Crime Data Assignment

Cesar Espitia

CUNY SPS Data 621

Table of Contents

Abstract

This assignment focused on census type data for the City of Boston. The dataset contains 466 records that encompass the entire area. The variables for the data are housing type variables such as number of rooms but the records do not give an indication of the part of the city it is representing. The purpose for this assignment is to analyze the data, perform any data manipulation / clean-up and build three (3) binary logistic regression models using only the data (or derivatives thereof) to predict if the region is above or below the median crime rate. The chosen model provided an AIC = 256.26.

Keywords: crime, data621

Crime Data Assignment

The following is the analysis and write-up based upon my interpretation of the data and predict if the region is above or below the median crime rate for the City of Boston.

Data Exploration

The purpose of this step is to get a 'feel' for the dataset. The following information describes the data from different angles including completeness, statistical summaries, visuals to determine the shape and effect of each variable and other items deemed pertinent.

Summary Statistics

The first step is to look at the data to determine some items including completeness and the shape of each variable. The following are the results of summarizing the data in a table and the visualization of each variables density function (PDF).

Table 1

Summary Statistics for Moneyball Training Data

Variable	Min	1Q	Med	Mean	3Q	Max
zn	0	0	0	11.58	16.25	100
indus	0.46	5.145	9.69	11.105	18.1	27.74
chas	0	0	0	0.0708	0	1
nox	0.389	0.448	0.538	0.5543	0.624	0.871
rm	3.863	5.887	6.21	6.291	6.63	8.78
age	2.9	43.88	77.15	68.37	94.1	100
dis	1.13	2.101	3.191	3.796	5.215	12.127
rad	1	4	5	9.53	24	24
tax	187	281	334.5	409.5	666	711
ptratio	12.6	16.9	18.9	18.4	20.2	22
black	0.32	375.61	391.34	357.12	396.24	396.9
lstat	1.73	7.043	11.35	12.631	16.93	37.97
medv	5	17.02	21.2	22.59	25	50
target	0	0	0	0.4914	1	1

Note: Source: moneyball-training-data.csv

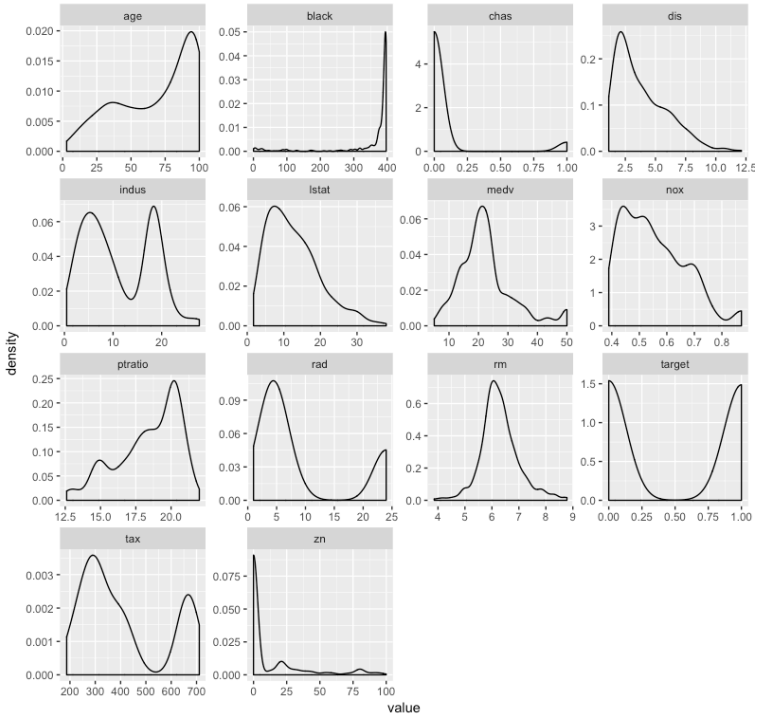


Figure 1. PDF for Each Dataframe Variable.

In looking at both, Table 1, Figure 1 and Appendix B (correlation matrix) together, we can note specific items that may skew our model building results. In this model, no data was missing when looking at nulls and complete cases.

PDF: Figure 1 shows the PDF of each variable, this allows us to see if the data is normal or night. Only one (1) variable (***rm*** – average rooms) shows the typical normal density function but all others like ***zn*** show left skewness and other show bimodality. For the purposes of this analysis, the variables ***zn, black and chas*** will be log transformed around each variable's mean to remove the effects of skewness. All other variables were left as is because the shape didn't warrant it.

