Data 621 Homework 3

Layla Quinones

10/24/2021

Libraries

```
library(tidyverse)
library(ggplot2)
library(VIM)
library(GGally)
library(caret)
library(broom)
```

EDA

```
# Load data
# Training
rawTrain <- read.csv("https://raw.githubusercontent.com/MsQCompSci/Data621Group4/main/HW3/crime-training

#Testing data
rawTest <- read.csv("https://raw.githubusercontent.com/MsQCompSci/Data621Group4/main/HW3/crime-evaluati
```

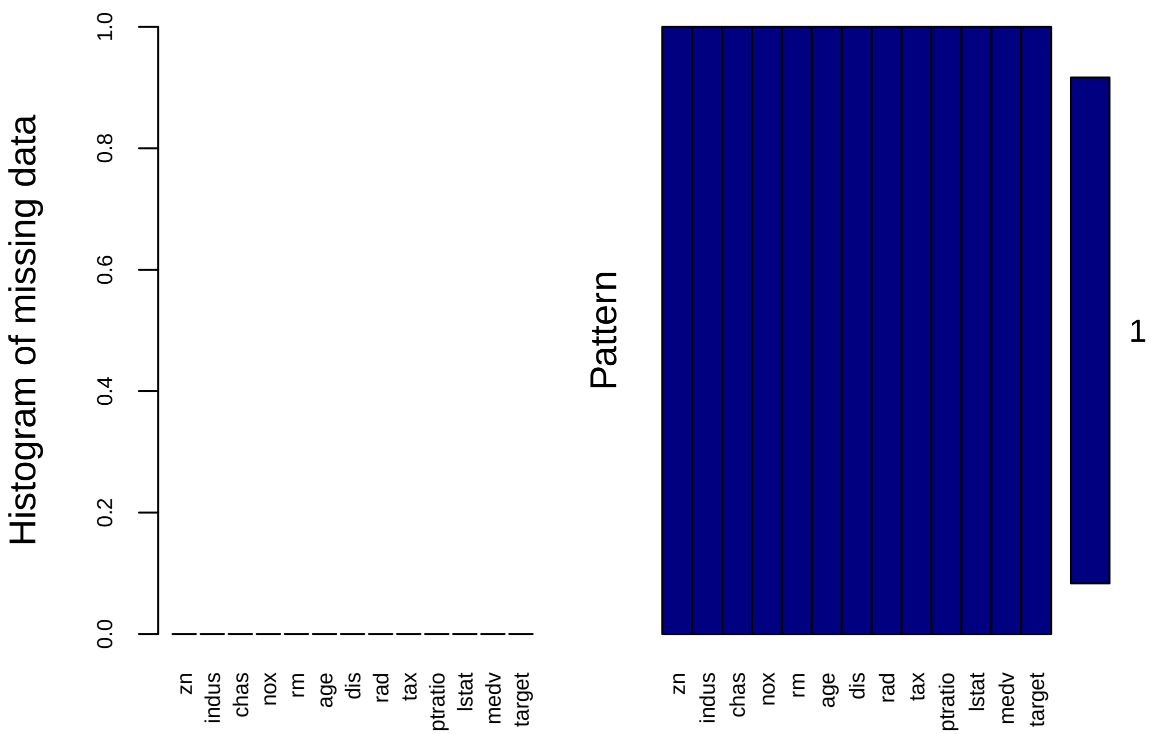
```
# check to see if we need to clean the data
# gives us a sense of what each predictor is
glimpse(rawTrain)
```

```
## Rows: 466
## Columns: 13
## $ zn      <dbl> 0, 0, 0, 30, 0, 0, 0, 0, 0, 80, 22, 0, 0, 22, 0, 0, 100, 20...
## $ indus   <dbl> 19.58, 19.58, 18.10, 4.93, 2.46, 8.56, 18.10, 18.10, 5.19, ...
## $ chas    <int> 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ nox     <dbl> 0.605, 0.871, 0.740, 0.428, 0.488, 0.520, 0.693, 0.693, 0.5...
## $ rm      <dbl> 7.929, 5.403, 6.485, 6.393, 7.155, 6.781, 5.453, 4.519, 6.3...
## $ age     <dbl> 96.2, 100.0, 100.0, 7.8, 92.2, 71.3, 100.0, 100.0, 38.1, 19...
## $ dis     <dbl> 2.0459, 1.3216, 1.9784, 7.0355, 2.7006, 2.8561, 1.4896, 1.6...
## $ rad     <int> 5, 5, 24, 6, 3, 5, 24, 24, 5, 1, 7, 5, 24, 7, 3, 3, 5, 5, 2...
## $ tax     <int> 403, 403, 666, 300, 193, 384, 666, 666, 224, 315, 330, 398,...
## $ ptratio <dbl> 14.7, 14.7, 20.2, 16.6, 17.8, 20.9, 20.2, 20.2, 20.2, 16.4,...
## $ lstat   <dbl> 3.70, 26.82, 18.85, 5.19, 4.82, 7.67, 30.59, 36.98, 5.68, 9...
## $ medv    <dbl> 50.0, 13.4, 15.4, 23.7, 37.9, 26.5, 5.0, 7.0, 22.2, 20.9, 2...
## $ target  <int> 1, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 1,...
```

```
# All varaibles are numeric
# categorical variables
# chas

#dicrete
#rad, zn, tax

#all others are continuous
```

