DATA621-Homework3-HoddeFarrisBurmood

Rob Hodde, Matt Farris, JeffreyBurmood

3/28/2017

DATA621 Homework #3

Team Members: Rob Hodde, Matt Farris, Jeffrey Burmood

Problem Description

Explore, analyze and model a data set containing information on crime for various neighborhoods of a major city. Using the data set build a binary logistic regression model on the training data set to predict whether the neighborhood will be at risk for high crime levels. Provide classifications and probabilities for the evaluation data set using the developed binary logistic regression model.

Data Exploration

```
# Load required libraries
library(ggplot2)
library(ROCR)
```

```
## Loading required package: gplots
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

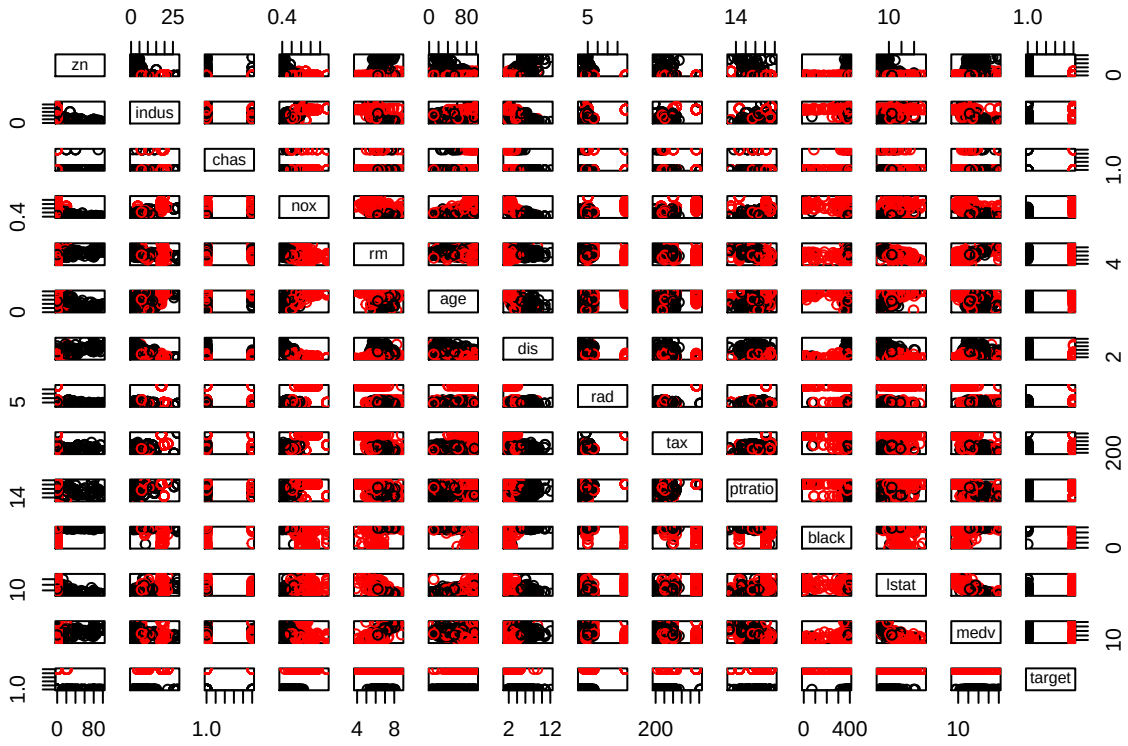
```
library(RCurl)
```

