Data 621 Homework Three

Zachary Safir,Mario Pena,Hector Santana

10/27/2021

Introduction

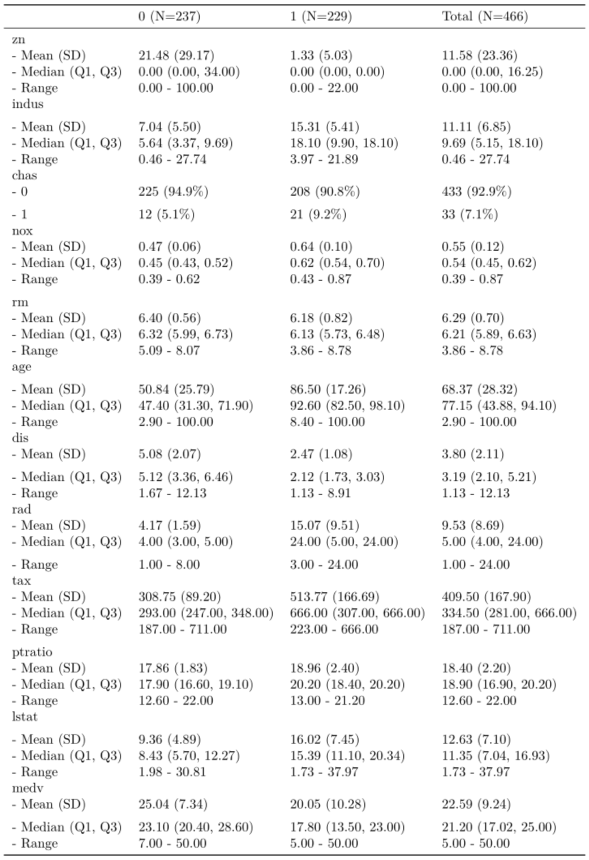
In this assignment we will explore, analyze and build a binary logistic regression model to predict wheter a particular neighborhood will be at risk for high crime levels.

We are provided with information on 466 neighborhoods, 12 predictor variables and 1 response variable. The response variable indicates whether the crime rate is above the median (1) or not (0).

Data Exploration

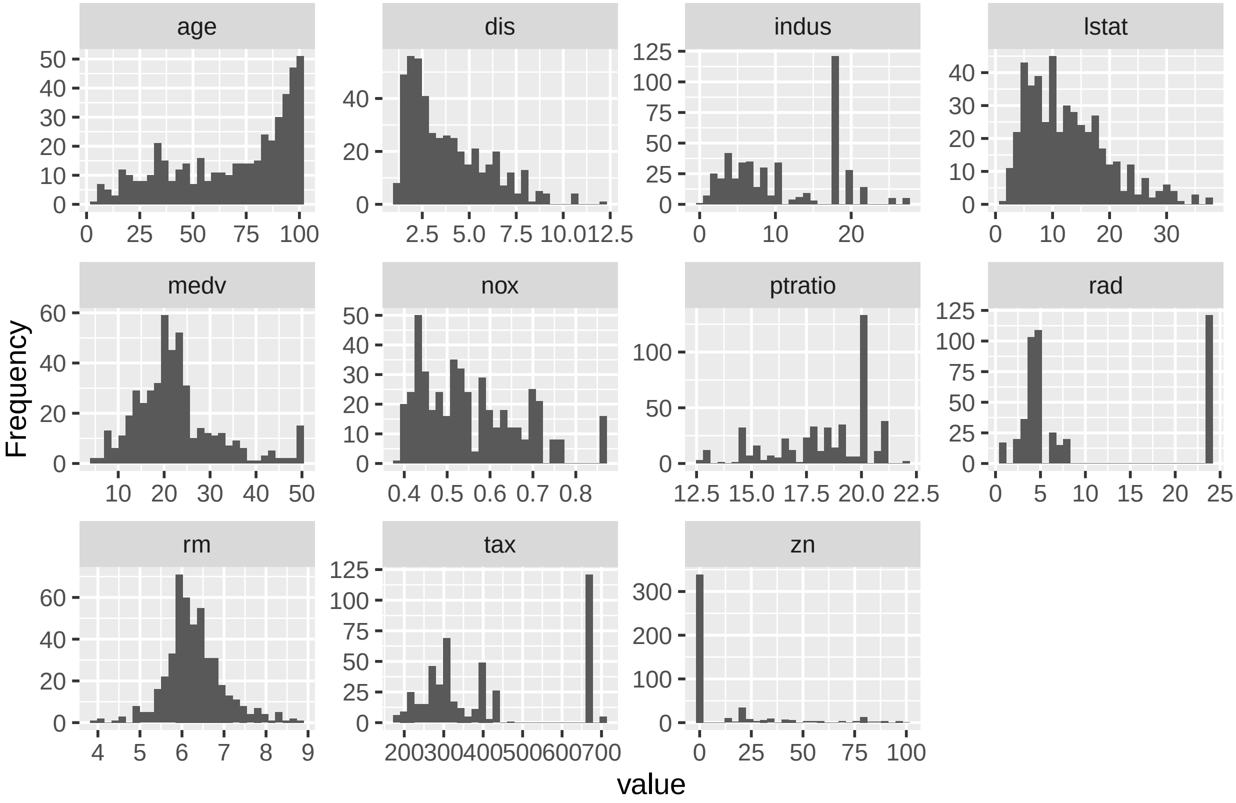
Below we have created a table with the summary statistics of our 12 predictor variables.
We ran a summary statistic analysis to determine if there were existing outliers, skew, and kurtosis in the data. The preliminary consensus was that the following variables showed skew, kurtosis, and outliers: “age”, “dis”, “indus”, “lstat”, “nox” ,“ptratio,” “rad”, “tax”.

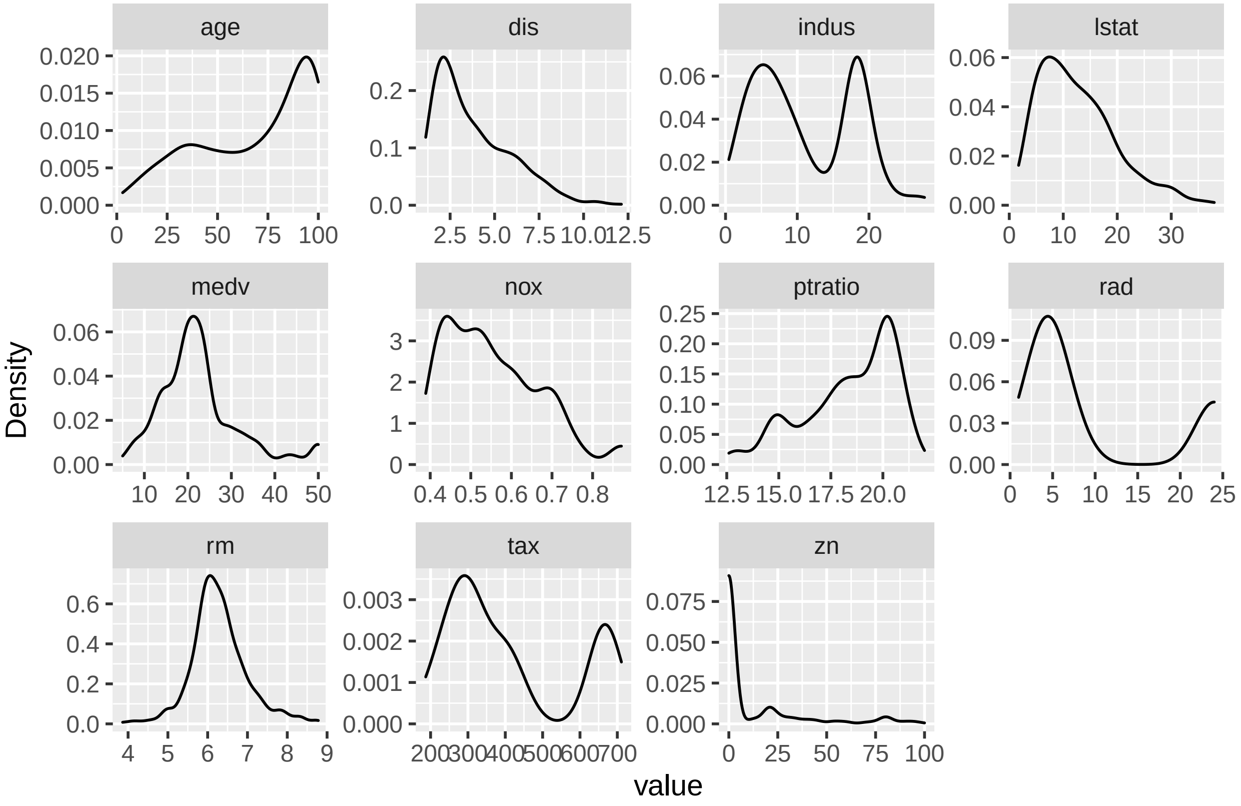
We can also observe that the “target” variable is roughly evenly distributed between the “0” and “1” responses.However, our variable chas has way too few 1 observations. It seems that will have to drop that value from consideration due to this.