No Missing Values

```
#plot missing values using VIM package
aggr(rawTrain , col=c('navyblue','red'), numbers=TRUE, sortVars=TRUE, labels=names(rawTrain), cex.axis=
```

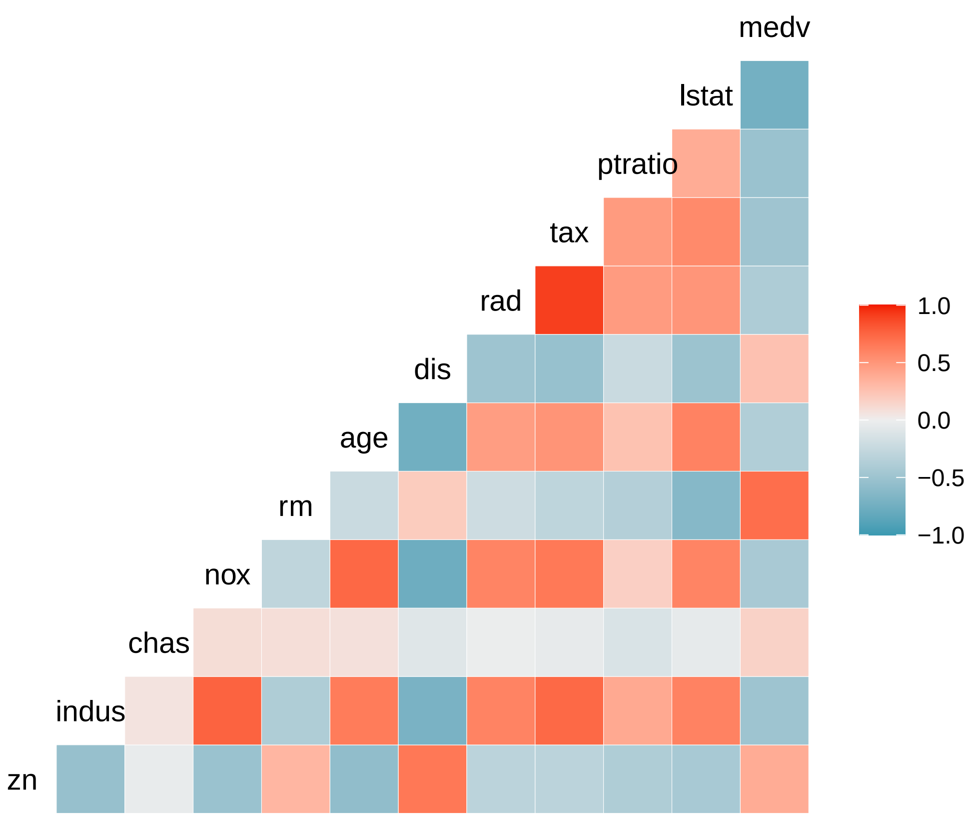



```
##
##  Variables sorted by number of missings:
##  Variable Count
##        zn     0
##     indus     0
##      chas     0
##       nox     0
##        rm     0
##       age     0
```

```
##       dis     0
##       rad     0
##       tax     0
##   ptratio     0
##     lstat     0
##      medv     0
##    target     0
```

Correlation

```
#correlation matrix for predictors
ggcorr(rawTrain%>% select(zn:medv))
```




```
#Lets look at some highly correlated variables and drop them
findCorrelation(cor(rawTrain%>% select(zn:medv)),
                cutoff = 0.75,
                verbose = TRUE,
                names = TRUE)
```

```
## Compare row 2  and column  4 with corr  0.76
##   Means:  0.539 vs 0.416 so flagging column 2
## Compare row 4  and column  7 with corr  0.769
##   Means:  0.487 vs 0.395 so flagging column 4
```