Correlation: We look for correlated variables that we can make decisions on and determine which variable might be closely related to others either due to collinearity or other

underlying factors that are visible at first glance in the dataset. Correlated variables bloat the model and don't produce any more insight than ignoring one of the two that show correlation. In our data, ***rad*** is correlated to ***tax*** with a Pearson correlation estimate of 0.91. In looking at the nature of the variables, accessibility to highways seems less important than the full-value property-tax rate for this exercise.

Data Preparation

The purpose of this step is to take the findings from the exploration and transform the data as needed. The following information describes the transformations done in order to prepare the data for model building and model selection.

NA: For this dataset, no missing values were found in any of the columns and therefore no imputations were needed for this analysis.

Log Transformation: For this dataset, three (3) variables were transformed that were deemed overtly skewed in comparison to other variables in the dataset. ***zn, black and chas*** were the variables transformed using the log base 10 function, for example for the ***zn*** column the transformation was *log(train$zn+1)*. The value 1 was added in order to ensure that values of 0 continue to be 0 after the transformation (as $\log_{10}(0)$ is not possible).

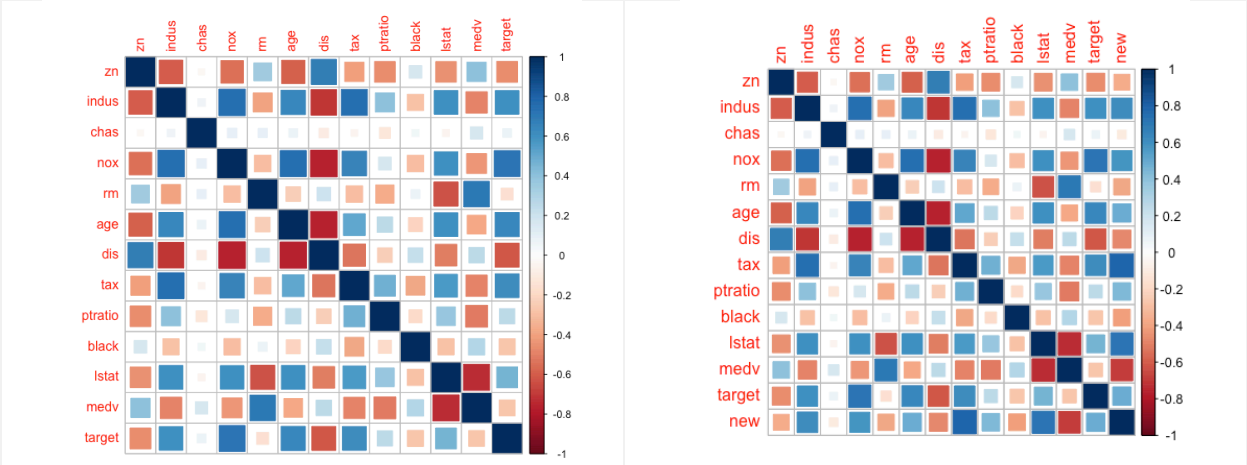
Variable Creation: For this dataset, one (1) variable was created that was deemed unique to the data. ***new*** was created by taking ***tax*** and dividing it by a modified ***medv***. The reasoning behind this variable is that there is a correlation to tax rate of a home to the median value of the home. This can help identify up and coming neighborhoods indirectly by determining if this ratio is less than 1 (undervalued) vs over 1 (overvalued). This was taking by using the following equation:

$$new = \frac{tax}{medv * 10}$$

Medv was multiplied by 10 as tax is in $10,000's while medv is in $1,000's.

Variable Deletion / Data Deletion: For this dataset, only one variable was removed that was deemed unnecessary for model building. As noted in the exploration step on Page 6, ***rad*** was removed from the dataset as it was highly correlated to ***tax***.

Correlation Check: Once these manipulations are done, a side-by-side comparison of the correlations matrix is done to ensure no inadvertent effects to the data.



Before | After

Figure 2. Correlation Comparison Before and After.

The addition of variable ***new*** did not affect the correlation for the data set and is appropriate to use in model building.

Model Building

The purpose of this step is to take the modified dataset and begin exploring potential models that will be used on the final dataset provided. The following information describes the

three (3) models built for this step and the relevant analysis to provide reasons for model selection in the next step.