```
## Loading required package: bitops
```

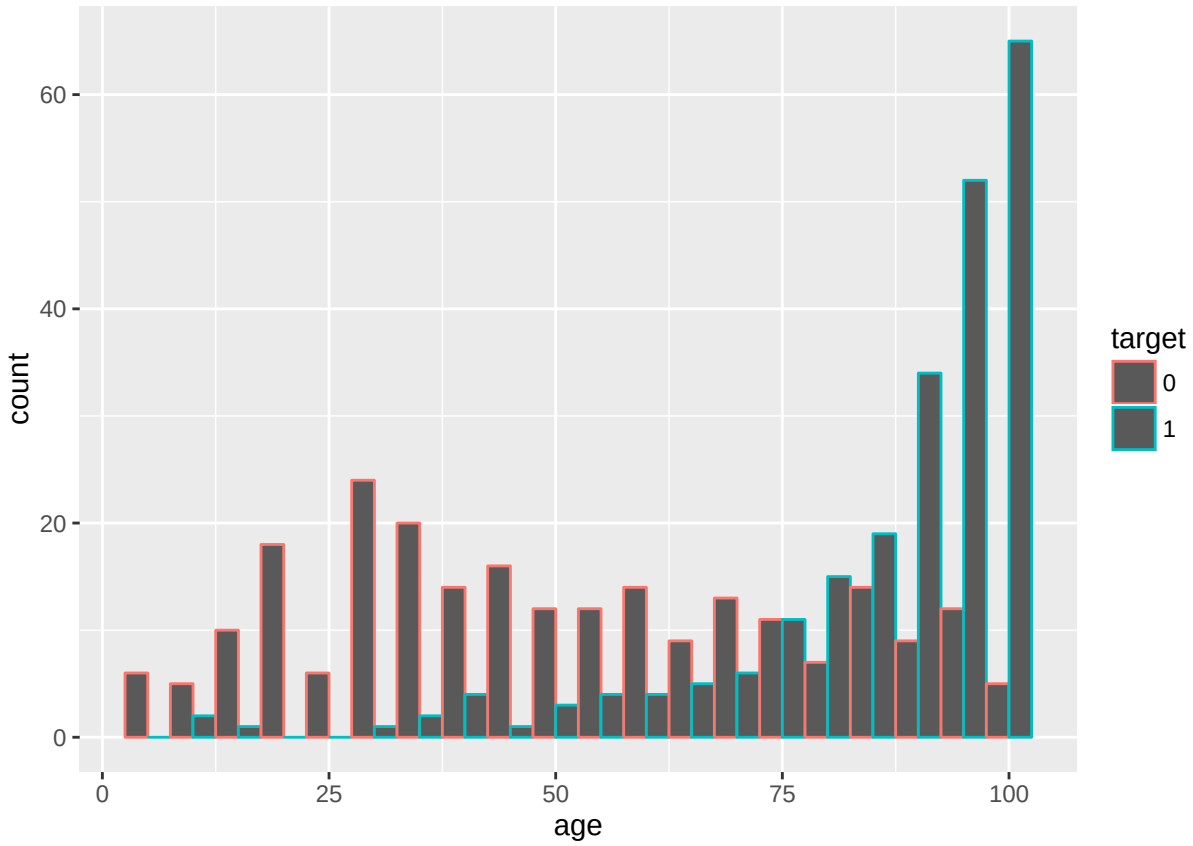
```
# Read in the dataset from github
crime <- read.csv(text=getURL("https://raw.githubusercontent.com/jeffreyburmood/data621/master/Homework3
crime_eval <- read.csv(text=getURL("https://raw.githubusercontent.com/jeffreyburmood/data621/master/Hom
# Need to set variables to a factor as required
crime$target <- as.factor(crime$target)
crime$chas <- as.factor(crime$chas)
crime_eval$chas <- as.factor(crime_eval$chas)
# Now generate some summary statistics
print(summary(crime))
```