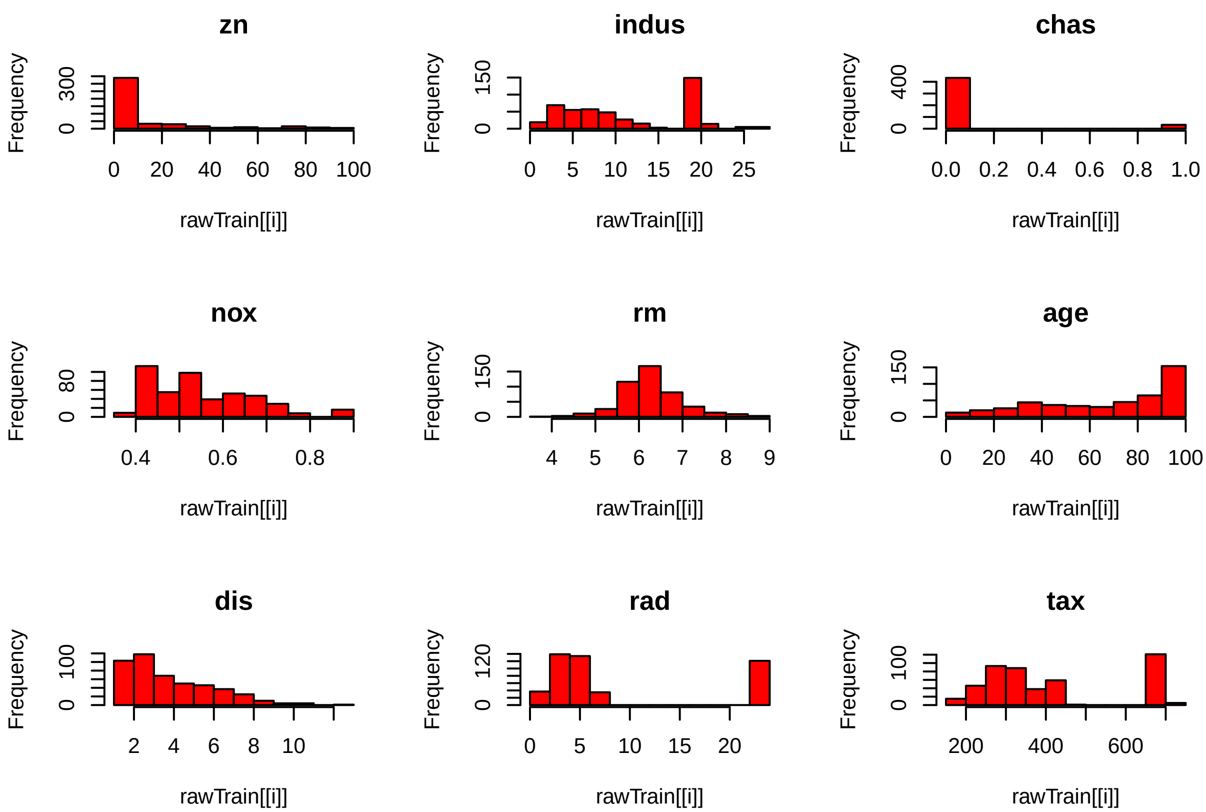
```
## Compare row 9  and column  8 with corr  0.906
##   Means:  0.46 vs 0.377 so flagging column 9
## Compare row 6  and column  7 with corr  0.751
##   Means:  0.417 vs 0.357 so flagging column 6
## All correlations <= 0.75

## [1] "indus" "nox"   "tax"   "age"
```

```
# There are 4 highly correlated variables
# I will drop the highest one which is tax which seems to be the most highly correlated
#tax and rad are 0.9 correlated lets look at their relationship to the predictor to see which one to dr
```

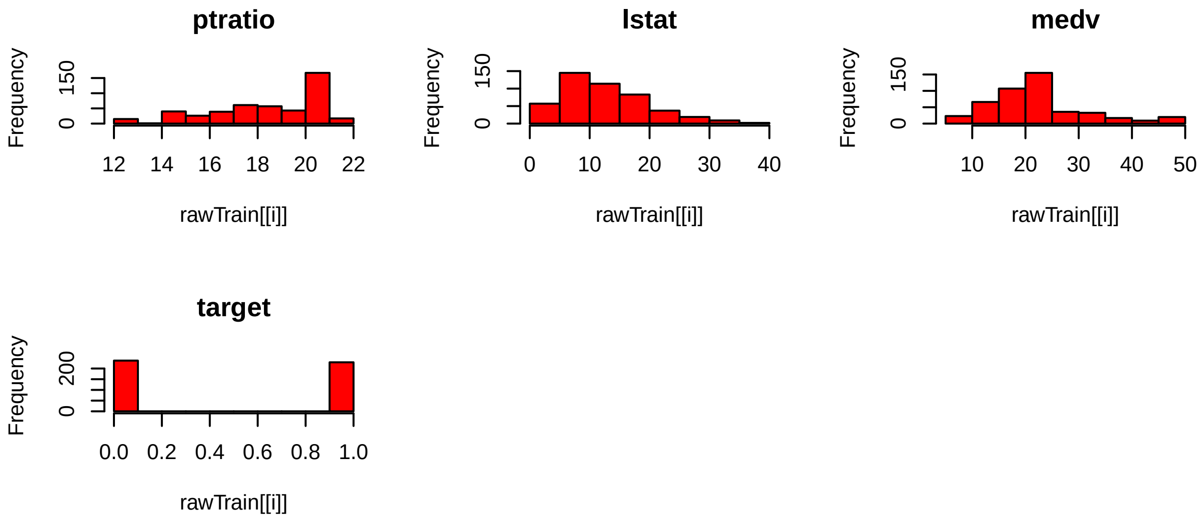
Distribution of Predictors

ADD VARIANCE AND INFLATION FACTORS TO THIS SECTION?

```
par(mfrow = c(3,3))
for(i in 1:ncol(rawTrain)) {#distribution of each variable
  hist(rawTrain[[i]], main = colnames(rawTrain[i]), col = "red")
}
```