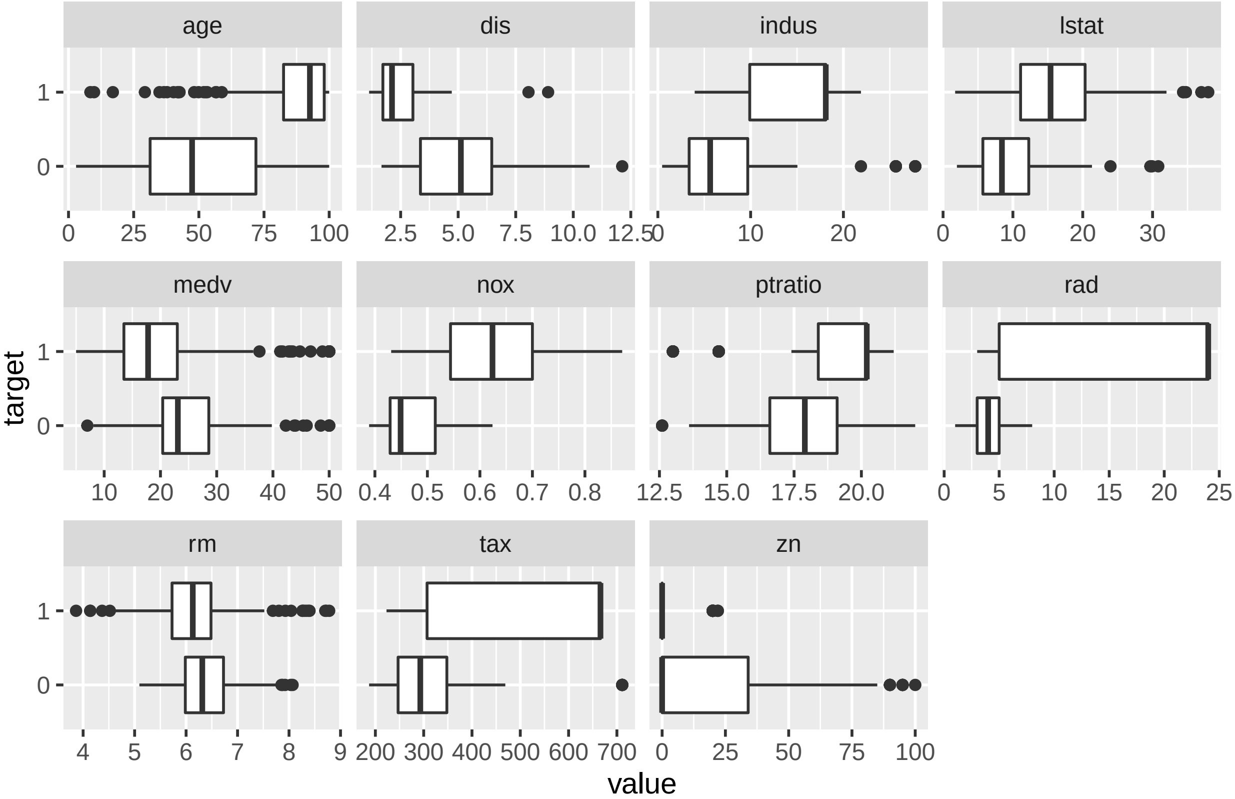


	0 (N=237)	1 (N=229)	Total (N=466)
zn			
- Mean (SD)	21.48 (29.17)	1.33 (5.03)	11.58 (23.36)
- Median (Q1, Q3)	0.00 (0.00, 34.00)	0.00 (0.00, 0.00)	0.00 (0.00, 16.25)
- Range	0.00 - 100.00	0.00 - 22.00	0.00 - 100.00
indus			
- Mean (SD)	7.04 (5.50)	15.31 (5.41)	11.11 (6.85)
- Median (Q1, Q3)	5.64 (3.37, 9.69)	18.10 (9.90, 18.10)	9.69 (5.15, 18.10)
- Range	0.46 - 27.74	3.97 - 21.89	0.46 - 27.74
chas			
- 0	225 (94.9%)	208 (90.8%)	433 (92.9%)
- 1	12 (5.1%)	21 (9.2%)	33 (7.1%)
nox			
- Mean (SD)	0.47 (0.06)	0.64 (0.10)	0.55 (0.12)
- Median (Q1, Q3)	0.45 (0.43, 0.52)	0.62 (0.54, 0.70)	0.54 (0.45, 0.62)
- Range	0.39 - 0.62	0.43 - 0.87	0.39 - 0.87
rm			
- Mean (SD)	6.40 (0.56)	6.18 (0.82)	6.29 (0.70)
- Median (Q1, Q3)	6.32 (5.99, 6.73)	6.13 (5.73, 6.48)	6.21 (5.89, 6.63)
- Range	5.09 - 8.07	3.86 - 8.78	3.86 - 8.78
age			
- Mean (SD)	50.84 (25.79)	86.50 (17.26)	68.37 (28.32)
- Median (Q1, Q3)	47.40 (31.30, 71.90)	92.60 (82.50, 98.10)	77.15 (43.88, 94.10)
- Range	2.90 - 100.00	8.40 - 100.00	2.90 - 100.00
dis			
- Mean (SD)	5.08 (2.07)	2.47 (1.08)	3.80 (2.11)
- Median (Q1, Q3)	5.12 (3.36, 6.46)	2.12 (1.73, 3.03)	3.19 (2.10, 5.21)
- Range	1.67 - 12.13	1.13 - 8.91	1.13 - 12.13
rad			
- Mean (SD)	4.17 (1.59)	15.07 (9.51)	9.53 (8.69)
- Median (Q1, Q3)	4.00 (3.00, 5.00)	24.00 (5.00, 24.00)	5.00 (4.00, 24.00)
- Range	1.00 - 8.00	3.00 - 24.00	1.00 - 24.00
tax			
- Mean (SD)	308.75 (89.20)	513.77 (166.69)	409.50 (167.90)
- Median (Q1, Q3)	293.00 (247.00, 348.00)	666.00 (307.00, 666.00)	334.50 (281.00, 666.00)
- Range	187.00 - 711.00	223.00 - 666.00	187.00 - 711.00
ptratio			
- Mean (SD)	17.86 (1.83)	18.96 (2.40)	18.40 (2.20)
- Median (Q1, Q3)	17.90 (16.60, 19.10)	20.20 (18.40, 20.20)	18.90 (16.90, 20.20)
- Range	12.60 - 22.00	13.00 - 21.20	12.60 - 22.00
lstat			
- Mean (SD)	9.36 (4.89)	16.02 (7.45)	12.63 (7.10)
- Median (Q1, Q3)	8.43 (5.70, 12.27)	15.39 (11.10, 20.34)	11.35 (7.04, 16.93)
- Range	1.98 - 30.81	1.73 - 37.97	1.73 - 37.97
medv			
- Mean (SD)	25.04 (7.34)	20.05 (10.28)	22.59 (9.24)
- Median (Q1, Q3)	23.10 (20.40, 28.60)	17.80 (13.50, 23.00)	21.20 (17.02, 25.00)
- Range	7.00 - 50.00	5.00 - 50.00	5.00 - 50.00

The insight gained from the statistical analysis permitted us to make note of further data of interest that needed to be analyzed in depth prior to the creation of our models. To confirm these irregularities we then constructed visual representations consisting of density plots, histograms, and boxplots.



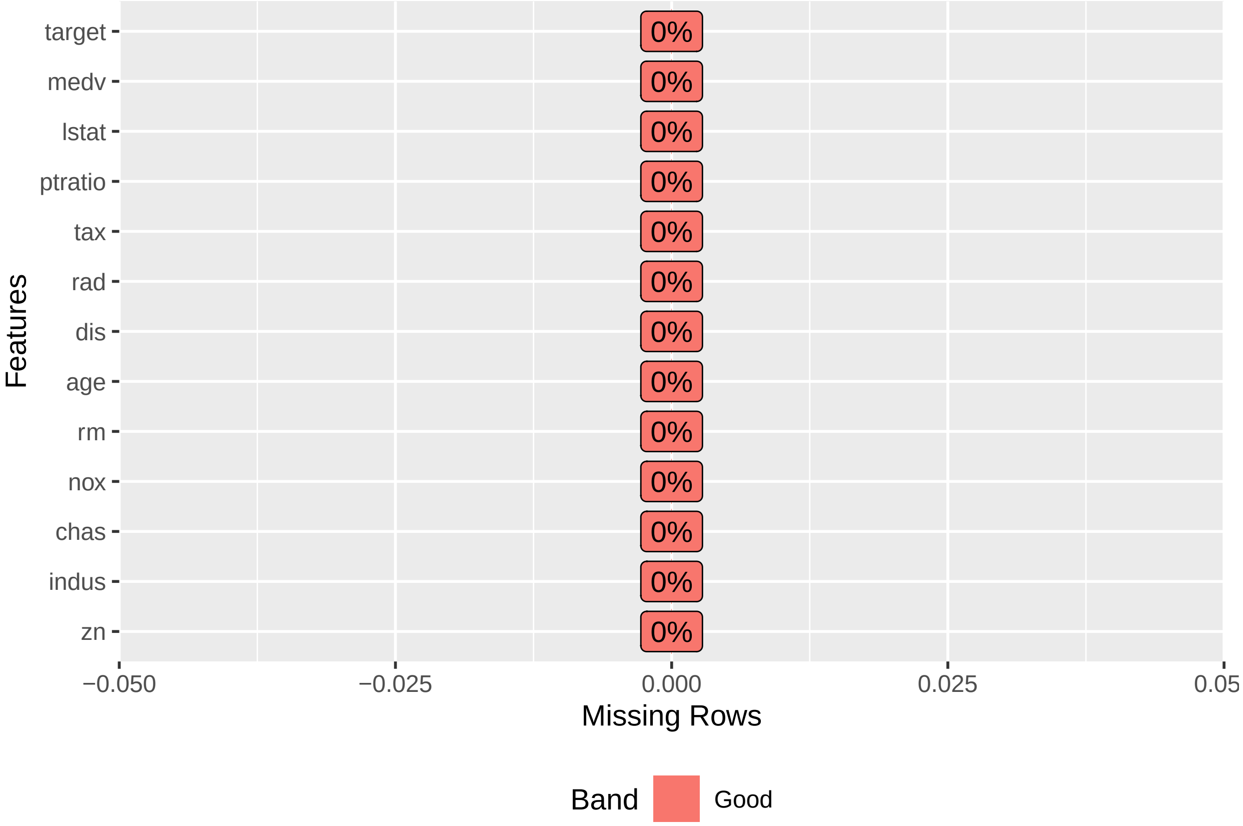




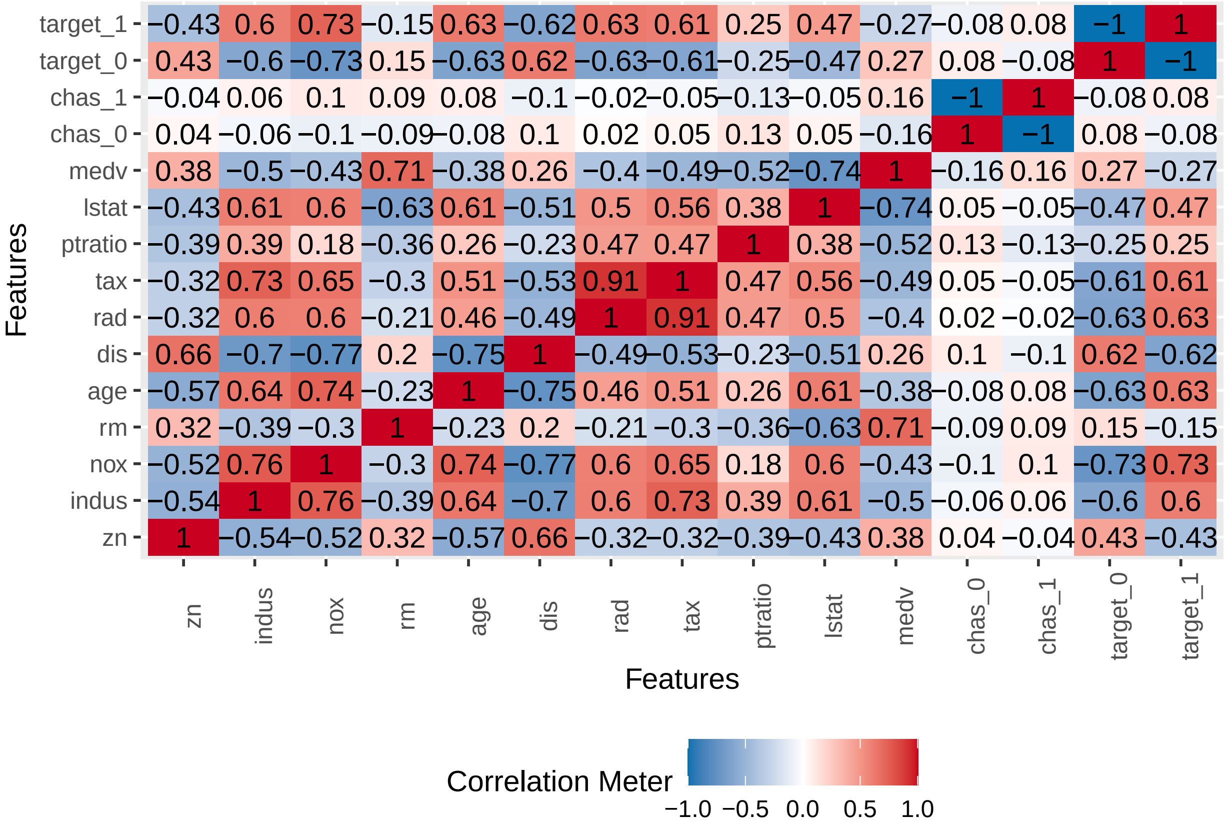
We have quite a few variables that are skewed or not normally distributed based on our plots above. We will have to do some type of transformation for some of the variables if we would like to include them in our models.

