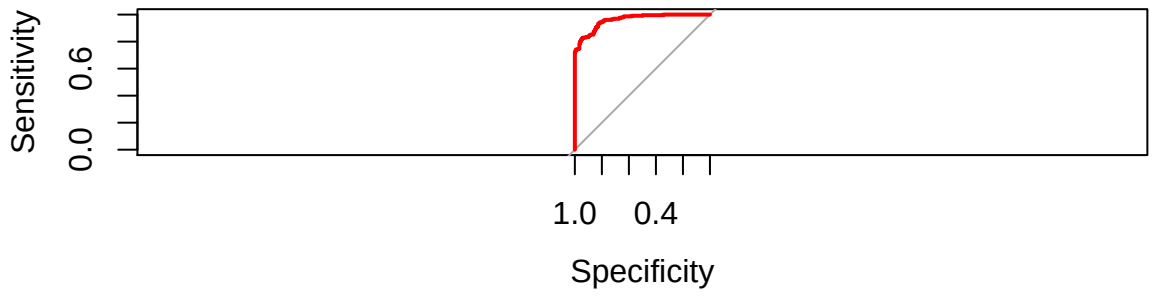



```
Call:
roc.formula(formula = target ~ predicted_model1, data = train_df)

Data: predicted_model1 in 237 controls (target 0) < 229 cases (target 1).
Area under the curve: 0.957

Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 214  23
         1  37 192

               Accuracy : 0.8712
```

```
               95% CI : (0.8374, 0.9003)
    No Information Rate : 0.5386
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.7421
 Mcnemar's Test P-Value : 0.09329

            Sensitivity : 0.8930
            Specificity : 0.8526
         Pos Pred Value : 0.8384
         Neg Pred Value : 0.9030
             Prevalence : 0.4614
         Detection Rate : 0.4120
   Detection Prevalence : 0.4914
      Balanced Accuracy : 0.8728

       'Positive' Class : 1
```

Model 2

```
Call:
glm(formula = target ~ nox + age + rad + ptratio + medv, family = binomial,
    data = train_df)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.96654  -0.29783  -0.03987   0.00769   2.80829

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -24.936540   3.683449  -6.770 1.29e-11 ***
nox          25.334778   4.084106   6.203 5.53e-10 ***
age           0.019403   0.009308   2.085  0.03711 *
rad           0.512600   0.114818   4.464 8.03e-06 ***
ptratio       0.274193   0.098737   2.777  0.00549 **
medv          0.085445   0.027979   3.054  0.00226 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 645.88  on 465  degrees of freedom
Residual deviance: 224.71  on 460  degrees of freedom
AIC: 236.71

Number of Fisher Scoring iterations: 8
```

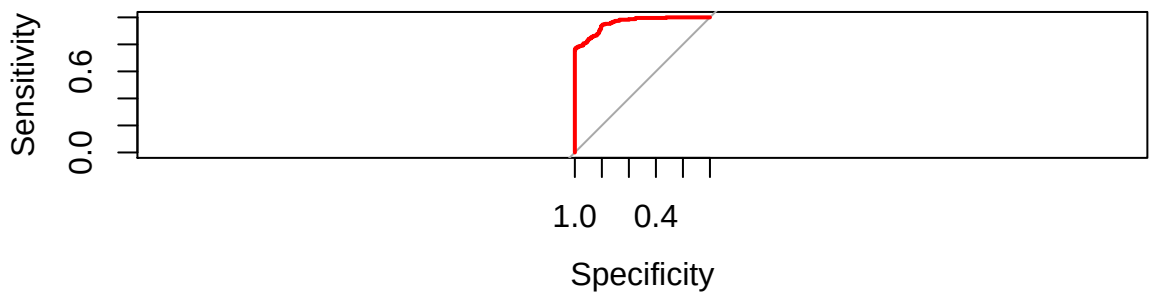


```
Call:
roc.formula(formula = factor(target) ~ predicted_model2, data = train_df)

Data: predicted_model2 in 237 controls (factor(target) 0) < 229 cases (factor(target) 1).
Area under the curve: 0.9605


Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 213  24
         1  37 192

               Accuracy : 0.8691
                 95% CI : (0.835, 0.8984)
    No Information Rate : 0.5365
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.7379
 Mcnemar's Test P-Value : 0.1244

            Sensitivity : 0.8889
            Specificity : 0.8520
         Pos Pred Value : 0.8384
         Neg Pred Value : 0.8987
             Prevalence : 0.4635
         Detection Rate : 0.4120
   Detection Prevalence : 0.4914
      Balanced Accuracy : 0.8704

       'Positive' Class : 1
```