```
##        zn              indus         chas          nox
##  Min.   :  0.00   Min.   : 0.460   0:433   Min.   :0.3890
##  1st Qu.:  0.00   1st Qu.: 5.145   1: 33   1st Qu.:0.4480
##  Median :  0.00   Median : 9.690           Median :0.5380
##  Mean   : 11.58   Mean   :11.105           Mean   :0.5543
##  3rd Qu.: 16.25   3rd Qu.:18.100           3rd Qu.:0.6240
##  Max.   :100.00   Max.   :27.740           Max.   :0.8710
##        rm              age              dis              rad
##  Min.   :3.863   Min.   :  2.90   Min.   : 1.130   Min.   : 1.00
##  1st Qu.:5.887   1st Qu.: 43.88   1st Qu.: 2.101   1st Qu.: 4.00
##  Median :6.210   Median : 77.15   Median : 3.191   Median : 5.00
##  Mean   :6.291   Mean   : 68.37   Mean   : 3.796   Mean   : 9.53
##  3rd Qu.:6.630   3rd Qu.: 94.10   3rd Qu.: 5.215   3rd Qu.:24.00
##  Max.   :8.780   Max.   :100.00   Max.   :12.127   Max.   :24.00
##       tax           ptratio          black            lstat
##  Min.   :187.0   Min.   :12.6   Min.   :  0.32   Min.   : 1.730
##  1st Qu.:281.0   1st Qu.:16.9   1st Qu.:375.61   1st Qu.: 7.043
##  Median :334.5   Median :18.9   Median :391.34   Median :11.350
##  Mean   :409.5   Mean   :18.4   Mean   :357.12   Mean   :12.631
##  3rd Qu.:666.0   3rd Qu.:20.2   3rd Qu.:396.24   3rd Qu.:16.930
##  Max.   :711.0   Max.   :22.0   Max.   :396.90   Max.   :37.970
##       medv       target
##  Min.   : 5.00   0:237
##  1st Qu.:17.02   1:229
##  Median :21.20
##  Mean   :22.59
##  3rd Qu.:25.00
##  Max.   :50.00
```

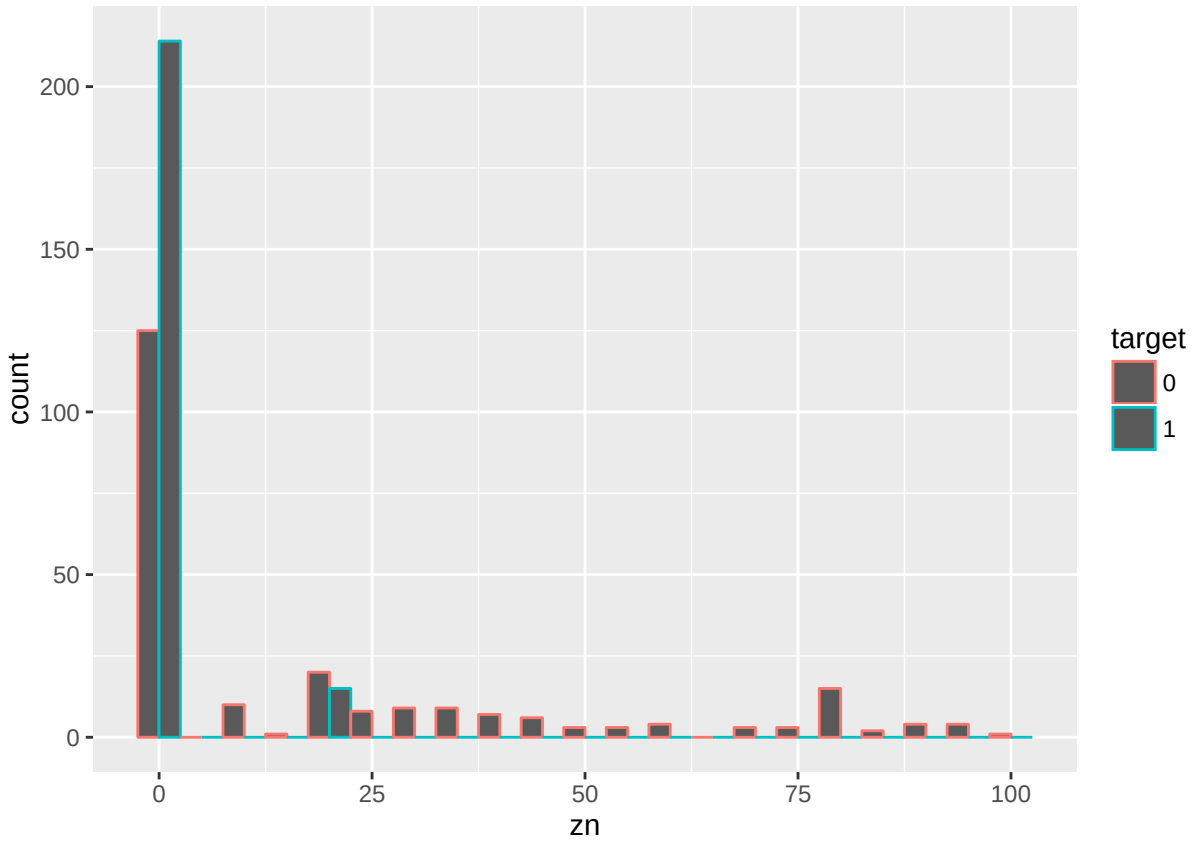
```r
# Visual check for obvious correlations
pairs(crime,col=crime$target)
```




```
#
# no NAs found so no missing values to remove or fix?
#
# Look over the variables checking for outliers/influencial points, correlation between variables, etc.