Additionally, according to our summary statistics above, and our graph below, there are no missing values in our dataset.

We note that our data has no missing values, and even upon further inspections, we find no obvious looking erroneous values in our data.



We can also observe the correlation of our variables with eachother on the next plot.



It seems that "target_1" has a strong correlation with the variables "nox", "rad", "age", "tax" and "indus". Target_0 has the inverse relationship for these same variables.

Additionally, it seems that "target_1" has a strong correlation with the variables "nox", "rad", "age", "tax" and "indus", whileTarget_0 has the inverse relationship for these same variables.

We also notice that there are some strong correlations between our predictor variables, possibly indicating a multicollinearity issue. Looking at the VIF scores, we can determine that the "medv" variable is redundant and can be removed from our data. Doing so lowers the VIF values for many of our predictors.

	VIF Score
zn	1.823146
indus	2.682271
chas	1.241479
nox	4.160497
rm	5.813851
age	2.569961
dis	3.887981
rad	1.942967
tax	2.144040
ptratio	2.275557
lstat	2.642656
medv	8.122037

	VIF Score
zn	1.677315
indus	2.625254
chas	1.174362
nox	3.596137
rm	2.327766
age	1.891963
dis	2.816354
rad	1.795920
tax	1.974189
ptratio	1.522170
lstat	2.630902

Data Preparation

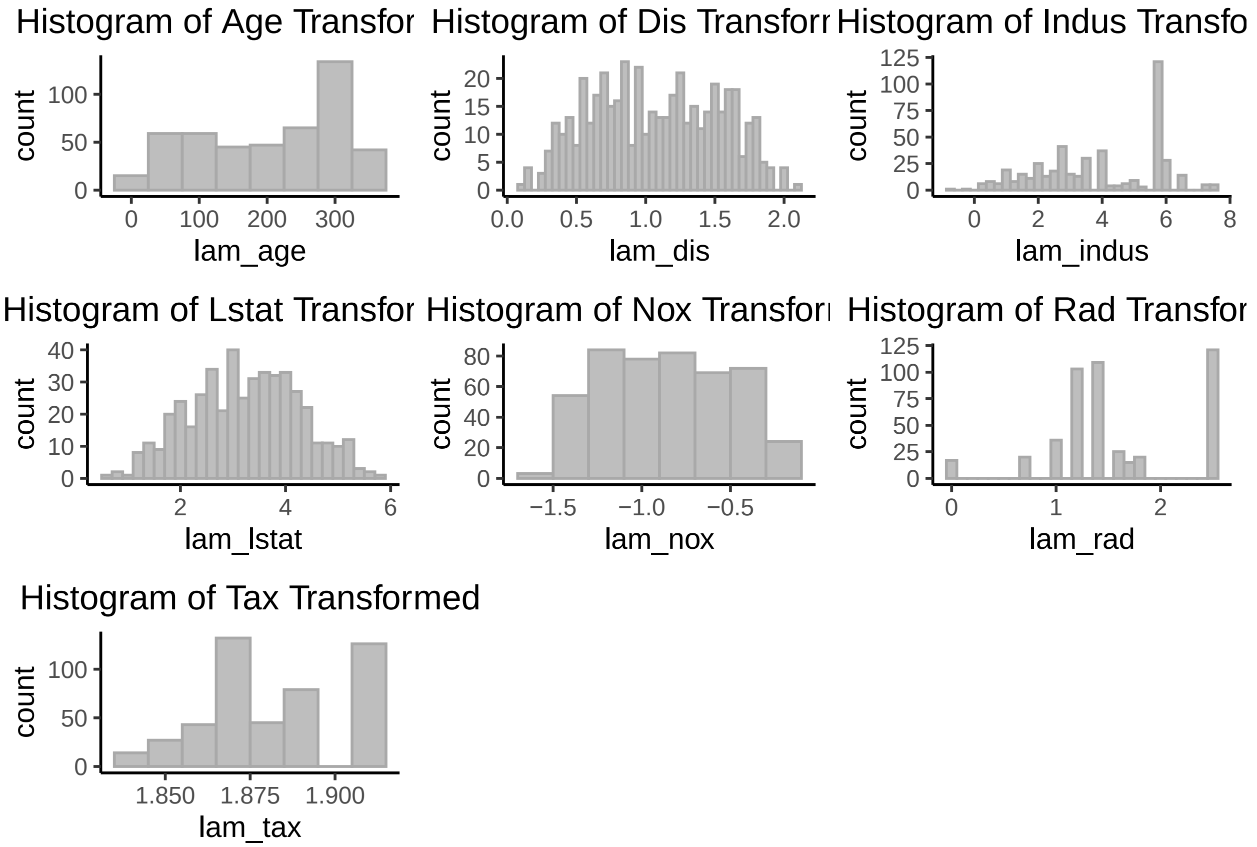
After the data exploration phase was completed, we transitioned into preparing the data for our regression models.

Our data looks relatively clean and it doesn't seem there is much need to do any modifications. However, we do have predictor variables that are skewed or do not follow a normal distribution. We can make use of both the log and the Box cox function in order to figure out what is the best transformation that can be applied to these variables in order to normalize them.

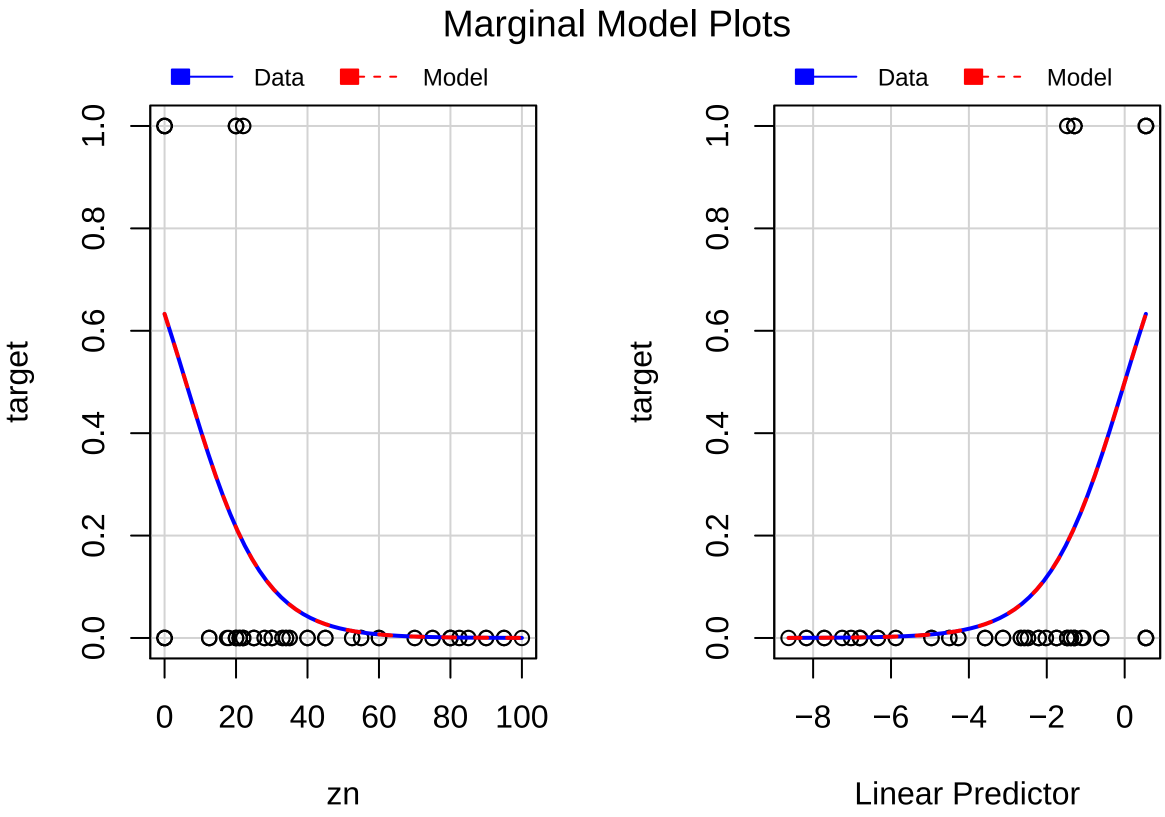
A few of the variables that seem skewed or don't follow a normal distribution include: "age", "dis", "indus", "lstat", "nox", "ptratio", and "rad".