Model 3

```
Call:
glm(formula = target ~ log(nox) + age + log(rad) + medv, family = binomial,
    data = train_df)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.75155  -0.31275  -0.02338   0.11521   2.75907

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.960177   1.925327   1.018   0.3086
log(nox)    13.000955   2.097257   6.199 5.68e-10 ***
age          0.016963   0.009147   1.855   0.0637 .
log(rad)     2.269365   0.449417   5.050 4.43e-07 ***
medv         0.048903   0.023537   2.078   0.0377 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 645.88  on 465  degrees of freedom
Residual deviance: 231.73  on 461  degrees of freedom
```

```
AIC: 241.73

Number of Fisher Scoring iterations: 7
```

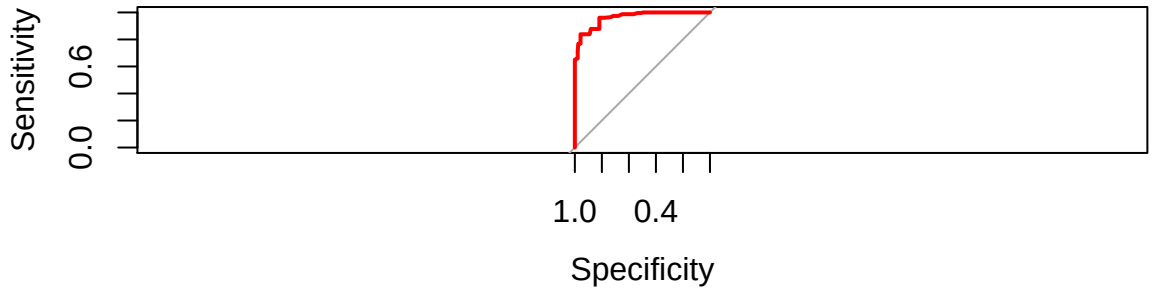



```
Call:
roc.formula(formula = factor(target) ~ predicted_model3, data = train_df)

Data: predicted_model3 in 237 controls (factor(target) 0) < 229 cases (factor(target) 1).
Area under the curve: 0.9584

Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 213  24
         1  37 192

               Accuracy : 0.8691
                 95% CI : (0.835, 0.8984)
    No Information Rate : 0.5365
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.7379
 Mcnemar's Test P-Value : 0.1244

            Sensitivity : 0.8889
            Specificity : 0.8520
         Pos Pred Value : 0.8384
         Neg Pred Value : 0.8987
             Prevalence : 0.4635
         Detection Rate : 0.4120
   Detection Prevalence : 0.4914
      Balanced Accuracy : 0.8704

       'Positive' Class : 1
```

Model4

```
Call:
glm(formula = target ~ log(nox) + log(rad) + tax, family = binomial,
    data = train_df)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.98241  -0.23038  -0.00753   0.14138   2.69904
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  9.385116   1.828731   5.132 2.87e-07 ***
log(nox)    19.347684   2.518525   7.682 1.56e-14 ***
log(rad)     3.356382   0.544715   6.162 7.20e-10 ***
tax         -0.008214   0.002335  -3.518 0.000435 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 645.88  on 465  degrees of freedom
Residual deviance: 223.54  on 462  degrees of freedom
AIC: 231.54

Number of Fisher Scoring iterations: 7
```




```
Call:
roc.formula(formula = factor(target) ~ predicted_model4, data = train_df)

Data: predicted_model4 in 237 controls (factor(target) 0) < 229 cases (factor(target) 1).
Area under the curve: 0.961

Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 211  26
         1  37 192

               Accuracy : 0.8648
                 95% CI : (0.8304, 0.8945)
    No Information Rate : 0.5322
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.7293
 Mcnemar's Test P-Value : 0.2077

            Sensitivity : 0.8807
            Specificity : 0.8508
         Pos Pred Value : 0.8384
         Neg Pred Value : 0.8903
             Prevalence : 0.4678
         Detection Rate : 0.4120
   Detection Prevalence : 0.4914
      Balanced Accuracy : 0.8658
```

```
'Positive' Class : 1
```

Model Selection and Prediction

The model selected used only the significant predictors (Model 3) was selected as the best model for prediction of `TARGET` in the crime data set. While the AUC value of this model the second-highest of the four models tested, its mean cross-validation error indicates that it is has the best predictive value for unseen data. Additionally, it is a parsimonious model, and the simplicity lends itself to easier understanding of the model by other users.

10-fold Cross Validation

Mean CV Error

Model1	36.6
Model2	46.69
Model3	15.48
Model4	22.14

The linear model is applied to an evaluation dataset containing response variables for 259 cases. A table of the predicted team wins is presented below.

0	1
250	216

0	1
237	229

Similar to the training dataset, the predictions for the test data set predictions are weighted more toward crime being below the median

A comparison of the full sets of predictions for the evaluation dataset is available in Appendix B.

Appendix A

Appendix B – Index-wise Results from Predictive Model

Appendix C – R Code