# Look at some histograms
# age
age.plot <- ggplot(crime, aes(x=age,color=target)) + geom_histogram(position="dodge",binwidth=5)
print(age.plot)
```




```
# zn
zn.plot <- ggplot(crime, aes(x=zn,color=target)) + geom_histogram(position="dodge",binwidth=5)
print(zn.plot)
```




```
#
```

Data Preparation

```
# Based on the data exploration results, identify any changes, transformations, and new or deleted vari
```

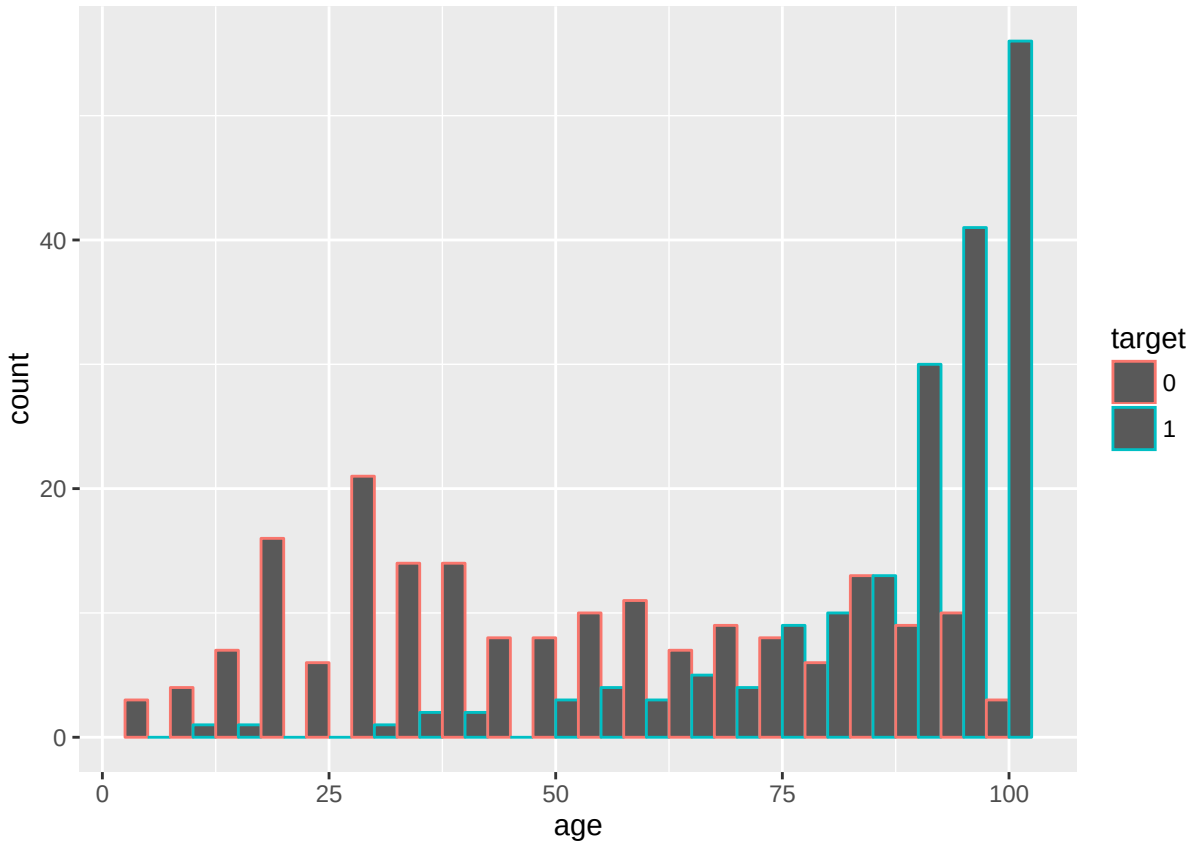
Build Models