MODEL 1

The first model takes in the data as manipulated in step two. In this first model, we have an AIC of 251.59. The data in Table 2, shows that the model has an accuracy of 90.1%.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1097  -0.2702  -0.0188   0.1965   3.5752

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -33.029280   8.409645  -3.928 8.58e-05 ***
zn           -0.681913   0.248420  -2.745 0.006051 **
indus        -0.152009   0.049530  -3.069 0.002148 **
chas          2.500407   0.970167   2.577 0.009958 **
nox          51.588709   7.272073   7.094 1.30e-12 ***
rm           -0.093566   0.620763  -0.151 0.880190
age           0.024907   0.012537   1.987 0.046958 *
dis           0.922461   0.220580   4.182 2.89e-05 ***
tax           0.008238   0.002404   3.427 0.000611 ***
ptratio       0.298144   0.113540   2.626 0.008642 **
black        -1.785172   1.031215  -1.731 0.083428 .
lstat         0.087574   0.052940   1.654 0.098088 .
medv          0.181463   0.061169   2.967 0.003011 **
new          -0.425892   0.194726  -2.187 0.028732 *
---
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: 223.59  on 452  degrees of freedom
AIC: 251.59

Number of Fisher Scoring iterations: 8
```

Variable Interpretation:

(Intercept)	zn	indus	chas	nox	age	dis	tax	ptratio	medv	new
-3.24	-0.05	-0.01	0.22	3.90	0.00	0.06	0.00	0.02	0.01	-0.02

Table 2. *Confusion Matrix Model 1*

True \ Pred	0	1
0	213	24
1	22	207

No variables seem peculiar expect for nox, ths variable stands out extensively in the data with its coefficient at over 3 meaning there is a 393% chance that the area has crime above the median in the city as this variable increase. This is an odd variable to have to indicate crime unless the theory that lack of clean air is a predictor of the crime potential in a city. For now, this variable will be left in.

MODEL 2

The second model only takes into account the variables noted of significance from Model 1 (p-value < 0.05). This means that ***chas, black, lstat and rm*** will be removed from the dataset for Model 2. In this second model, we have an AIC of 256.26. The data in Table 3, shows that the model has an accuracy of 89.7%.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3341  -0.3000  -0.0271   0.2216   3.5120

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -40.559519   5.624242  -7.212 5.53e-13 ***
zn           -0.570579   0.227004  -2.514 0.011953 *
indus        -0.141446   0.049062  -2.883 0.003939 **
chas          2.791745   0.970675   2.876 0.004026 **
nox          48.903546   6.885627   7.102 1.23e-12 ***
age           0.027388   0.010057   2.723 0.006465 **
dis           0.804755   0.203539   3.954 7.69e-05 ***
tax           0.008311   0.002432   3.418 0.000632 ***
ptratio       0.289929   0.107363   2.700 0.006925 **
medv          0.130716   0.033228   3.934 8.36e-05 ***
new          -0.280439   0.208775  -1.343 0.179187
---
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: 234.26  on 455  degrees of freedom
AIC: 256.26

Number of Fisher Scoring iterations: 7
```

Variable Interpretation:

(Intercept)	zn	indus	chas	nox	age	dis	tax	ptratio	medv	new
-3.24	-0.05	-0.01	0.22	3.90	0.00	0.06	0.00	0.02	0.01	-0.02

Table 3. *Confusion Matrix Model 2*

True \ Pred	0	1
0	214	27
1	25	204

In this model, the created variable ***new*** no longer has significance due to the fact that ***medv*** and ***tax*** are still in the model.

MODEL 3

The third model is my personal PC model. This dataset struck a nerve with me because as upcoming Data Scientists we can inadvertently introduce our own unconscious bias into datasets, models and algorithms. Similar to the first kodak films that were unconsciously configured for lighter skin tones, data and models can be inadvertently created to disenfranchise one particular segment of the population vs the other especially when these models or algorithms impact individuals livelihoods in the forms of loans or other items in society. An example of this is the increase use of models/algorithms in presetting bail values which have already raised alerts in the community for potential bias that needs to be controlled for. For this model, all variables related in any way shape or form to racial bias were removed (***black and lstat***) and the dummy variable chas was also removed. In this third model, we have an AIC of 264.17. The data in Table 3, shows that the model has an accuracy of 88.6%.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3623  -0.3016  -0.0318   0.2470   3.4208

