DATA 621 Project 3

Michael Muller

April 06, 2018

Overview

Explore, analyze, and model the crime dataset

Four parts [Data Exploration, Data Preparation, Model Building, Model Selection]

1. DATA EXPLORATION

Briefing

There are 466/466 complete cases in the training dataset.

The dataset contains 466 observations with 12 features we will use for predicting the 13th feature, our target variable.

Our target variable indicates whether or not the crime rate in this particular observation is above the 'median crime rate' of the whole.

The testing dataset contains 40 observations, and our training dataset contains equal number positive/negative targets. So tracking our ROC curve will be of critical importance.

Feature notes

[zn,rad,tax,target] are all discrete variables

[indus,nox,rm,age,dis,ptratio,lstat,medv] are all continuous variables

4 Discrete, 7 Continuous, 1 Categorical predictor features (chas = dummy variable) - (Binning will be considered for the continous features with low correlation to target)

Various feature statistics

The following two charts show similar distributional traits among most features with Highway accessibility, Residential Zoning, and Property Tax lying outside the norm with disproportionate mean/medians.

High skew for residential zoning, median value of homes, distance to employment centers and whether or not they're near the river. The nature of these variables is that groups of similar values are clustered together geographically. (The house farthest away from employment centers is most probably the closest house to the house 2nd farthest away from employment centers.) Something to consider, can't draw a conclusion yet.

In cases where the median is not 50% of the range, we can expect to see some proportionate skew. Lets confirm that in figure 2, I predict skew in almost every variable. Which may mean significant transformations need to be made.

. . . .

. . . ..

. . . . . .

Figure 1 : Statistics

	Feature	Description
1	zn	Residential land zoning
2	indus	Industry land zoning
3	chas	Borders River
4	nox	Nitrogen Oxide levels
5	rm	Avg rooms per dwelling
6	age	Amt of old buildings
7	dis	Distance to employment centers
8	rad	Highway accessibility
9	tax	Property tax
10	ptratio	Pupil-teacher ratio
11	lstat	Lower status of population
12	medv	Median value of residential homes (In 1000 increments)
13	target	Target value

vars	mean	sd	median	range	skew	kurtosis	se
1	11.5772532	23.3646511	0.00000	100.0000	2.1768152	3.8135765	1.0823466
2	11.1050215	6.8458549	9.69000	27.2800	0.2885450	-1.2432132	0.3171281
3	0.0708155	0.2567920	0.00000	1.0000	3.3354899	9.1451313	0.0118957
4	0.5543105	0.1166667	0.53800	0.4820	0.7463281	-0.0357736	0.0054045
5	6.2906738	0.7048513	6.21000	4.9170	0.4793202	1.5424378	0.0326516
6	68.3675966	28.3213784	77.15000	97.1000	-0.5777075	-1.0098814	1.3119625
7	3.7956929	2.1069496	3.19095	10.9969	0.9988926	0.4719679	0.0976026
8	9.5300429	8.6859272	5.00000	23.0000	1.0102788	-0.8619110	0.4023678
9	409.5021459	167.9000887	334.50000	524.0000	0.6593136	-1.1480456	7.7778214
10	18.3984979	2.1968447	18.90000	9.4000	-0.7542681	-0.4003627	0.1017669
11	12.6314592	7.1018907	11.35000	36.2400	0.9055864	0.5033688	0.3289887
12	22.5892704	9.2396814	21.20000	45.0000	1.0766920	1.3737825	0.4280200
13	0.4914163	0.5004636	0.00000	1.0000	0.0342293	-2.0031131	0.0231835