```
## 75% of the sample size
smp_size <- floor(0.80 * nrow(crime))

## set the seed to make your partition reproductible
train_ind <- sample(seq_len(nrow(crime)), size = smp_size)

train <- crime[train_ind, ]
test <- crime[-train_ind, ]

# quick look at model with all variables
qm <- glm(target ~ .,family=binomial(link='logit'),data=train)
print(summary(qm))
```

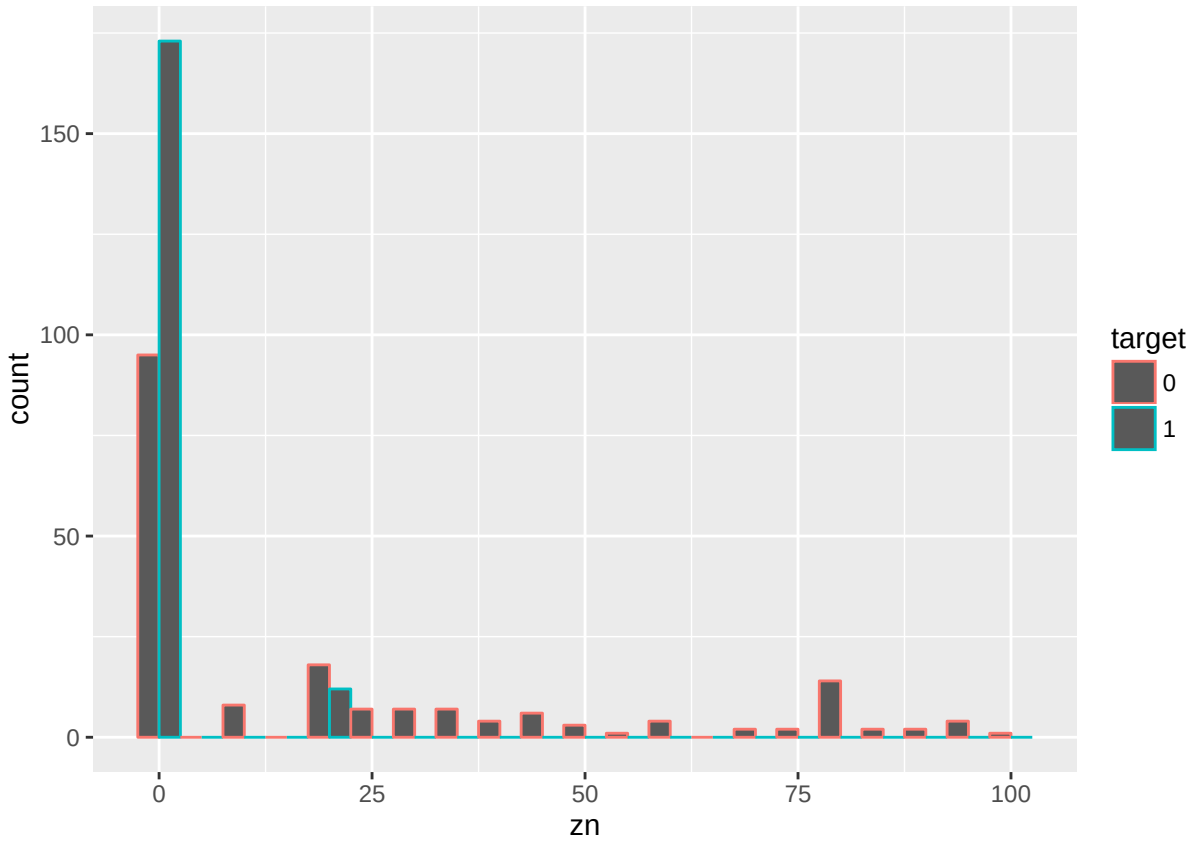
```
##
```