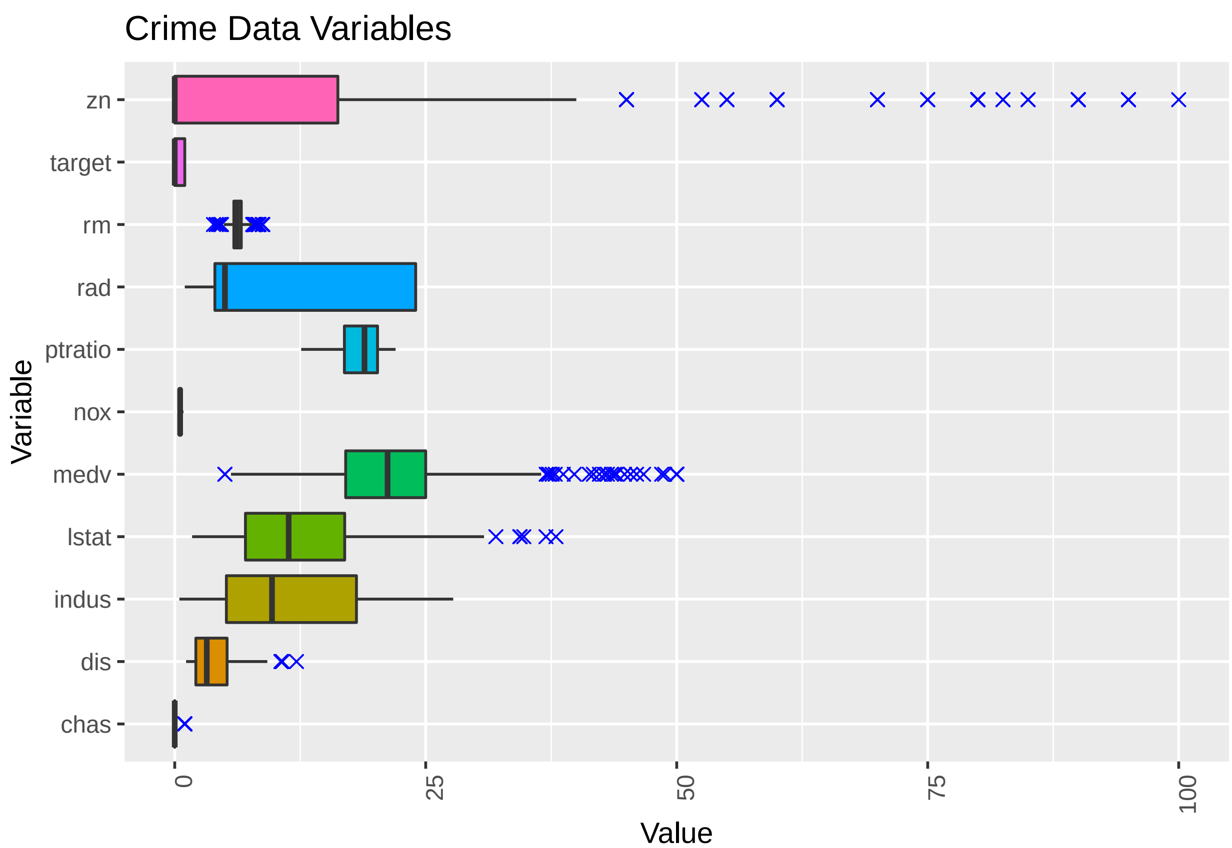



```
#binomial data
# indus, tax and rad

#all other variables ar skewed excpet RM
```