Figure 2 : Histograms

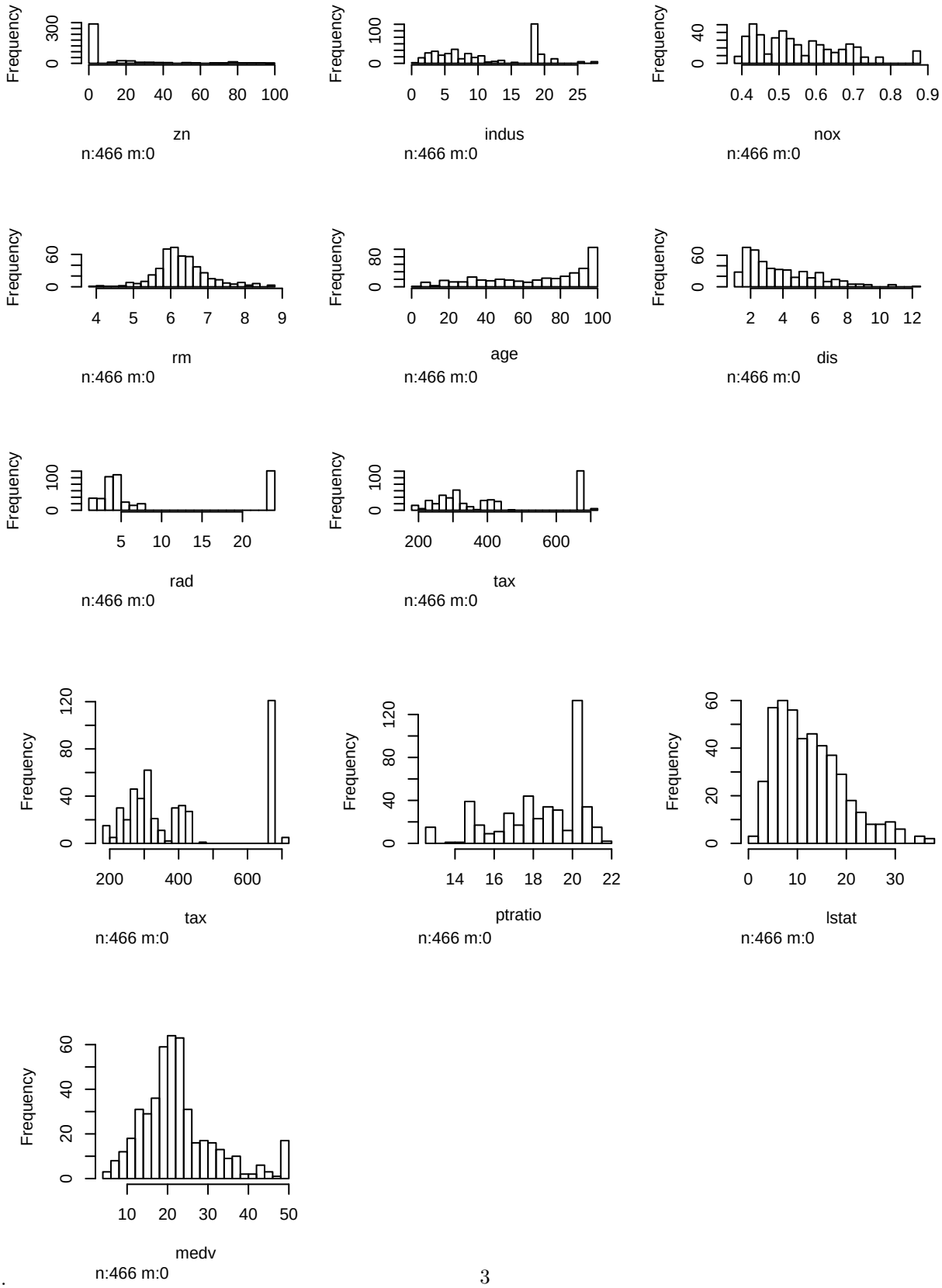
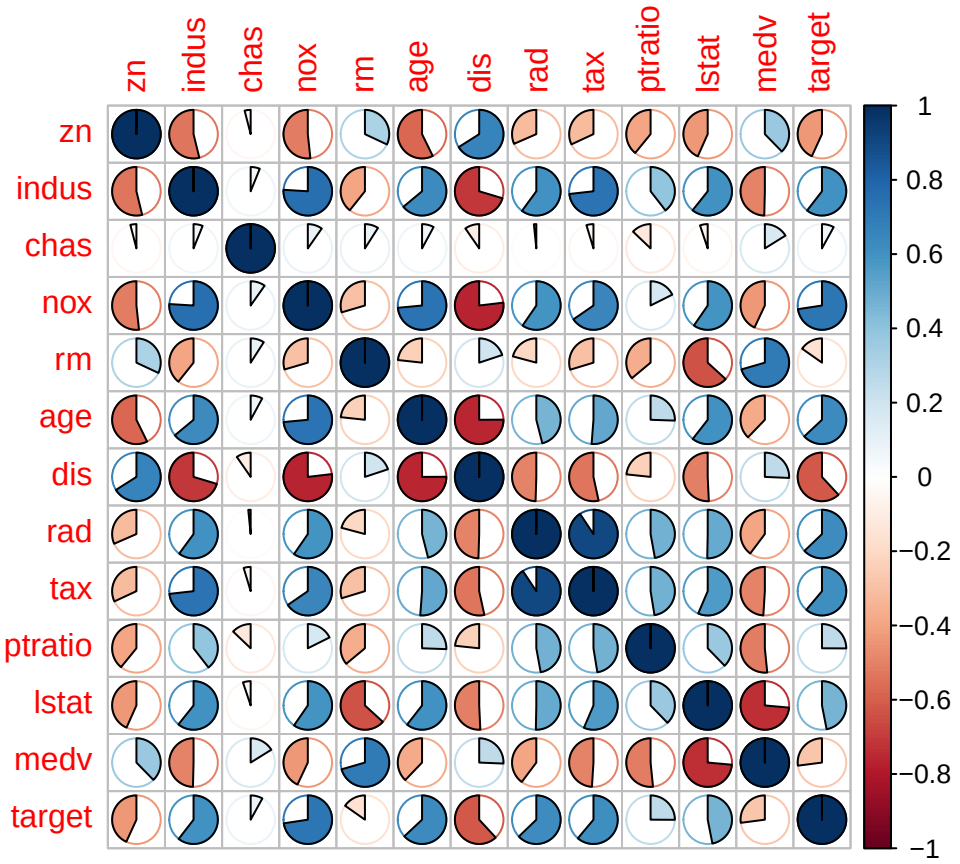