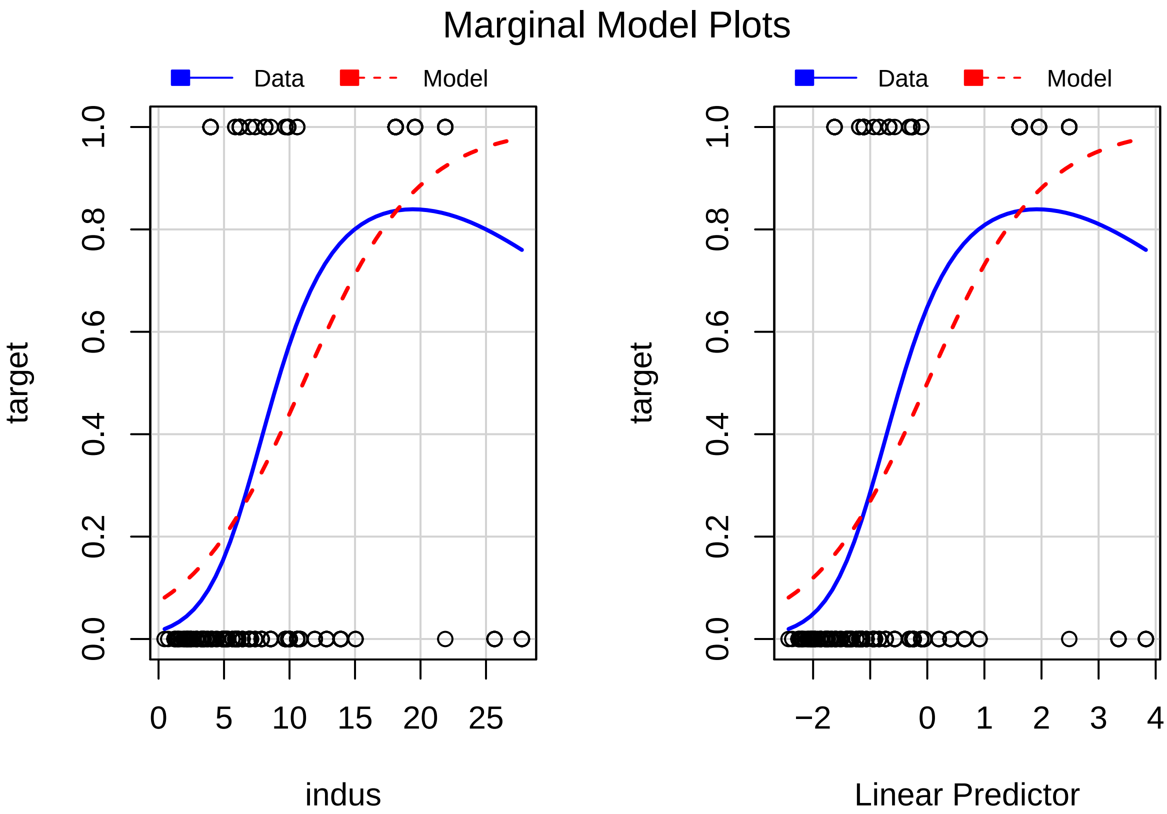
We will use the "boxcoxfit" function from the "geoR" and " forecast" package to extract the fitted parameters and use the value of lambda for the transformations of each of the variables mentioned above. We will then use

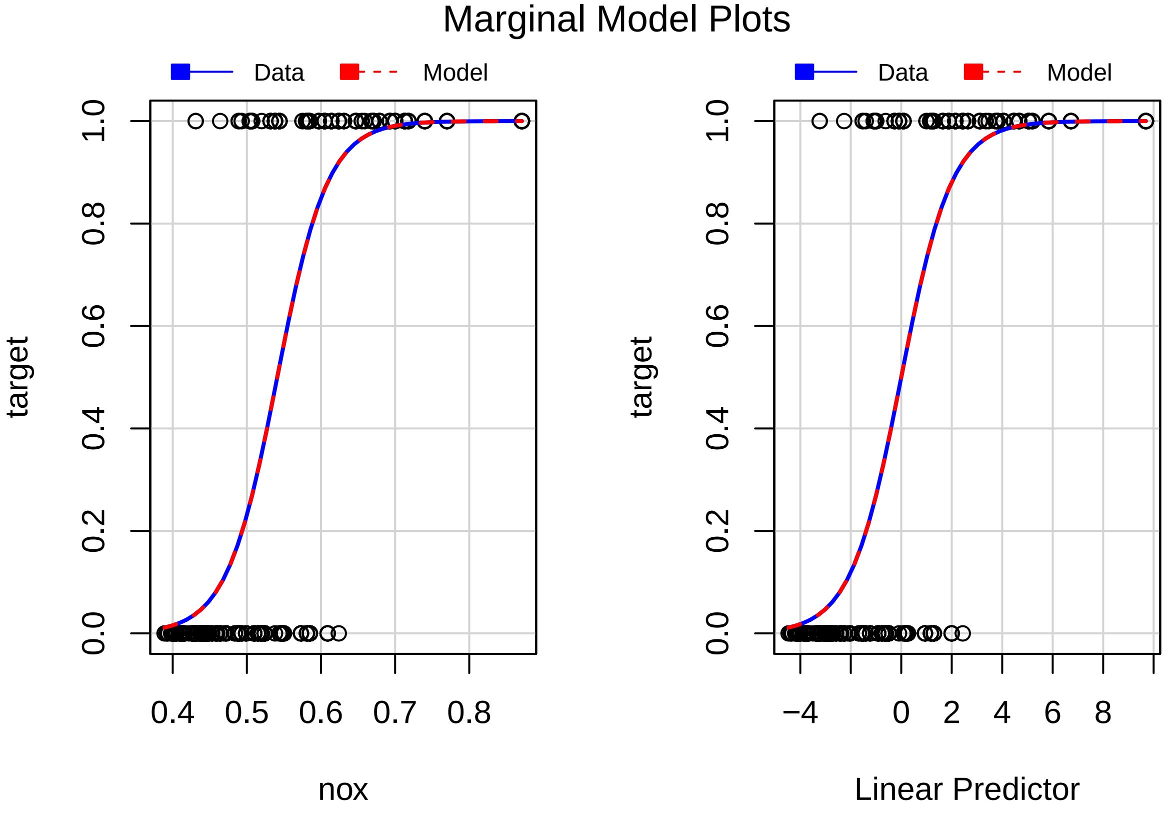
Below, we can observe that even after transforming the variables, some of them do not follow a nearly normal distribution still, but we were at least able to bring them closer to normalization

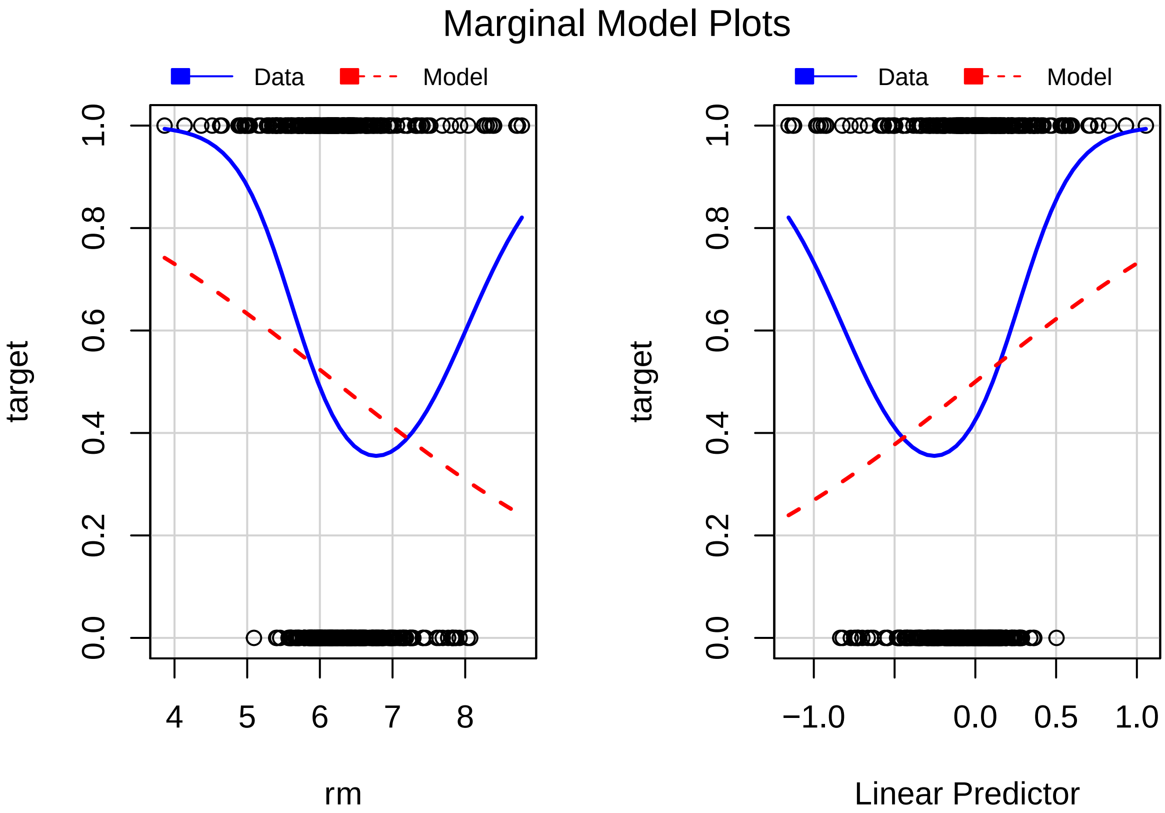


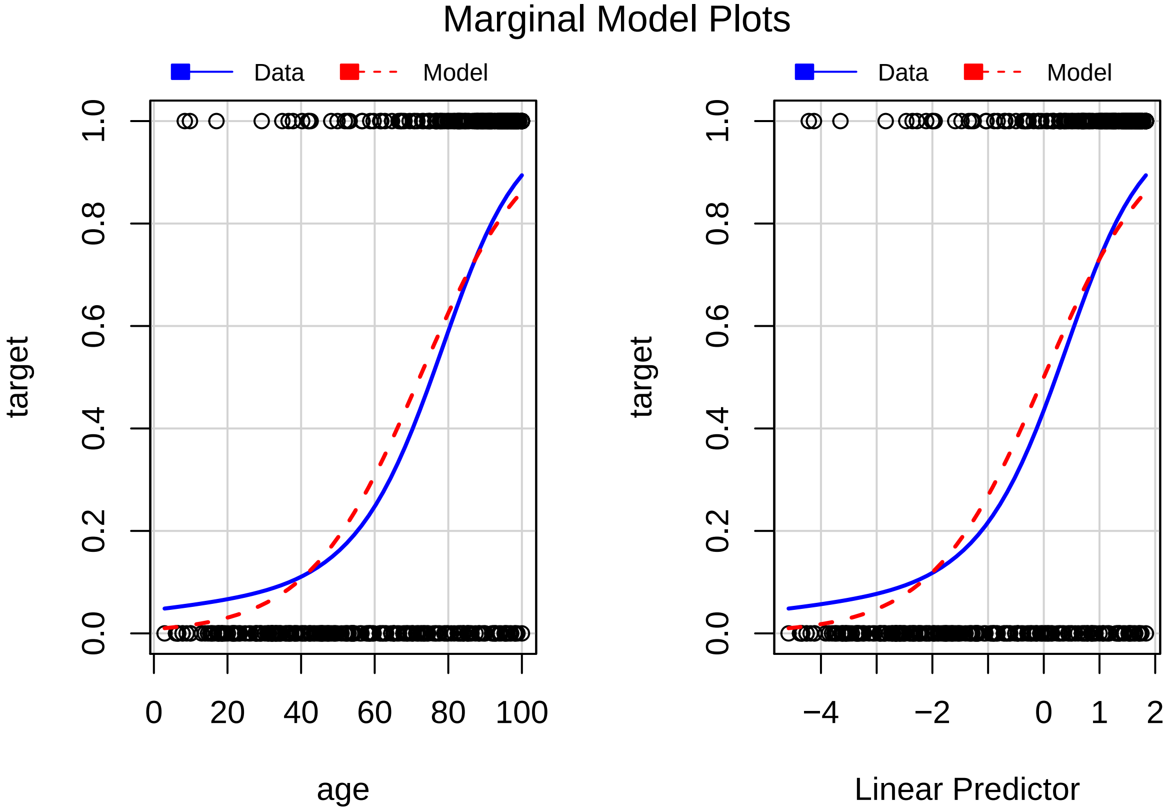
We can also observe the marginal model plots below, which show us how well our data fits with our logistic model. We can see that, some of our variables in our data do not fit well with the model, such as “indus”, “rm”, and “ptrartio”. We will try using log and Box Croft to see if we can make the badly performing predictors fit the model better.