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -35.640886   5.181413  -6.879 6.04e-12 ***
zn           -0.607165   0.230874  -2.630 0.008542 **
indus        -0.122350   0.045692  -2.678 0.007413 **
nox          44.778082   6.398918   6.998 2.60e-12 ***
rm           -0.332101   0.508476  -0.653 0.513673
age           0.030652   0.010570   2.900 0.003732 **
dis           0.771242   0.203098   3.797 0.000146 ***
tax           0.007710   0.002318   3.326 0.000880 ***
ptratio       0.231692   0.105821   2.189 0.028563 *
medv          0.151192   0.054648   2.767 0.005663 **
new          -0.250426   0.197261  -1.270 0.204257
---
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: 242.17  on 455  degrees of freedom
AIC: 264.17

Number of Fisher Scoring iterations: 7
```

Variable Interpretation:

(Intercept)	zn	indus	nox	rm	age	dis	tax	ptratio	medv	new
-2.94	-0.05	-0.01	3.69	-0.03	0.00	0.06	0.00	0.02	0.01	-0.02

Table 4. *Confusion Matrix Model 3*

True \ Pred	0	1
0	210	27
1	26	203

The increase in AIC was not expected and seems counter intuitive. Although the accuracy only decreases by another 1.1% any of these three models would seem the most appropriate. My specific mantra is that the least amount of variables (simplicity) is always better than complexity is part of the criteria in addition to AIC and the accuracy in prediction. In this case, Model 2 still ties with Model 3 with the least amount of variables at 10 but other variables make Model 2

more appropriate as described in the next section. With this in mind, Model 2 so far seems the most appropriate so far.

Model Selection

The purpose of this step is to take the models built and develop a ranking criterion to select the final model for use with the test data set. The following information describes the methodology and the results on the test data.

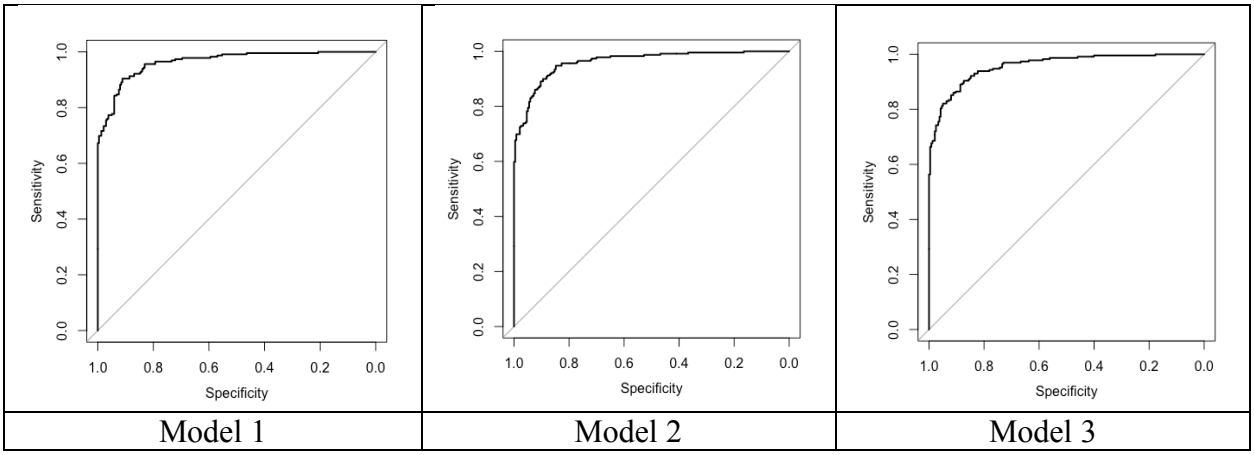
METHODOLOGY

Familiarity with the dataset subject is low and therefore the methodology will be more closely related to the statistical information presented. In this case, a combination of four (4) factors (AIC, Percent Accuracy, ROC Curve and Number of Variables) will be the criteria to select the model. The reason for this is that the significance of each variable is high in Model 1 through 3 as the adjustments for correlation and log transformations were already taken care of in Step 2 of the process. If Step 2 had not been done, then it would have been hidden in the model building and taken care of between Model 1 and Model 2. In addition, because this is a binary predictive exercise accuracy is also important for this exercise as seen in Table 5 below.

Table 5. *Model Criteria Selection*

Criteria	Model 1 (All Variables)	**Model 2 (Significant Variables Only)**	Model 3 (Politically Correct)
AIC	251.59	**256.26**	264.17
Accuracy %	90.1	**89.7**	88.6
No. of Vars	13	**10**	10

Of importance also is the ROC Curves for each model which tell us if the model predictive capability is better than just chance (a coin-toss at 50/50). In looking at each curve blow in Figure 3, we can see that the ROC curve for model 2 has a better smoother transition in comparison to Model 1 and Model 3. Overall, the ROC curve for Model 2 trends to the upper left quadrant in a more evenly distributed manner versus the other two (2).



Model 1	Model 2	Model 3

Figure 3. ROC Curves for Each Model (Model 1 through 3).

With this in mind, Model 2 is best model with an AIC of 256.26.

TEST DATA

The dataset had 40 entries and 13 columns and was modified to fit the final variables and scaling used in Model 2 from above. This means that the same process of adjustments and log transformations was done in order to be able to use the model correctly. The final predicted values are based upon a normalized value from the test data. The data is shown as follows with the corresponding summaries for the spread of the data.

Table 6. *Predicted Statstics vs Summary of Model 2 Predicted Values*

	Min	Q1	Med	Mean	Q3	Max
Model 2	0.0002	0.0330	0.4686	0.4914	0.9663	0.9999995
On Test Data	0.0011	0.0515	0.4513	0.4887	0.9067	0.999981

This table is only meant as a comparison but it does highlight that the test data has a higher set of values that would be deemed 0 (crime is less than the city median). The spread of the data for test is also a lot tighter than the training values which may be a function of the lower number of cases available for test (90% training vs 10% test). This should definitely be closer to a 70%/30% split between training and test. For the test data, 21 cases were deemed a value of 0 (less than the median of the city) and 19 were deemed a value of 1 (more than the median of the city). This is concurrent with the confusion matrix in Table 3 that shows that 214 cases were deemed a value of 0 (true positive) and 204 were deemed a value of 1 (true negative).

Conclusion

Three models were presented after exploring and manipulating the data as necessary. With using a multi-criteria approach for this exercise, it became clear that the Model 2 was selected and provided an AIC of 256.26 which was adequate for the data but doesn't necessarily indicate the best model if it were solely based upon AIC (Model 1 would have been chosen) which is the equivalent of R-squared for binary regression models. If more time were available, the creation of other new variables that were not correlated could have been generated with better insight into the data set.

Appendix A: R Code