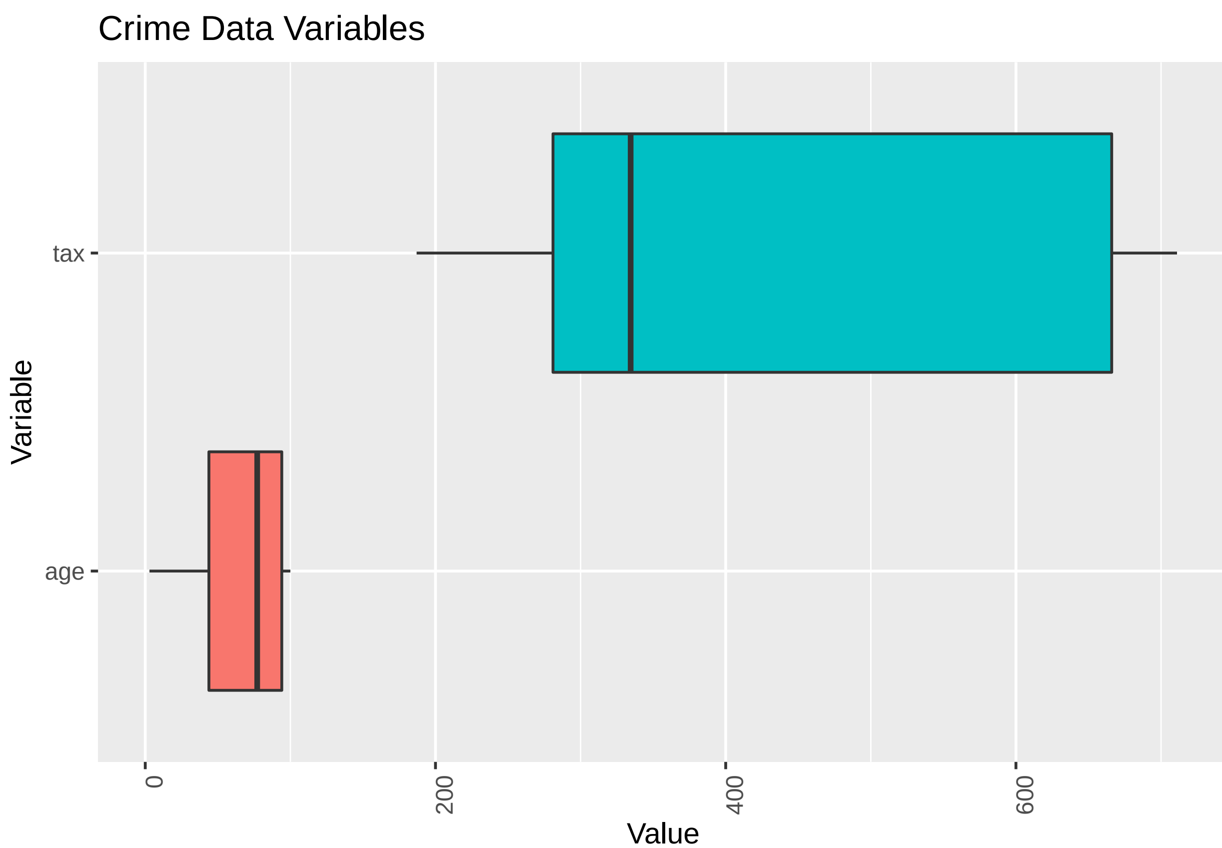
Box Plots

```
#make long
#tax and age has a much different scale so we are seperating it here
longData <- rawTrain %>%
  select(-tax, -age) %>%
  gather(key = Variable, value = Value)

# generate boxplot to identify outliers
ggplot(longData, aes(Variable, Value, fill = Variable)) +
  geom_boxplot(outlier.colour="blue",
               outlier.shape=4,
               outlier.size=2,
               show.legend=FALSE) +
   theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
   coord_flip()+
  labs(title="Crime Data Variables", y="Value")
```




```
#we can see that zn, medv and lstat has MANY outliers
```

```
#make long
#tax and age has a much different scale so we are seperating it here
longData <- rawTrain %>%
  select(tax, age) %>%
  gather(key = Variable, value = Value)

# generate boxplot to identify outliers
ggplot(longData, aes(Variable, Value, fill = Variable)) +
  geom_boxplot(outlier.colour="blue",
               outlier.shape=4,
               outlier.size=2,
               show.legend=FALSE) +
   theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
   coord_flip()+
  labs(title="Crime Data Variables", y="Value")
```




```
# no outliers for tax and age
```

```
#Train/Test Split

dt <- createDataPartition(iris$Species, p = .8,
                                  list = FALSE,
                                  times = 1)
train<-rawTrain[dt,]
test<-rawTrain[-dt,]
```