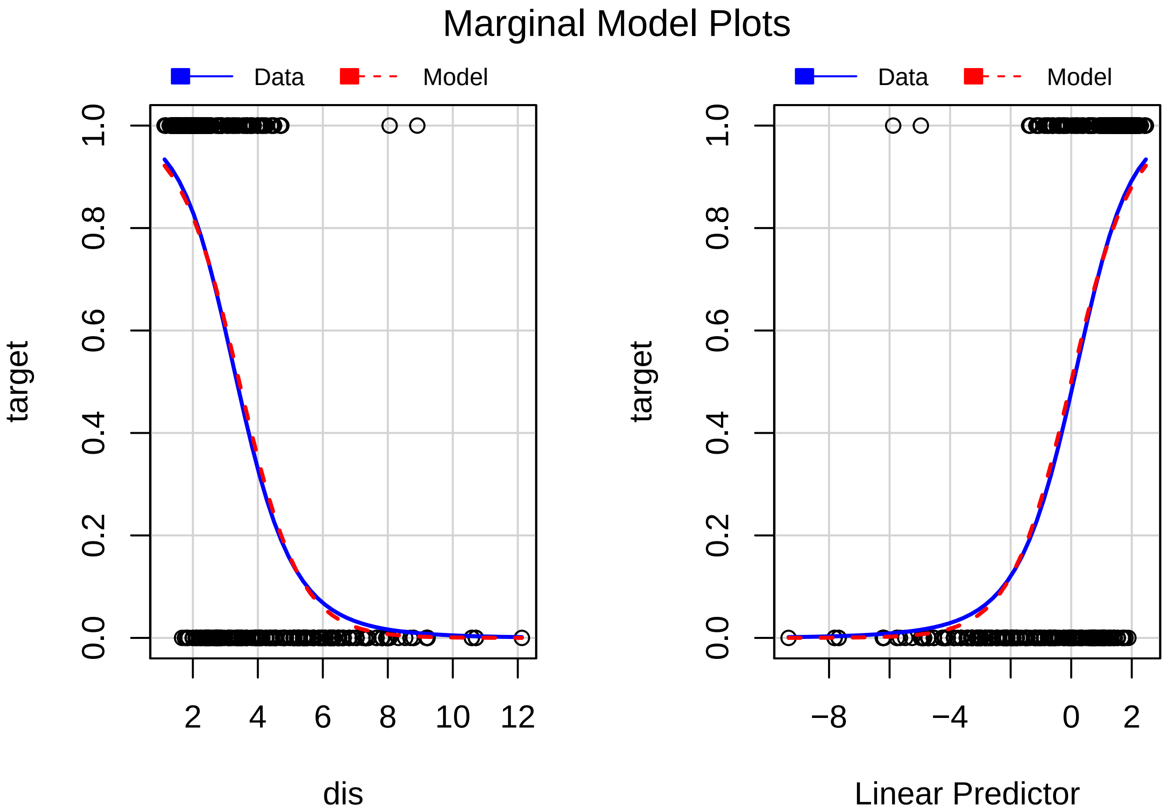


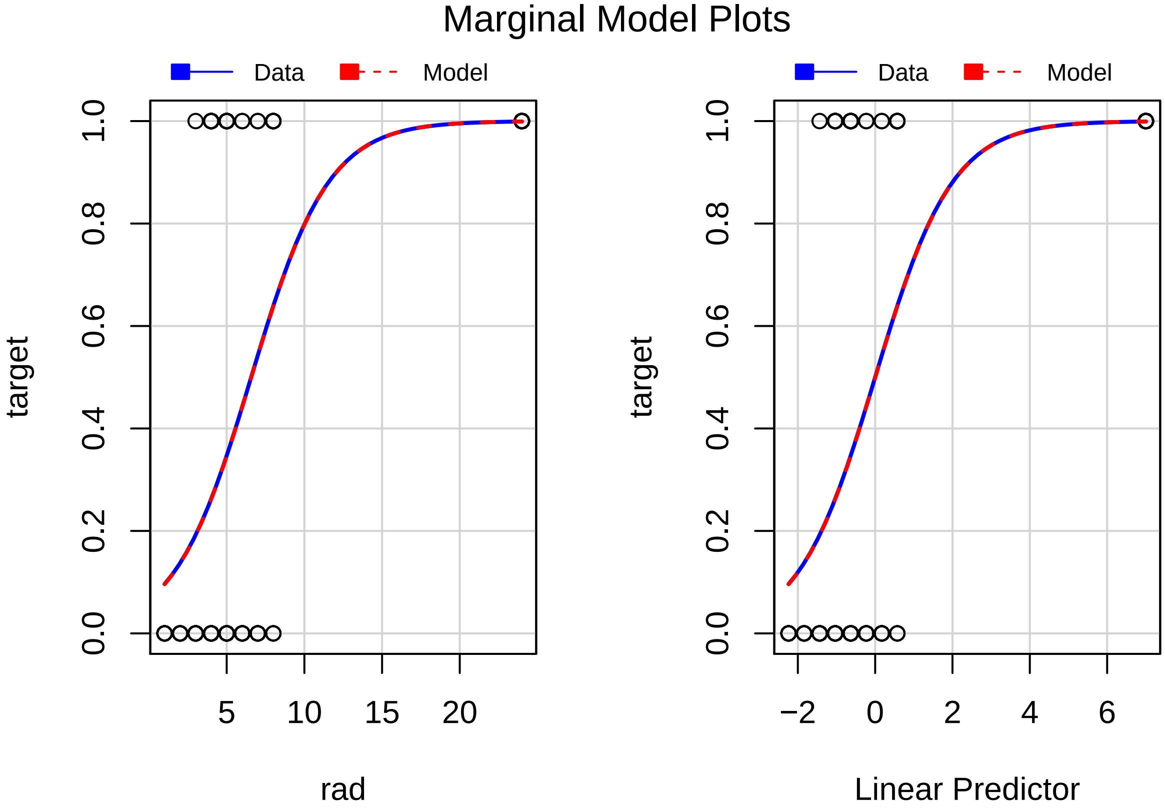

Marginal Model Plots
Data
Model
target
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
indus
−2
0
1
2
3
4
Linear Predictor

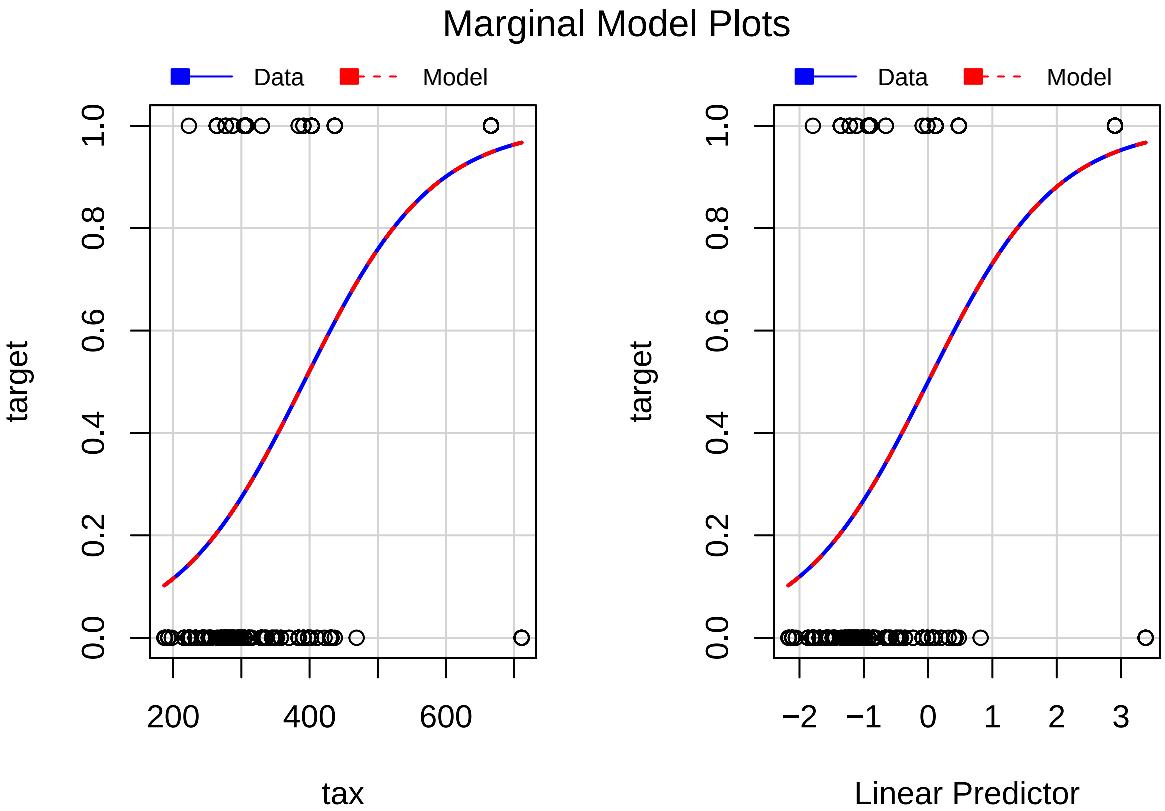

Marginal Model Plots
Data
Model
target
0.0
0.2
0.4
0.6
0.8
1.0
0.4
0.5
0.6
0.7
0.8
nox
−4
0
2
4
6
8
Linear Predictor