Figure 3 : Correlation matrix plot



In *Figure 2* we can see near normal distributions for medv and rm. With lstat,dis,rad,tax, and nox farther off. We may need to transform them. In particular we see dis,lstat, and age victim of a stride and true right skew.

In *Figure 3*

We can see strong positive correlations in the following sets.

indus, age, rad, tax, lstat, target :: nox

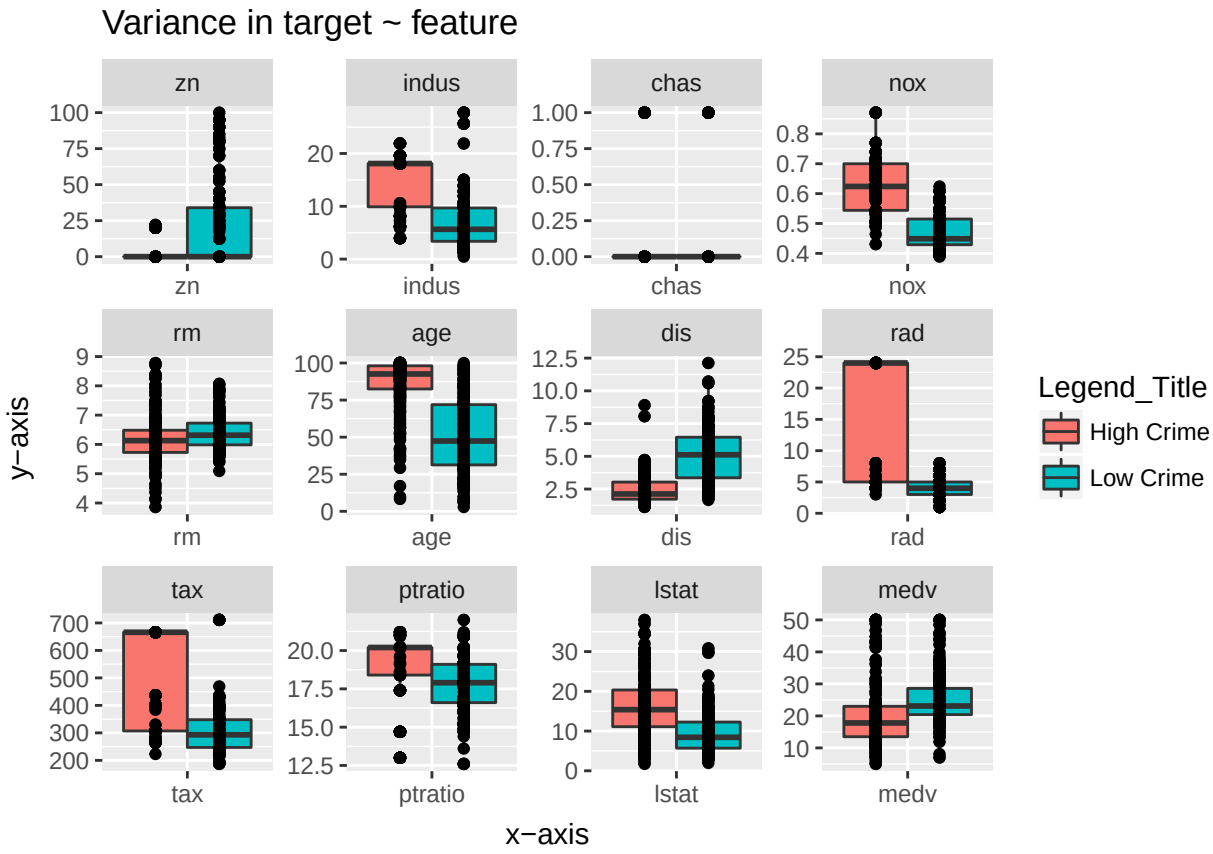
We can see strong negative correlations in the following sets.