Model Building

```
#remove Tax due to high correlation with other variables
modelOne <- glm(target ~ zn + indus + chas + nox + rm + age + dis + rad + ptratio + lstat + medv , data

modelOne
```

```
##
## Call:  glm(formula = target ~ zn + indus + chas + nox + rm + age + dis +
##     rad + ptratio + lstat + medv, family = "binomial", data = train)
##
## Coefficients:
```

```
## (Intercept)          zn       indus        chas         nox          rm
##  -52.655519   -0.026868   -0.002801    2.257303   62.174802   -1.205111
##         age         dis         rad     ptratio       lstat        medv
##    0.039700    0.849402    0.564755    0.595263   -0.017266    0.299196
##
## Degrees of Freedom: 119 Total (i.e. Null);  108 Residual
## Null Deviance:       166.2
## Residual Deviance: 52.5  AIC: 76.5
```

```
#remove Tax squared age and log lstat
modelTwo <- glm(target ~ zn + indus + chas + nox + rm + age^2 + dis + rad + ptratio + log2(lstat) + medv

modelTwo
```

```
##
## Call:  glm(formula = target ~ zn + indus + chas + nox + rm + age^2 +
##     dis + rad + ptratio + log2(lstat) + medv, family = "binomial",
##     data = train)
##
## Coefficients:
## (Intercept)          zn       indus        chas         nox          rm
##  -5.090e+01  -2.508e-02  -2.442e-04   2.400e+00   6.311e+01  -1.349e+00
##         age         dis         rad     ptratio  log2(lstat)        medv
##   4.512e-02   8.636e-01   5.497e-01   5.960e-01  -5.277e-01   2.912e-01
##
## Degrees of Freedom: 119 Total (i.e. Null);  108 Residual
## Null Deviance:       166.2
## Residual Deviance: 52.32     AIC: 76.32
```

```
#This one has a litter lower AIC
```

```
summary(modelTwo)
```

```
##
## Call:
## glm(formula = target ~ zn + indus + chas + nox + rm + age^2 +
##     dis + rad + ptratio + log2(lstat) + medv, family = "binomial",
##     data = train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.91856  -0.21156   0.00001   0.00988   2.64672
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.090e+01  1.751e+01  -2.907  0.00365 **
## zn          -2.508e-02  5.945e-02  -0.422  0.67310
## indus       -2.442e-04  1.054e-01  -0.002  0.99815
## chas         2.400e+00  1.703e+00   1.409  0.15883
## nox          6.311e+01  2.023e+01   3.119  0.00181 **
## rm          -1.349e+00  1.570e+00  -0.859  0.39038
## age          4.512e-02  3.214e-02   1.404  0.16036
## dis          8.636e-01  4.890e-01   1.766  0.07738 .
```

```
## rad          5.497e-01  3.077e-01   1.786  0.07404 .
## ptratio      5.960e-01  2.775e-01   2.147  0.03176 *
## log2(lstat) -5.277e-01  1.188e+00  -0.444  0.65689
## medv         2.912e-01  1.507e-01   1.933  0.05327 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 166.222  on 119  degrees of freedom
## Residual deviance:  52.324  on 108  degrees of freedom
## AIC: 76.324
##
## Number of Fisher Scoring iterations: 9
```

```r
#log10(zn + 1), log10(dis) and deleted log2(lstat) - not significant
modelThree <- glm(target ~ log10(zn + 1) + indus + chas + nox + rm + age^2 + log10(dis) + rad + ptratio

modelThree
```

```
##
## Call:  glm(formula = target ~ log10(zn + 1) + indus + chas + nox + rm +
##     age^2 + log10(dis) + rad + ptratio + medv, family = "binomial",
##     data = train)
##
## Coefficients:
##   (Intercept)  log10(zn + 1)          indus           chas            nox
##     -57.68516       -0.04904        0.03638        2.04326       64.78409
##            rm            age     log10(dis)            rad        ptratio
##      -1.22613        0.03834        8.58431        0.62559        0.64765
##          medv
##       0.32052
##
## Degrees of Freedom: 119 Total (i.e. Null);  109 Residual
## Null Deviance:       166.2
## Residual Deviance: 51.74     AIC: 73.74
```

```r
#AIC is lower again (not sure if age^2 ishelpful)
```

```r
summary(modelThree)
```

```
##
## Call:
## glm(formula = target ~ log10(zn + 1) + indus + chas + nox + rm +
##     age^2 + log10(dis) + rad + ptratio + medv, family = "binomial",
##     data = train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.89617  -0.22508   0.00000   0.00645   2.73161
##
## Coefficients:
```

```
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -57.68516   18.39615  -3.136  0.00171 **
## log10(zn + 1)  -0.04904    0.97256  -0.050  0.95978
## indus           0.03638    0.10537   0.345  0.72987
## chas            2.04326    1.64699   1.241  0.21475
## nox            64.78409   20.10051   3.223  0.00127 **
## rm             -1.22613    1.52159  -0.806  0.42035
## age             0.03834    0.02504   1.531  0.12572
## log10(dis)      8.58431    4.57746   1.875  0.06075 .
## rad             0.62559    0.32538   1.923  0.05452 .
## ptratio         0.64765    0.29525   2.194  0.02827 *
## medv            0.32052    0.15540   2.063  0.03915 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 166.222  on 119  degrees of freedom
## Residual deviance:  51.736  on 109  degrees of freedom
## AIC: 73.736
##
## Number of Fisher Scoring iterations: 9
```

```r
#combine rad and rm (multiplied) - they seemed to correspond in their distributions
modelFour<- glm(target ~ log10(zn + 1) + indus + chas + nox +  age^2 + log10(dis) + rad*rm + ptratio  +
```