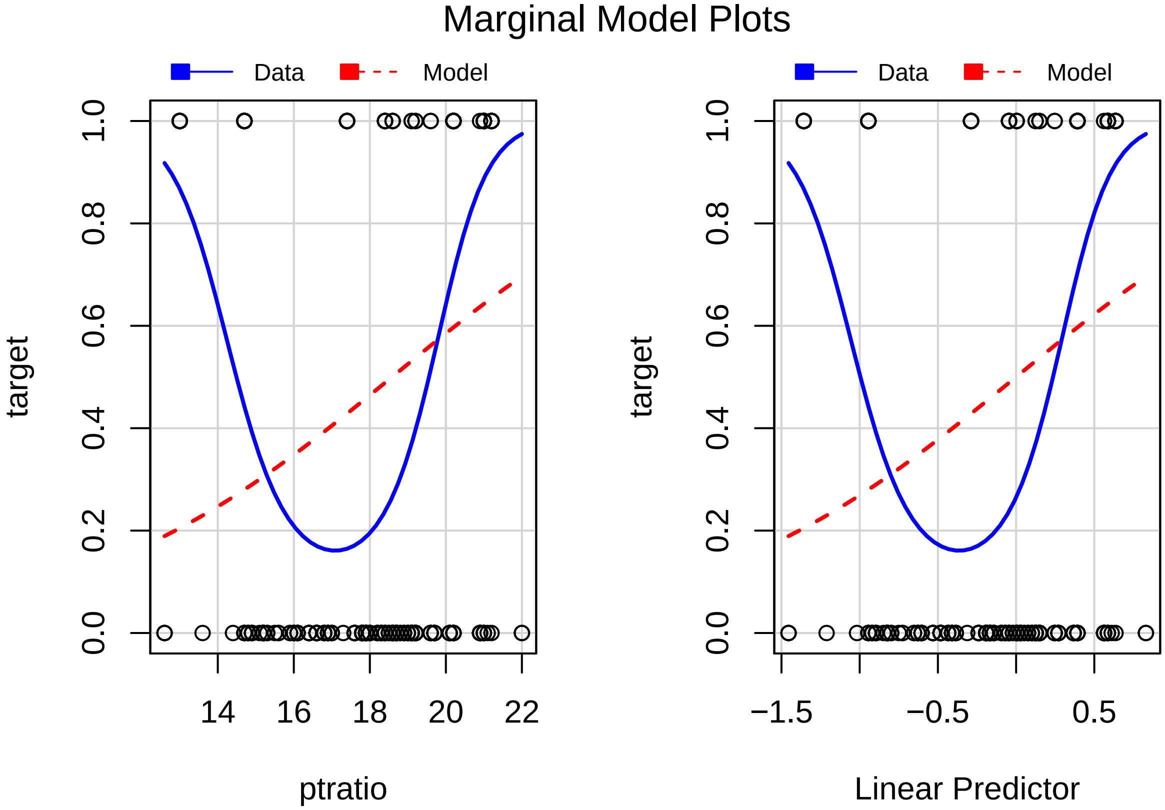

Marginal Model Plots
Data
Model
Data
Model
target
target
1.0
0.8
0.6
0.4
0.2
0.0
1.0
0.8
0.6
0.4
0.2
0.0
4
5
6
7
8
−1.0
0.0
0.5
1.0
rm
Linear Predictor

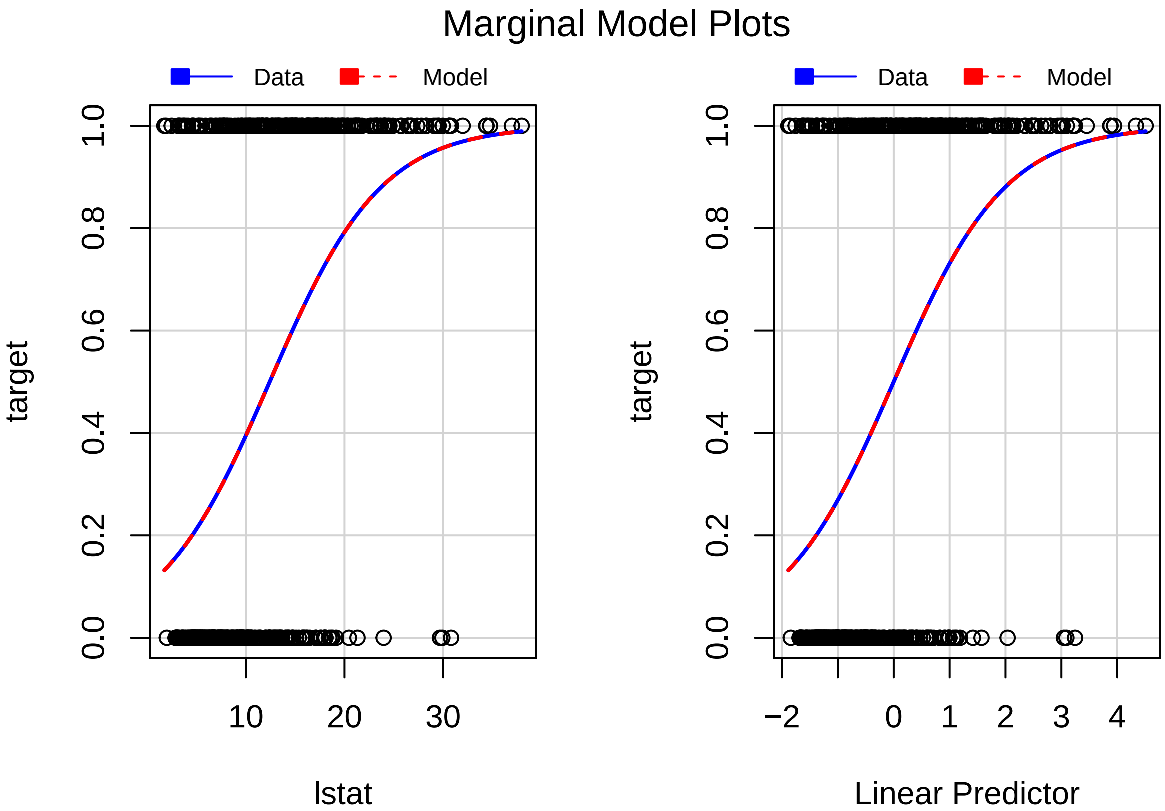

Marginal Model Plots
Data
Model
Data
Model
target
target
1.0
0.8
0.6
0.4
0.2
0.0
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
100
−4
−2
0
1
2
age
Linear Predictor

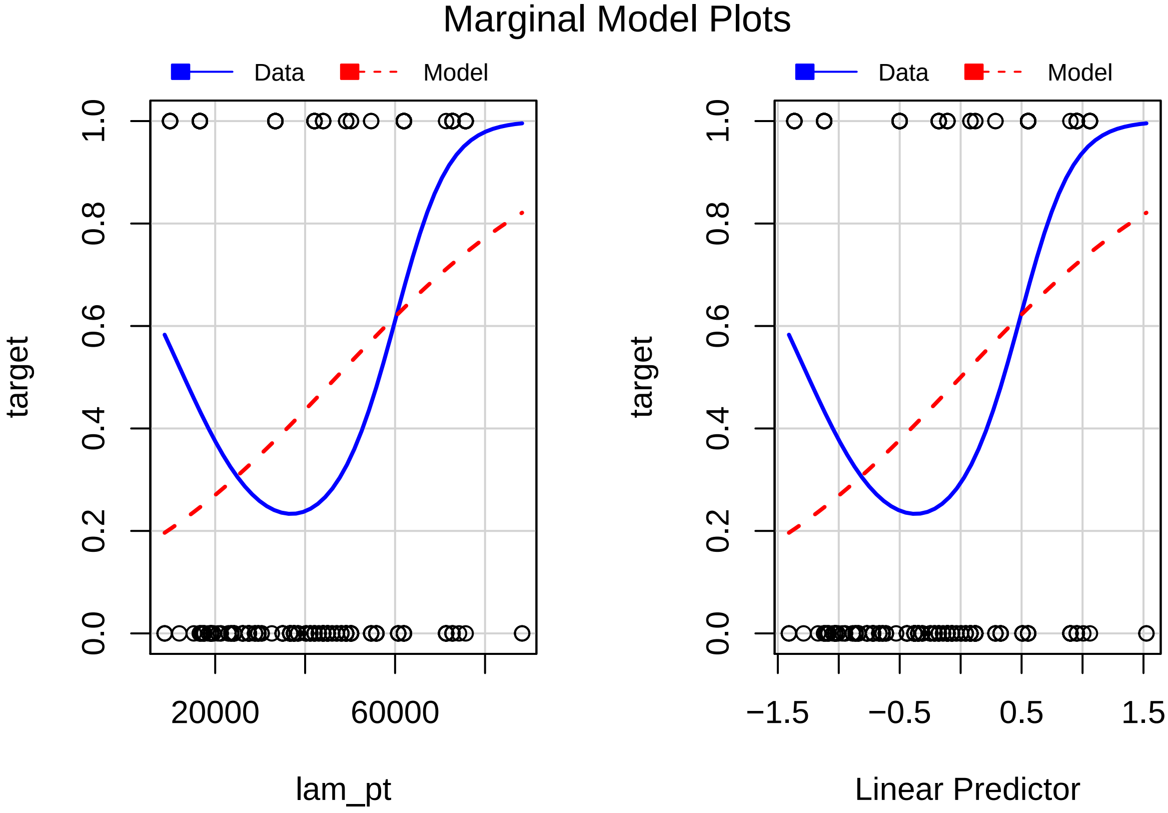

Marginal Model Plots
Data
Model
target
0.0
0.2
0.4
0.6
0.8
1.0
2
4
6
8
10
12
dis
target
0.0
0.2
0.4
0.6
0.8
1.0
−8
−4
0
2
Linear Predictor

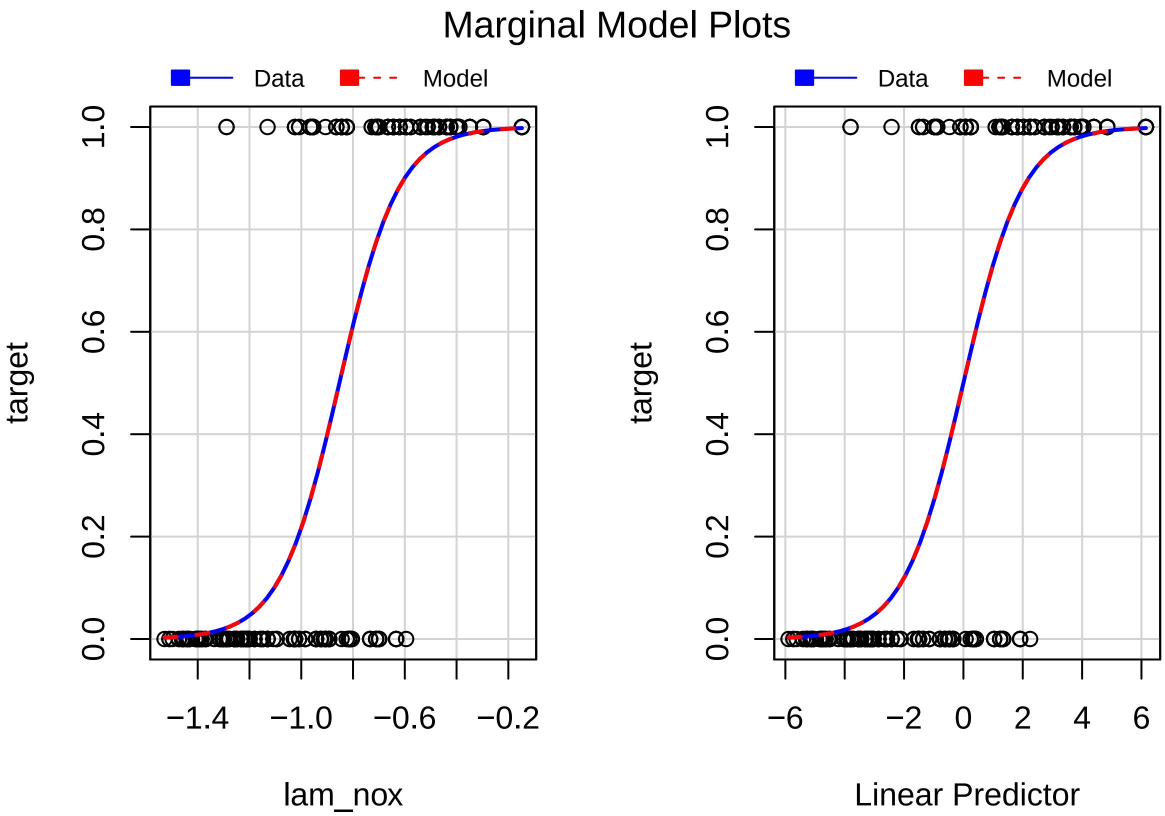

Marginal Model Plots
Data
Model
target
0.0
0.2
0.4
0.6
0.8
1.0
5
10
15
20
rad
target
0.0
0.2
0.4
0.6
0.8
1.0
−2
0
2
4
6
Linear Predictor