```
---
title: "Data 621"
author: 'Cesar Espitia HW #3'
date: "7/1/2018"
output: html_document
---

## Data Exploration


```{r dataexploration, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
library(e1071)
library(dplyr)
library(purrr)
library(tidyr)
library(ggplot2)
library(corrplot)
library(FactoMineR)
library(VIF)
library(knitr)
library(kableExtra)
library(Hmisc)

# read data
train = read.csv(file="data/crime-training-data.csv")
train2<-train
dim(train)

#check data
summary(train) %>% kable() %>% kable_styling()
sapply(train, function(x) sum(is.na(x)))

ntrain<-select_if(train, is.numeric)
ntrain %>%
  keep(is.numeric) %>%                     # Keep only numeric columns
  gather() %>%                             # Convert to key-value pairs
  ggplot(aes(value)) +                     # Plot the values
    facet_wrap(~ key, scales = "free") +   # In separate panels
    geom_density()

rcorr(as.matrix(train))
corrplot(cor(train), method="square")

# correlation test 1
cor.test(train$rad,train$tax,method="pearson")
#significant ignore


```

Data Preparation

```{r datapreparation, include=FALSE}
# transform data using log for skewed

train$zn <- log(train$zn+1)
train$black <- log(train$black+1)
train$chas <- log(train$chas+1)


#remove rad per correlation in prior section

train <- train[, !(colnames(train) %in% c("rad"))]

#create variable
train$new <- train$tax / (train$medv*10)

rcorr(as.matrix(train))
corrplot(cor(train), method="square")

```


Build Models
```{r buildmodels, include=FALSE}

#MODEL 1
logit <- glm(target ~ ., data=train, family = "binomial" (link="logit"))
summary(logit)
exp(logit$coefficients)
logitscalar <- mean(dlogis(predict(logit, type = "link")))
logitscalar * coef(logit)

confint.default(model)
```
```

```
predlogit <- predict(logit, type="response")
summary(predlogit)

table(true = train$target, pred = round(fitted(logit)))