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

```r
modelFour
```

```
##
## Call:  glm(formula = target ~ log10(zn + 1) + indus + chas + nox + age^2 +
##     log10(dis) + rad * rm + ptratio + medv, family = "binomial",
##     data = train)
##
## Coefficients:
##   (Intercept)  log10(zn + 1)          indus           chas            nox
##     -47.99349        0.06731        0.02142        1.43536       84.80346
##           age     log10(dis)            rad             rm        ptratio
##       0.05672       10.60139       -3.21353       -6.89961        0.97156
##          medv         rad:rm
##       0.55480        0.65463
##
## Degrees of Freedom: 119 Total (i.e. Null);  108 Residual
## Null Deviance:       166.2
## Residual Deviance: 43.27     AIC: 67.27
```

```r
#AIC is lower #Not sure what the rationale is for this working but it lowered the AIC nummber and Resid
```

```r
summary(modelFour)
```

```
##
```

```
## Call:
## glm(formula = target ~ log10(zn + 1) + indus + chas + nox + age^2 +
##     log10(dis) + rad * rm + ptratio + medv, family = "binomial",
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2936  -0.1251   0.0000   0.0256   2.3054
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -47.99349   18.20694  -2.636  0.00839 **
## log10(zn + 1)   0.06731    1.21579   0.055  0.95585
## indus           0.02142    0.12406   0.173  0.86292
## chas            1.43536    1.65112   0.869  0.38467
## nox            84.80346   26.02913   3.258  0.00112 **
## age             0.05672    0.02855   1.987  0.04697 *
## log10(dis)     10.60139    5.11300   2.073  0.03813 *
## rad            -3.21353    1.22734  -2.618  0.00884 **
## rm             -6.89961    2.94837  -2.340  0.01928 *
## ptratio         0.97156    0.36410   2.668  0.00762 **
## medv            0.55480    0.20955   2.648  0.00811 **
## rad:rm          0.65463    0.24015   2.726  0.00641 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 166.222  on 119  degrees of freedom
## Residual deviance:  43.275  on 108  degrees of freedom
## AIC: 67.275
##
## Number of Fisher Scoring iterations: 9
```

```
#delte indus
modelFive<-glm(target ~ log10(zn+1)+ nox +  age^2 + log10(dis) + rad*rm + ptratio  + medv, data = train
```

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

```
modelFive
```

```
##
## Call:  glm(formula = target ~ log10(zn + 1) + nox + age^2 + log10(dis) +
##     rad * rm + ptratio + medv, family = "binomial", data = train)
##
## Coefficients:
##   (Intercept)  log10(zn + 1)            nox            age     log10(dis)
##     -42.71566       -0.15640       83.51526        0.06028       10.48536
##           rad             rm        ptratio           medv         rad:rm
##      -3.46991       -7.65572        0.93527        0.58392        0.70641
##
## Degrees of Freedom: 119 Total (i.e. Null);  110 Residual
## Null Deviance:       166.2
## Residual Deviance: 44.28     AIC: 64.28
```

```
#AIC is higher #resiudal deviance is lower
# I looked at the histograms and looked for complementary shapes to decide what to multiply
```

Variable importance

```
summary(modelFive)
```

```
##
## Call:
## glm(formula = target ~ log10(zn + 1) + nox + age^2 + log10(dis) +
##     rad * rm + ptratio + medv, family = "binomial", data = train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.42393  -0.13335   0.00000   0.02269   2.09935
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -42.71566   15.80686  -2.702 0.006885 **
## log10(zn + 1)  -0.15640    1.15468  -0.135 0.892255
## nox            83.51526   24.53273   3.404 0.000663 ***
## age             0.06028    0.02787   2.163 0.030549 *
## log10(dis)     10.48536    4.73263   2.216 0.026722 *
## rad            -3.46991    1.24937  -2.777 0.005481 **
## rm             -7.65572    2.91597  -2.625 0.008654 **
## ptratio         0.93527    0.35468   2.637 0.008365 **
## medv            0.58392    0.20464   2.853 0.004325 **
## rad:rm          0.70641    0.24556   2.877 0.004019 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 166.222  on 119  degrees of freedom
## Residual deviance:  44.278  on 110  degrees of freedom
## AIC: 64.278
##
## Number of Fisher Scoring iterations: 9
```

```
#indus and zn are not important
```

```
#multiply ptratio*nox (remove squared from age)
modelSix<- glm(target ~ log10(zn + 1) + age  + ptratio*nox + log10(dis) + rad*rm + medv, data = train, f
```

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

```
modelSix
```

```
##
```

```
## Call:  glm(formula = target ~ log10(zn + 1) + age + ptratio * nox +
##     log10(dis) + rad * rm + medv, family = "binomial", data = train)
##
## Coefficients:
##   (Intercept)  log10(zn + 1)            age        ptratio            nox
##     -56.70701       -0.16656        0.06145        1.66511      107.77925
##    log10(dis)            rad             rm           medv    ptratio:nox
##      10.43924       -3.43004       -7.58642        0.58249       -1.31103
##        rad:rm
##       0.70009
##
## Degrees of Freedom: 119 Total (i.e. Null);  109 Residual
## Null Deviance:       166.2
## Residual Deviance: 44.21     AIC: 66.21
```

```r
#AIC is lower
```

```r
summary(modelSix)
```

```
##
## Call:
## glm(formula = target ~ log10(zn + 1) + age + ptratio * nox +
##     log10(dis) + rad * rm + medv, family = "binomial", data = train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.40341  -0.13098   0.00000   0.02148   2.14347
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -56.70701   55.90754  -1.014  0.31044
## log10(zn + 1)  -0.16656    1.18003  -0.141  0.88775
## age             0.06145    0.02835   2.167  0.03021 *
## ptratio         1.66511    2.80987   0.593  0.55345
## nox           107.77925   96.72027   1.114  0.26513
## log10(dis)     10.43924    4.74580   2.200  0.02783 *
## rad            -3.43004    1.26249  -2.717  0.00659 **
## rm             -7.58642    2.94975  -2.572  0.01011 *
## medv            0.58249    0.20669   2.818  0.00483 **
## ptratio:nox    -1.31103    4.98898  -0.263  0.79272
## rad:rm          0.70009    0.24742   2.830  0.00466 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 166.222  on 119  degrees of freedom
## Residual deviance:  44.207  on 109  degrees of freedom
## AIC: 66.207
##
## Number of Fisher Scoring iterations: 9
```