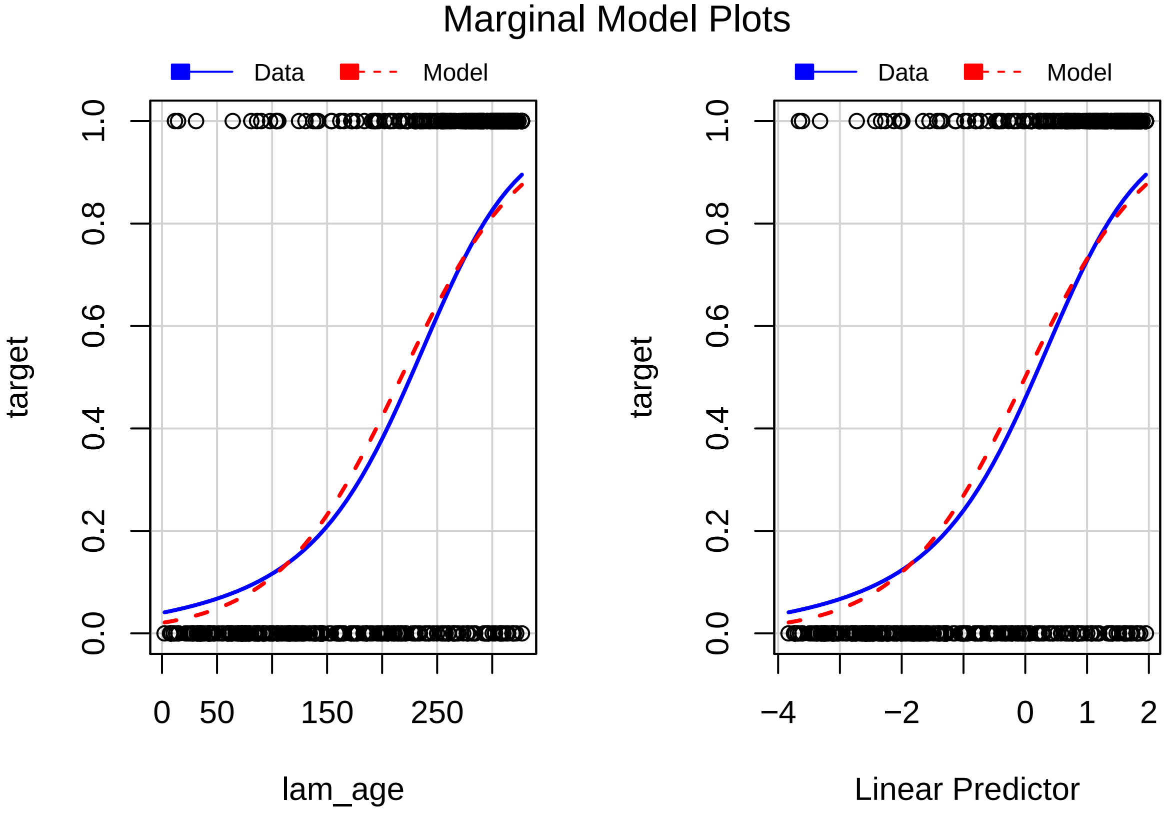

Marginal Model Plots
Data
Model
target
1.0
0.8
0.6
0.4
0.2
0.0
200
400
600
tax
Data
Model
target
1.0
0.8
0.6
0.4
0.2
0.0
−2
−1
0
1
2
3
Linear Predictor

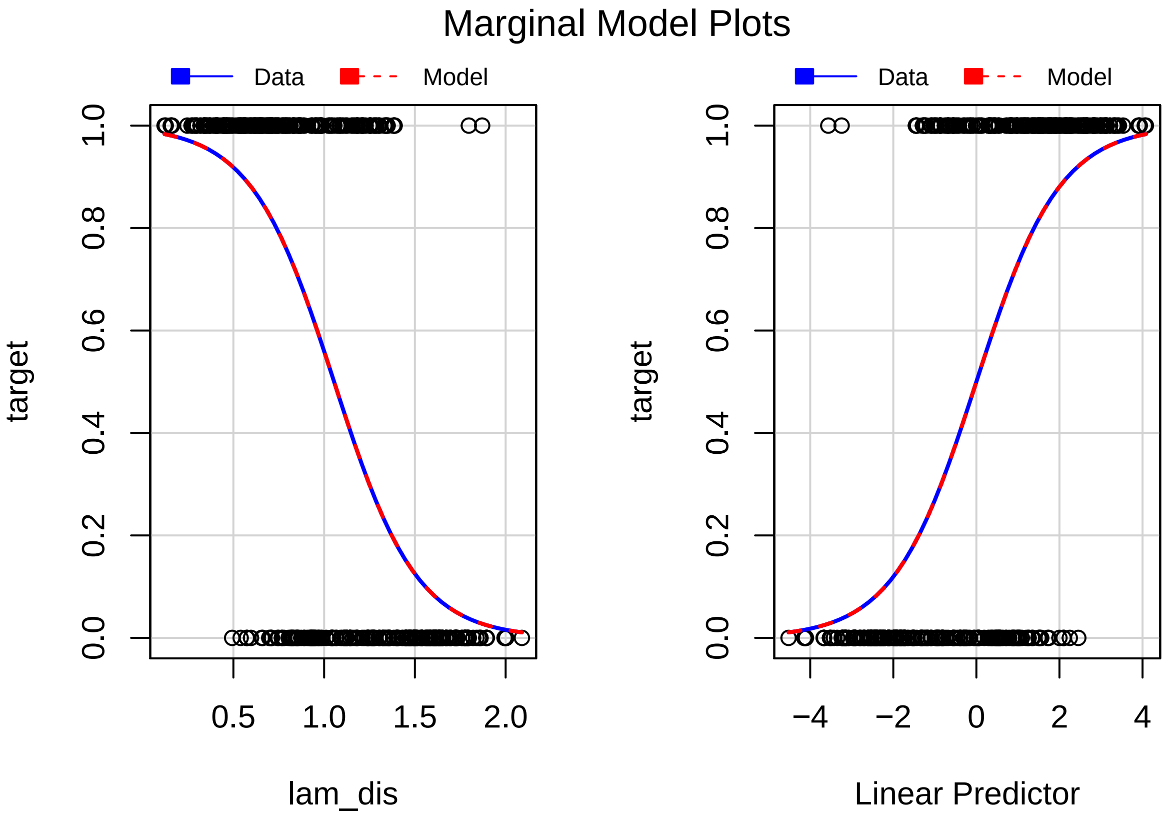

Marginal Model Plots
Data
Model
target
1.0
0.8
0.6
0.4
0.2
0.0
14
16
18
20
22
ptratio
Data
Model
target
1.0
0.8
0.6
0.4
0.2
0.0
−1.5
−0.5
0.5
Linear Predictor

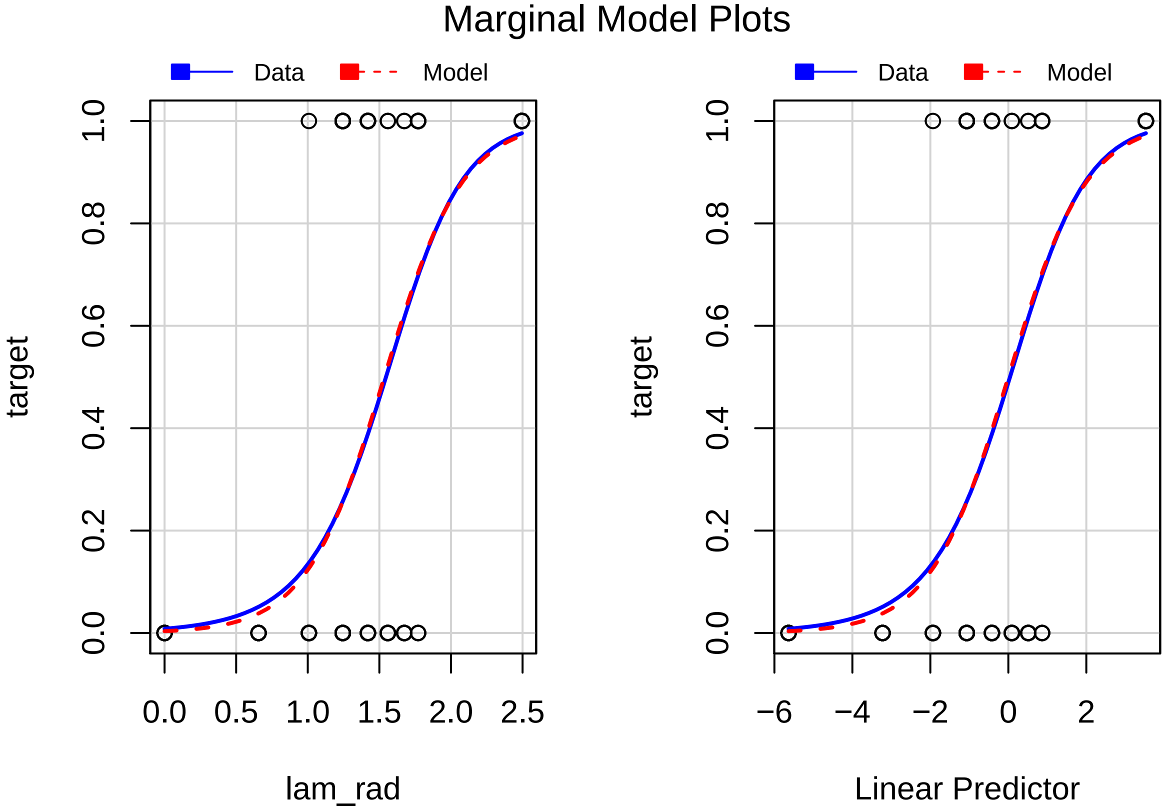

Marginal Model Plots
Data
Model
target
1.0
0.8
0.6
0.4
0.2
0.0
10
20
30
lstat
Data
Model
target
1.0
0.8
0.6
0.4
0.2
0.0
−2
0
1
2
3
4
Linear Predictor