indus, nox, age, rad, tax, target :: dis

Nitrogen Oxide levels and Distance to employment centers seem to be two major players in this soon-to-be classifier. Both opposed to each other.

There appears to be some issues of multi-collinearity between Highway accessibility index and Property Tax seeing as their correlations are a little too strong. Lets keep an eye on that. I don't think we'll need to take a variable out or transform either one to compensate. When we get to the GLM I would recommend using an interactive term.

Figure 4 : Boxplots of target ~ feature variance



The following figure shows us that in addition to fixing the distributional right skews of dis, lstat and age we should consider fixing zn, nox, tax, and rad through transformations to minimize their target variation.

2. DATA PREPERATION

To prep this data, I'm going to make three datasets. I didn't find anything worthy of binning; no insane skews worthy of datapoint capping and 0 missing data to work with.

Variables we don't touch [zn,chas,target]

1) Original dataset with transformed [dis,lstat,age] with log
2) A log normalized dataset as a lazy baseline
3) Transform [dis,lstat,age,zn,nox,tax,rad] with both quadratic and log terms. Together that is more than half the dataset which may make me a perpetrator of overfitting. I should be able to find a working model with these features + base dataset.

3. BUILD MODELS

I want to build 4 models and branch out from there.
1) Standard dataset model, all features included 2) Log transformed dataset model, all features included 3) Added transformed log and quadratic terms ontop of standard dataset 4) Standard dataset model important

features (found from previous 3 models) + interactive property tax and highway accessibility term as noted in the data exploration

```
##
## Call:
## glm(formula = target ~ ., family = binomial, data = dataset1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8464  -0.1445  -0.0017   0.0029   3.4665
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -40.822934   6.632913  -6.155 7.53e-10 ***
## zn           -0.065946   0.034656  -1.903  0.05706 .
## indus        -0.064614   0.047622  -1.357  0.17485
## chas          0.910765   0.755546   1.205  0.22803
## nox          49.122297   7.931706   6.193 5.90e-10 ***
## rm           -0.587488   0.722847  -0.813  0.41637
## age           0.034189   0.013814   2.475  0.01333 *
## dis           0.738660   0.230275   3.208  0.00134 **
## rad           0.666366   0.163152   4.084 4.42e-05 ***
## tax          -0.006171   0.002955  -2.089  0.03674 *
## ptratio       0.402566   0.126627   3.179  0.00148 **
## lstat         0.045869   0.054049   0.849  0.39608
## medv          0.180824   0.068294   2.648  0.00810 **
## ---
## 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: 192.05  on 453  degrees of freedom
## AIC: 218.05
##
## Number of Fisher Scoring iterations: 9
```

Our original dataset model with all variables has an AIC of 218.05 with a sample over 400, meaning it may be an accurate metric over BIC.