Test Models

```
#Make predictions
predOne = predict(modelOne,test, type = "response")
predTwo = predict(modelTwo,test, type = "response")
predThree = predict(modelThree,test, type = "response")
predFour = predict(modelFour,test, type = "response")
predFive = predict(modelFive,test, type = "response")
predSix = predict(modelSix,test, type = "response")
```

```
#Error Measures
data.frame(modelOne = postResample(pred = predOne, obs = test$target), modelTwo = postResample(pred = pr
```

```
##           modelOne  modelTwo modelThree modelFour modelFive  modelSix
## RMSE     0.2761763 0.2761245  0.2715408 0.2801529 0.2776759 0.2763455
## Rsquared 0.6982113 0.6984770  0.7079434 0.6933252 0.6978356 0.7006364
## MAE      0.1254664 0.1247825  0.1220058 0.1168108 0.1170055 0.1159272
```

```
#We can see RMSE is increasing which means the fit is better for every model - This doesnt reflect very
```

Confusion Matrix and Accuracy Measurment

```
#Extract Accuracy

#Model One
#format predictions to binary
resultsFitOne <- ifelse(predOne > 0.5,1,0)
resultsFitOne <- as.factor(resultsFitOne)

#Confusion Matrix to Extract Accuracy
cOne <- confusionMatrix(as.factor(test$target),resultsFitOne)
accOne <- as.data.frame(cOne$overall)[1]
accOne<- accOne %>%
  slice(1)


#Model Two
#format predictions to binary
resultsFitTwo <- ifelse(predTwo > 0.5,1,0)
resultsFitTwo <- as.factor(resultsFitTwo)

#Confusion Matrix to Extract Accuracy
cTwo <- confusionMatrix(resultsFitTwo, as.factor(test$target))
accTwo <- as.data.frame(cTwo$overall)[1]
accTwo<- accTwo %>%
  slice(1)

#Model Three
#format predictions to binary
```

```r
resultsFitThree<- ifelse(predThree > 0.5,1,0)
resultsFitThree <- as.factor(resultsFitThree)

#Confusion Matrix to Extract Accuracy
cThree <- confusionMatrix(resultsFitThree, as.factor(test$target))
accThree <- as.data.frame(cThree$overall)[1]
accThree<- accThree%>%
  slice(1)

#Model Four
#format predictions to binary
resultsFitFour<- ifelse(predFour > 0.5,1,0)
resultsFitFour <- as.factor(resultsFitFour)

#Confusion Matrix to Extract Accuracy
cFour <- confusionMatrix(resultsFitFour, as.factor(test$target))
accFour <- as.data.frame(cFour$overall)[1]
accFour<- accFour%>%
  slice(1)

#Model Five
#format predictions to binary
resultsFitFive<- ifelse(predFive > 0.5,1,0)
resultsFitFive <- as.factor(resultsFitFive)

#Confusion Matrix to Extract Accuracy
cFive <- confusionMatrix(resultsFitFive, as.factor(test$target))
accFive <- as.data.frame(cFive$overall)[1]
accFive<- accFive%>%
  slice(1)


#Model Six
#format predictions to binary
resultsFitSix<- ifelse(predSix > 0.5,1,0)
resultsFitSix <- as.factor(resultsFitSix)

#Confusion Matrix to Extract Accuracy
cSix<- confusionMatrix(resultsFitSix, as.factor(test$target))
accSix <- as.data.frame(cSix$overall)[1]
accSix<- accSix%>%
  slice(1)
```

```r
#create a table with accuracies
data.frame(c(accOne, accTwo, accThree, accFour,accFive, accSix))
```

```
##   cOne.overall cTwo.overall cThree.overall cFour.overall cFive.overall
## 1    0.8872832    0.8872832      0.8959538     0.8872832     0.8988439
##   cSix.overall
## 1    0.8988439
```

```
#Here we see that our best models are Five and Six in terms of accuracy
```

WE NEED QQ PLOTS OR SOME OTHER VISUAL TO HELP US TALK ABOUT GOODNESS OF FIT GETTING HIGHER ALTHOUGH THE ACCURACY IS NOT CHANGING SO WE CAN CHOOSE ONE (FIVE OR SIX)

AUC or ROC curve