#extract variables that are significant and rerun model
sigvars <- data.frame(summary(logit)$coef[summary(logit)$coef[,4] <= .05, 4])
sigvars <- add_rownames(sigvars, "vars")
colist<-dplyr::pull(sigvars, vars)
colist<-colist[2:11]

idx <- match(colist, names(train))
trainmod2 <- cbind(train[,idx], train2['target'])

#MODEL 2
logit2 <- glm(target ~ ., data=trainmod2, family = "binomial" (link="logit"))
summary(logit2)
exp(logit2$coefficients)
logit2scalar <- mean(dlogis(predict(logit2, type = "link")))
logit2scalar * coef(logit2)

predlogit2 <- predict(logit2, type="response")
summary(predlogit2)

table(true = train$target, pred = round(fitted(logit2)))

#MODEL 3
#PC Model no racial bias
logit3<-model <- glm(target ~ zn + indus + nox + rm + age + dis + tax + ptratio + medv + new, data=train, family = "binomial" (link="logit"))
summary(logit3)
exp(logit3$coefficients)

predlogit3 <- predict(logit3, type="response")
summary(predlogit3)

table(true = train$target, pred = round(fitted(logit3)))

logit3scalar <- mean(dlogis(predict(logit3, type = "link")))
logit3scalar * coef(logit3)

round(logitscalar * coef(logit),2)
round(logit2scalar * coef(logit2),2)
round(logit3scalar * coef(logit3),2)
```

Select Models

```{r selectmodels, include=FALSE}

dim(test)

test = read.csv(file="data/crime-evaluation-data.csv")
test2<- test

# transform data using log for skewed

test$zn <- log(test$zn+1)
test$black <- log(test$black+1)
test$chas <- log(test$chas+1)


#remove rad per correlation in prior section

test <- test[, !(colnames(test) %in% c("rad"))]

#create variable
test$new <- test$tax / (test$medv*10)

idx <- match(colist, names(test))
testNorm2 <- test[,idx]

summary(testNorm2)

modelfinal <- predict(logit2, newdata = testNorm2, type="response")

y_pred_num <- ifelse(modelfinal > 0.5, 1, 0)
y_pred <- factor(y_pred_num, levels=c(0, 1))
summary(y_pred)

rbind(summary(predlogit2), summary(modelfinal)) %>%

```

Appendix B: CORRELATION MATRIX

	zn	indus	chas	nox	rm	age	dis	tax	ptratio	black	lstat	medv	target
zn	1	-0.6	-0.04	-0.55	0.34	-0.58	0.69	-0.4	-0.46	0.17	-0.46	0.4	-0.47
indus	-0.6	1	0.06	0.76	-0.39	0.64	-0.7	0.73	0.39	-0.28	0.61	-0.5	0.6
chas	-0.04	0.06	1	0.1	0.09	0.08	-0.1	-0.05	-0.13	0.05	-0.05	0.16	0.08
nox	-0.55	0.76	0.1	1	-0.3	0.74	-0.77	0.65	0.18	-0.29	0.6	-0.43	0.73
rm	0.34	-0.39	0.09	-0.3	1	-0.23	0.2	-0.3	-0.36	0.09	-0.63	0.71	-0.15
age	-0.58	0.64	0.08	0.74	-0.23	1	-0.75	0.51	0.26	-0.21	0.61	-0.38	0.63
dis	0.69	-0.7	-0.1	-0.77	0.2	-0.75	1	-0.53	-0.23	0.23	-0.51	0.26	-0.62
tax	-0.4	0.73	-0.05	0.65	-0.3	0.51	-0.53	1	0.47	-0.38	0.56	-0.49	0.61
ptratio	-0.46	0.39	-0.13	0.18	-0.36	0.26	-0.23	0.47	1	-0.18	0.38	-0.52	0.25
black	0.17	-0.28	0.05	-0.29	0.09	-0.21	0.23	-0.38	-0.18	1	-0.28	0.29	-0.28
lstat	-0.46	0.61	-0.05	0.6	-0.63	0.61	-0.51	0.56	0.38	-0.28	1	-0.74	0.47
medv	0.4	-0.5	0.16	-0.43	0.71	-0.38	0.26	-0.49	-0.52	0.29	-0.74	1	-0.27
target	-0.47	0.6	0.08	0.73	-0.15	0.63	-0.62	0.61	0.25	-0.28	0.47	-0.27	1