Null deviance of almost 650. Significant variables include [rad,nox, dis,medv, ptratio].

```
##
## Call:
## glm(formula = target ~ ., family = binomial, data = dataset2)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.90738  -0.16285  -0.00161   0.08746   3.13907
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -15.96595   11.72606  -1.362  0.17333
## zn           -0.03216    0.02687  -1.197  0.23145
## indus         0.24931    0.52151   0.478  0.63262
## chas          0.86893    0.73997   1.174  0.24028
```

```
## nox          24.13637    3.91814   6.160 7.27e-10 ***
## rm            2.11032    3.70451   0.570  0.56891
## age           0.85132    0.63686   1.337  0.18131
## dis           2.69643    0.84395   3.195  0.00140 **
## rad           2.98142    0.68476   4.354 1.34e-05 ***
## tax          -1.51088    1.12043  -1.348  0.17750
## ptratio       5.52739    2.12059   2.607  0.00915 **
## lstat         0.17708    0.66971   0.264  0.79146
## medv          2.32688    1.48654   1.565  0.11751
## ---
## 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: 206.10  on 453  degrees of freedom
## AIC: 232.1
##
## Number of Fisher Scoring iterations: 8
```

AIC up, residual deviance up. Again significant variables are nox and tax.

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
##
## Call:
## glm(formula = target ~ ., family = binomial, data = dataset3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1703  -0.1209   0.0000   0.0089   3.8317
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.339e+01  2.913e+01   0.460 0.645677
## zn           1.068e-01  2.030e-01   0.526 0.598755
## indus        2.002e-01  1.169e-01   1.712 0.086861 .
## chas        -5.510e-01  9.694e-01  -0.568 0.569783
## nox         -2.920e+02  1.182e+02  -2.469 0.013537 *
## rm          -1.962e+00  1.024e+00  -1.916 0.055373 .
## age          8.780e-02  3.888e-02   2.258 0.023925 *
## dis         -3.540e+00  1.268e+00  -2.791 0.005251 **
## rad          1.150e-01  4.250e-01   0.271 0.786639
## tax          2.332e-01  7.049e-02   3.308 0.000939 ***
## ptratio      7.702e-01  2.055e-01   3.748 0.000178 ***
## lstat        3.121e-01  1.723e-01   1.812 0.070047 .
## medv         3.201e-01  1.105e-01   2.898 0.003757 **
## ldis         1.598e+01  4.899e+00   3.263 0.001104 **
## llstat      -3.835e+00  2.219e+00  -1.728 0.083914 .
## lage        -1.873e+00  1.530e+00  -1.224 0.221133
## zn2         -4.618e-03  7.091e-03  -0.651 0.514853
## nox2         3.243e+02  1.136e+02   2.854 0.004315 **
## tax2        -3.718e-04  1.099e-04  -3.382 0.000719 ***
## rad2         8.662e-02  3.181e-02   2.723 0.006471 **
## ---
```

```
## 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: 130.53  on 446  degrees of freedom
## AIC: 170.53
##
## Number of Fisher Scoring iterations: 12
```

As predicted, model 3 is overfitted (I believe thats what the error message means 'Fitted probabilities numerically 0 or 1 occurred). However the residual deviance is at an all time low of 130. I may want to go back here and use a step-wise model fitting algorithm. Lets see the last base model.

```
##
## Call:
## glm(formula = target ~ dis + tax:rad + medv + dis + age, data = dataset1)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.64490  -0.20830  -0.03949   0.20205   1.09900
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.683e-02  1.186e-01   0.142 0.887182
## dis         -4.478e-02  1.170e-02  -3.826 0.000148 ***
## medv         3.782e-03  1.924e-03   1.966 0.049856 *
## age          5.779e-03  8.722e-04   6.626 9.66e-11 ***
## tax:rad      3.141e-05  3.008e-06  10.442  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1128629)
##
##     Null deviance: 116.47  on 465  degrees of freedom
## Residual deviance:  52.03  on 461  degrees of freedom
## AIC: 312.81
##
## Number of Fisher Scoring iterations: 2
```

Model 4 seems to be a clear winner (for now) With residual deviance roughly 1/4 the size of model 1, with estimated 150% dataloss (judging from AIC) I like this model because it doesn't overfit like model 3. A very rough and right estimator... Im proud of this one. Lets not give up on model 3 though.