```
## Call:
## glm(formula = target ~ ., family = binomial(link = "logit"),
##     data = train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.93129  -0.06933  -0.00018   0.00037   3.04856
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -47.331489   9.366185  -5.053 4.34e-07 ***
## zn           -0.090253   0.049299  -1.831 0.067139 .
## indus        -0.105597   0.059737  -1.768 0.077110 .
## chas1         0.591425   0.925456   0.639 0.522782
## nox          62.653703  11.166666   5.611 2.01e-08 ***
## rm           -1.096003   0.917488  -1.195 0.232255
## age           0.051179   0.018754   2.729 0.006354 **
## dis           0.950282   0.319388   2.975 0.002927 **
## rad           0.808860   0.221810   3.647 0.000266 ***
## tax          -0.006474   0.003631  -1.783 0.074598 .
## ptratio       0.518743   0.167168   3.103 0.001915 **
## black        -0.009236   0.005927  -1.558 0.119133
## lstat         0.048067   0.065650   0.732 0.464063
## medv          0.249263   0.090158   2.765 0.005697 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 515.69  on 371  degrees of freedom
## Residual deviance: 129.94  on 358  degrees of freedom
## AIC: 157.94
##
## Number of Fisher Scoring iterations: 9
```

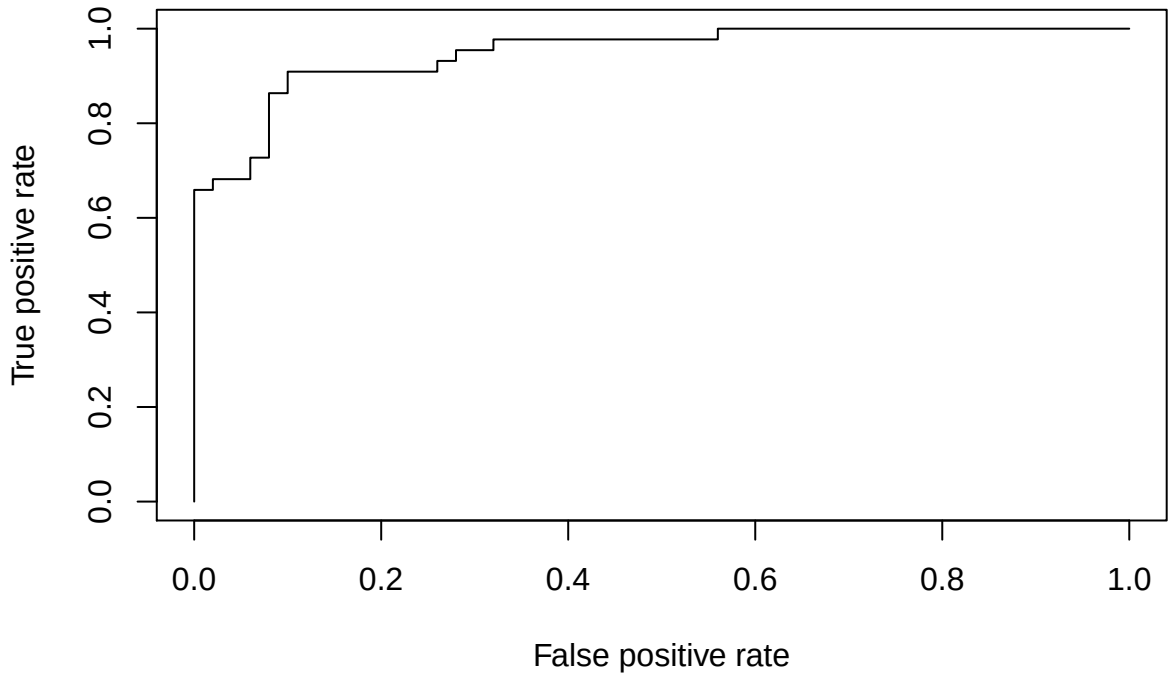
```r
# Look at some histograms
# age
age.plot <- ggplot(train, aes(x=age,color=target)) + geom_histogram(position="dodge",binwidth=5)
print(age.plot)
```




```
# zn
zn.plot <- ggplot(train, aes(x=zn,color=target)) + geom_histogram(position="dodge",binwidth=5)
print(zn.plot)
```




```
#

p <- predict(qm, newdata=subset(test,select=c(1,2,3,4,5,6,7,8,9,10,11,12,13)), type="response")
pr <- prediction(p, test$target)
prf <- performance(pr, measure = "tpr", x.measure = "fpr")
plot(prf)
```




```
#
auc <- performance(pr, measure = "auc")
auc <- auc@y.values[[1]]
auc
```

```
## [1] 0.9490909
```

Select Models

All Done!