```
## Start:  AIC=170.53
## target ~ zn + indus + chas + nox + rm + age + dis + rad + tax +
##     ptratio + lstat + medv + ldis + llstat + lage + zn2 + nox2 +
##     tax2 + rad2
```

```
##
## Call:  glm(formula = target ~ zn + indus + chas + nox + rm + age + dis +
##     rad + tax + ptratio + lstat + medv + ldis + llstat + lage +
##     zn2 + nox2 + tax2 + rad2, family = binomial, data = dataset3)
##
## Coefficients:
## (Intercept)           zn        indus         chas          nox
```

```
##   1.339e+01    1.068e-01    2.002e-01   -5.510e-01   -2.920e+02
##          rm          age          dis          rad          tax
##  -1.962e+00    8.780e-02   -3.540e+00    1.150e-01    2.332e-01
##     ptratio        lstat         medv         ldis       llstat
##   7.702e-01    3.121e-01    3.201e-01    1.598e+01   -3.835e+00
##        lage          zn2         nox2         tax2         rad2
##  -1.873e+00   -4.618e-03    3.243e+02   -3.718e-04    8.662e-02
##
## Degrees of Freedom: 465 Total (i.e. Null);  446 Residual
## Null Deviance:       645.9
## Residual Deviance: 130.5     AIC: 170.5
```

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
461	52.02979	NA	NA	NA
446	130.53317	15	-78.50338	NA

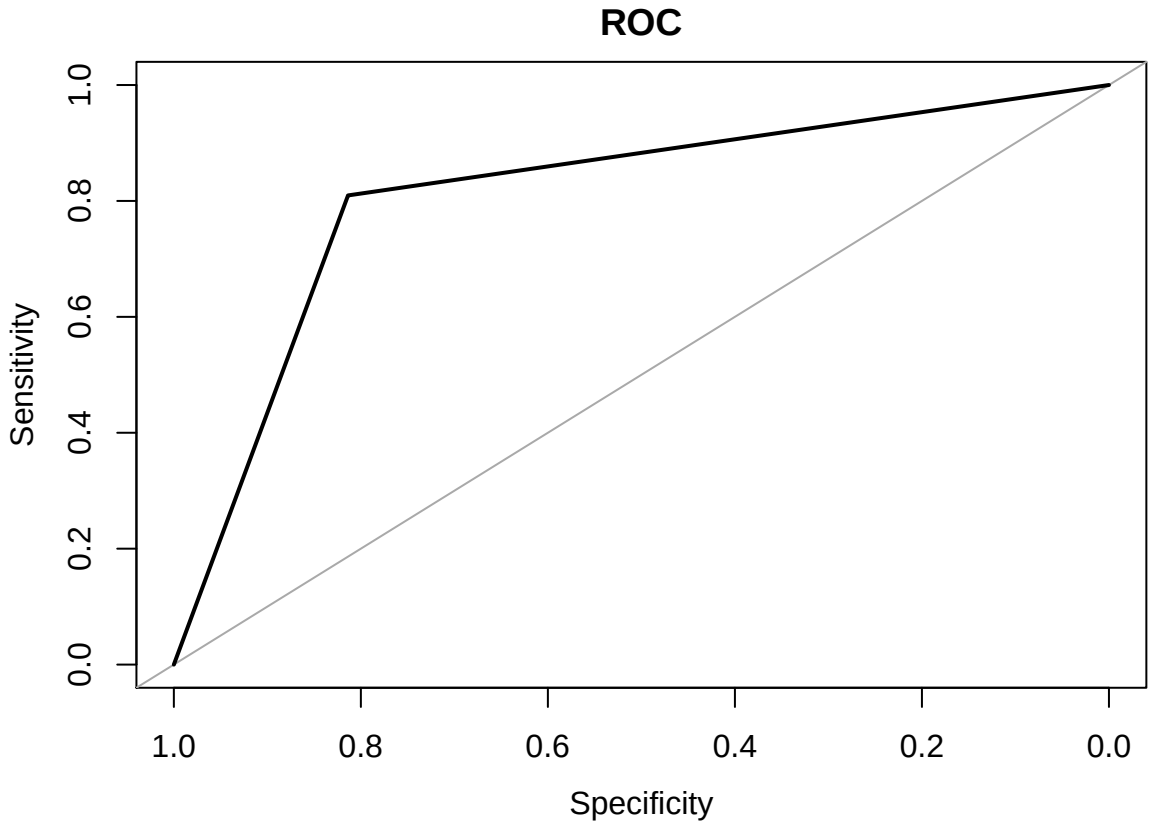
Verdict : Model 4 (Interactive tax:rad + significant features) has almost double the AIC and a third of the residual deviance with no bad indicators from ANOVA. While our 19 feature freak-beast overfitted model 3 has a lower AIC after forward step selection, it fails in comparison to goodness of fit to model 4 and ultimately, we want to shy away from too many features in any model.

4. SELECT MODELS

We've selected model 4. Less residual deviance and less features. Lets evaluate it.

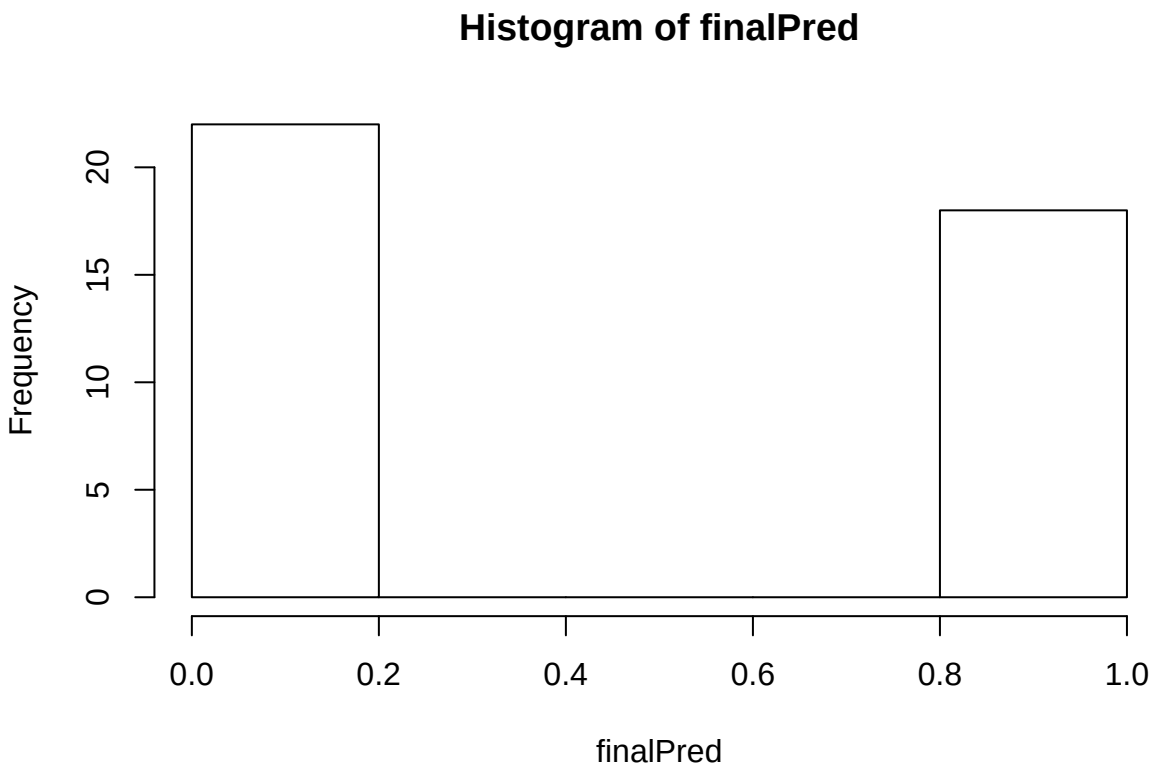
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 83 16
##          1 19 68
##
##                Accuracy : 0.8118
##                  95% CI : (0.7481, 0.8653)
##     No Information Rate : 0.5484
##     P-Value [Acc > NIR] : 3.925e-14
##
##                   Kappa : 0.6213
##  Mcnemar's Test P-Value : 0.7353
##
##             Sensitivity : 0.8095
##             Specificity : 0.8137
##          Pos Pred Value : 0.7816
##          Neg Pred Value : 0.8384
##              Prevalence : 0.4516
##          Detection Rate : 0.3656
##    Detection Prevalence : 0.4677
##       Balanced Accuracy : 0.8116
##
##        'Positive' Class : 1
##
```

Accuracy of .81 with minimum type I & II errors, 0 P-value. Around .8 Sensitivity and Specificity. All this using near 4 features.




```
## Area under the curve: 0.8116
```

AUC at .81 is fantastic at a .5 threshhold. Lets send in our predictions.



Code Appendix

```
knitr::opts_chunk$set(echo = FALSE)
#https://stackoverflow.com/questions/9341635/check-for-installed-packages-before-running-install-packag
requiredPackages = c('knitr','prettydoc','kableExtra','ggplot2','tidyr','plyr','dplyr','psych','corrplo
for(p in requiredPackages){
  if(!require(p,character.only = TRUE)) install.packages(p)
  library(p,character.only = TRUE)
}
#Load Data
train = read.csv('crime-train.csv')
#Simple way to find missing data metrics (too many variables for the mice plot to map on a PDF)
descriptions = c('Residential land zoning','Industry land zoning','Borders River','Nitrogen Oxide level
myVars = data.frame(
  abbre = names(train),
  descr = descriptions
)
columnNames = c('Feature','Description')
colnames(myVars) = columnNames
knitr::kable(myVars,row.names = TRUE)
ting = describe(train)
ting = ting[,-c(2,6,7,8,9)]
knitr::kable(ting,row.names = FALSE)
#load Hmisc for a multi-hist plot
library(Hmisc)
hist(train[1:9],na.big = FALSE)
hist(train[9:13],na.big = FALSE)
```

```r
#Creating a correlation matrix to address multi-colinearity issues
correlationMatrix = cor(train, use='complete.obs')
corrplot(correlationMatrix, method="pie")
#https://stackoverflow.com/questions/14604439/plot-multiple-boxplot-in-one-graph/14606549
train2= train
train2$target = ifelse(train$target==1,'High Crime','Low Crime')
train.m = melt(train2,id.var='target')
require(ggplot2)
p = ggplot(data = train.m, aes(x=variable, y=value))
p = p + geom_boxplot(aes(fill = target))
# if you want color for points replace group with colour=Label
p = p + geom_point(aes(y=value, group=target), position = position_dodge(width=0.75))
p = p + facet_wrap( ~ variable, scales="free")
p = p + xlab("x-axis") + ylab("y-axis") + ggtitle("Variance in target ~ feature")
p = p + guides(fill=guide_legend(title="Legend_Title"))
p

trainT = data.frame(
ldis = log(train$dis),
llstat = log(train$lstat),
lage = log(train$age),
zn2 = train$zn^2,
nox2 = train$nox^2,
tax2 = train$tax^2,
rad2 = train$rad^2)
dataset1 = train
dataset2 = log(train)
dataset2$zn = train$zn
dataset2$target = train$target
dataset2$chas = train$chas
dataset3 = cbind(train,trainT)
#Model 1 : Original dataset
m1 = glm(data=dataset1,target~.,family=binomial)
summary(m1)
#Model 2 : Log transformed dataset
m2 = glm(data=dataset2,target~.,family=binomial)
summary(m2)
#Mddel 3 : Log and Quad transformed dataset
m3 = glm(data=dataset3,target~.,family=binomial)
summary(m3)
m4 = glm(data=dataset1,target~ dis+tax:rad+medv+dis+age)
summary(m4)
m5 = step(m3,direction = 'forward')
m5
anova(m4,m5, test="Chisq")
#https://www.r-bloggers.com/evaluating-logistic-regression-models/
trainE = createDataPartition(train$target,p=.6,list=FALSE)
trainingData = train[ trainE, ]
testingData = train[ -trainE, ]
pred4 = predict(m4, newdata=testingData)
pred4 = ifelse(pred4<.5,0,1)
theMatrix = confusionMatrix(data=pred4,testingData$target,positive = '1')
theMatrix
```

```
theRock = roc(testingData$target, pred4)
plot(theRock,asp=NA,main='ROC')
theRock$auc
#Load Data
test = read.csv('crime-eval.csv')
finalPred =predict(m4, newdata=test)
finalPred = ifelse(finalPred<.5,0,1)
hist(finalPred)
write.csv(finalPred,'MMullerPredictions.csv')
```