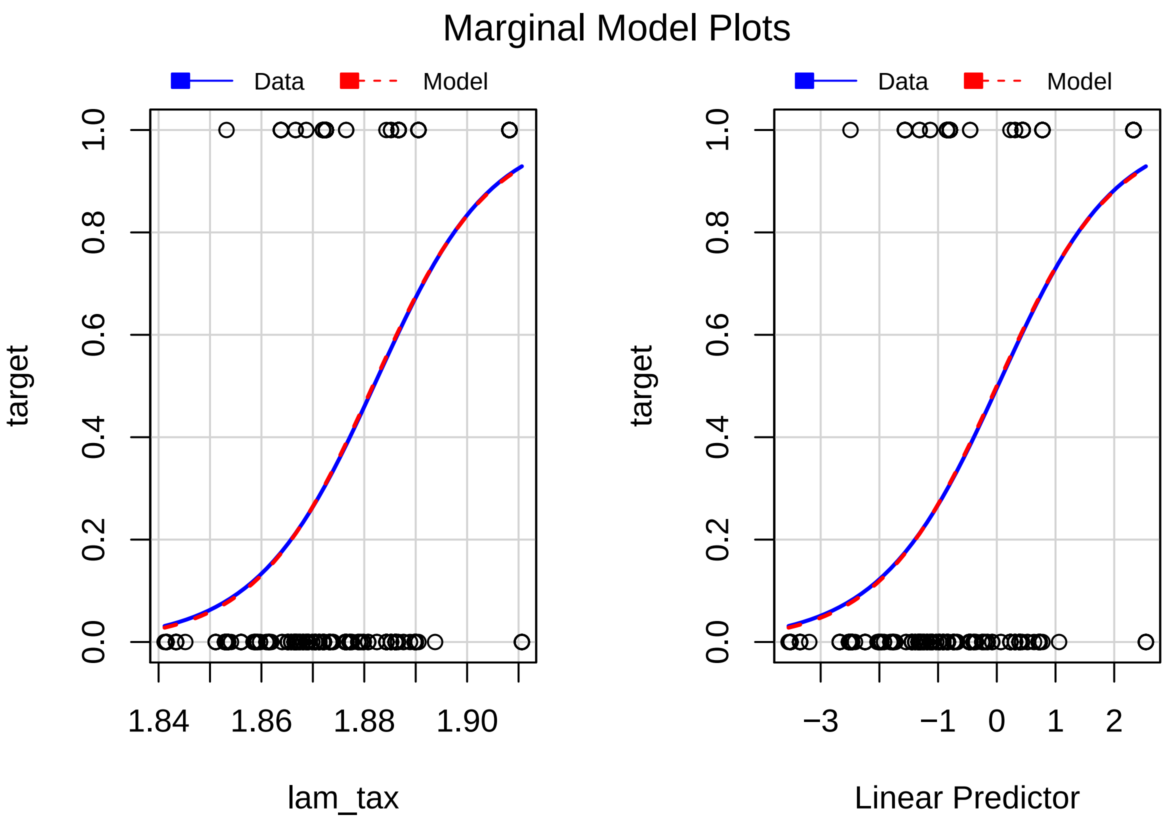

Marginal Model Plots
Data
Model
target
1.0
0.8
0.6
0.4
0.2
0.0
20000
60000
lam_pt
Data
Model
target
1.0
0.8
0.6
0.4
0.2
0.0
−1.5
−0.5
0.5
1.5
Linear Predictor

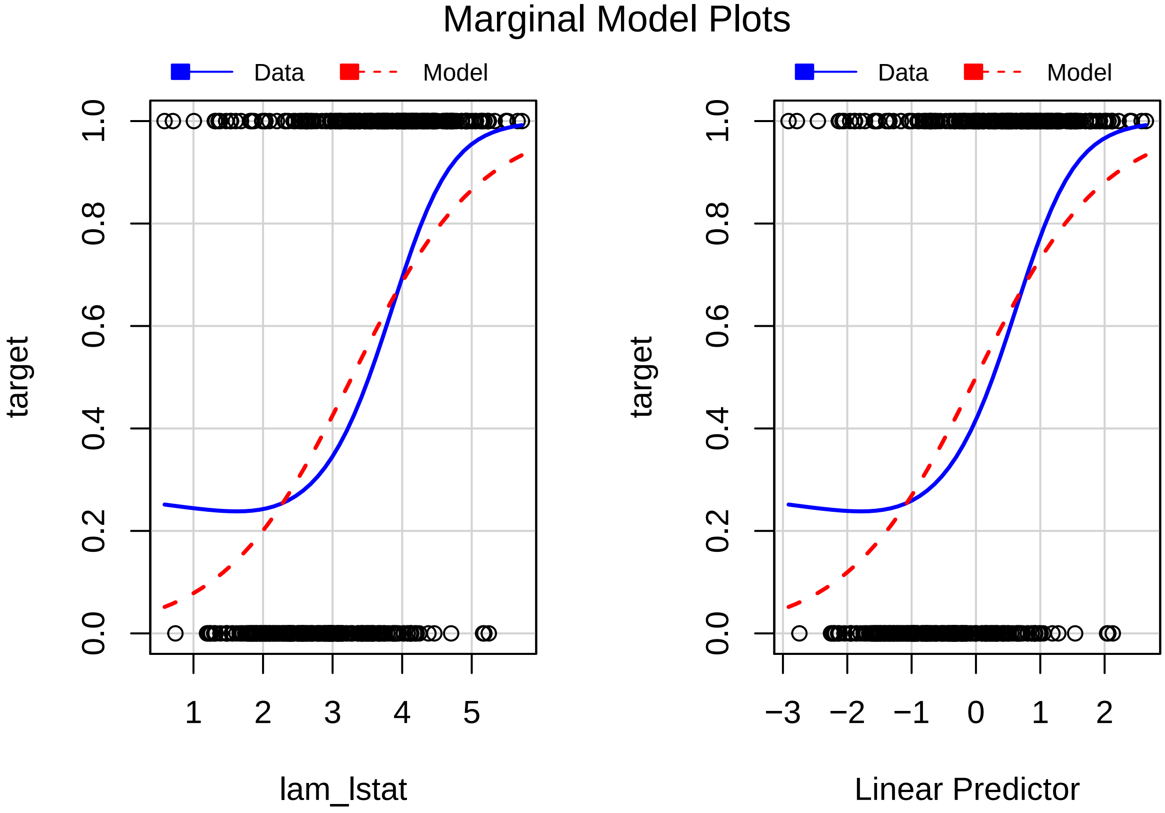

Marginal Model Plots
Data
Model
target
1.0
0.8
0.6
0.4
0.2
0.0
−1.4
−1.0
−0.6
−0.2
lam_nox
Data
Model
target
1.0
0.8
0.6
0.4
0.2
0.0
−6
−2
0
2
4
6
Linear Predictor

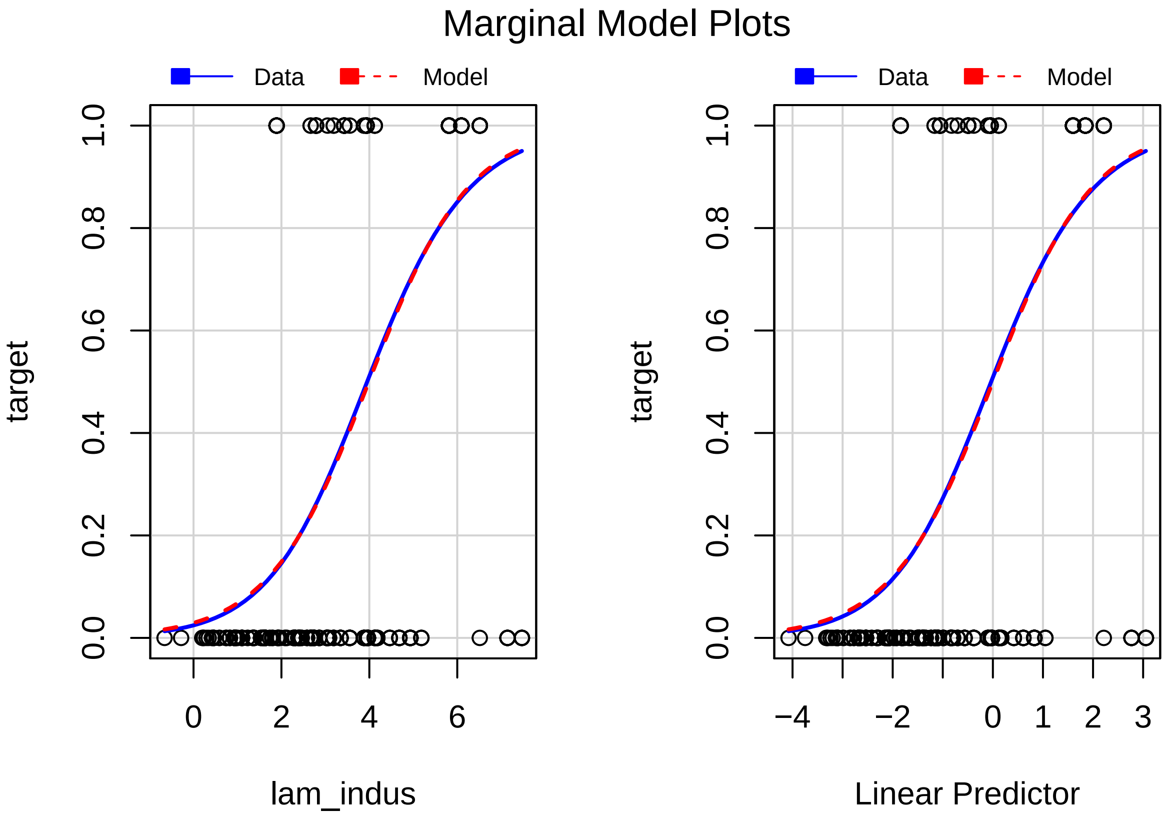

Marginal Model Plots
Data
Model
target
1.0
0.8
0.6
0.4
0.2
0.0
0
50
150
250
lam_age
Data
Model
target
1.0
0.8
0.6
0.4
0.2
0.0
−4
−2
0
1
2
Linear Predictor


Marginal Model Plots
Data
Model
Data
Model
target
0.0
0.2
0.4
0.6
0.8
1.0
0.5
1.0
1.5
2.0
lam_dis
target
0.0
0.2
0.4
0.6
0.8
1.0
−4
−2
0
2
4
Linear Predictor

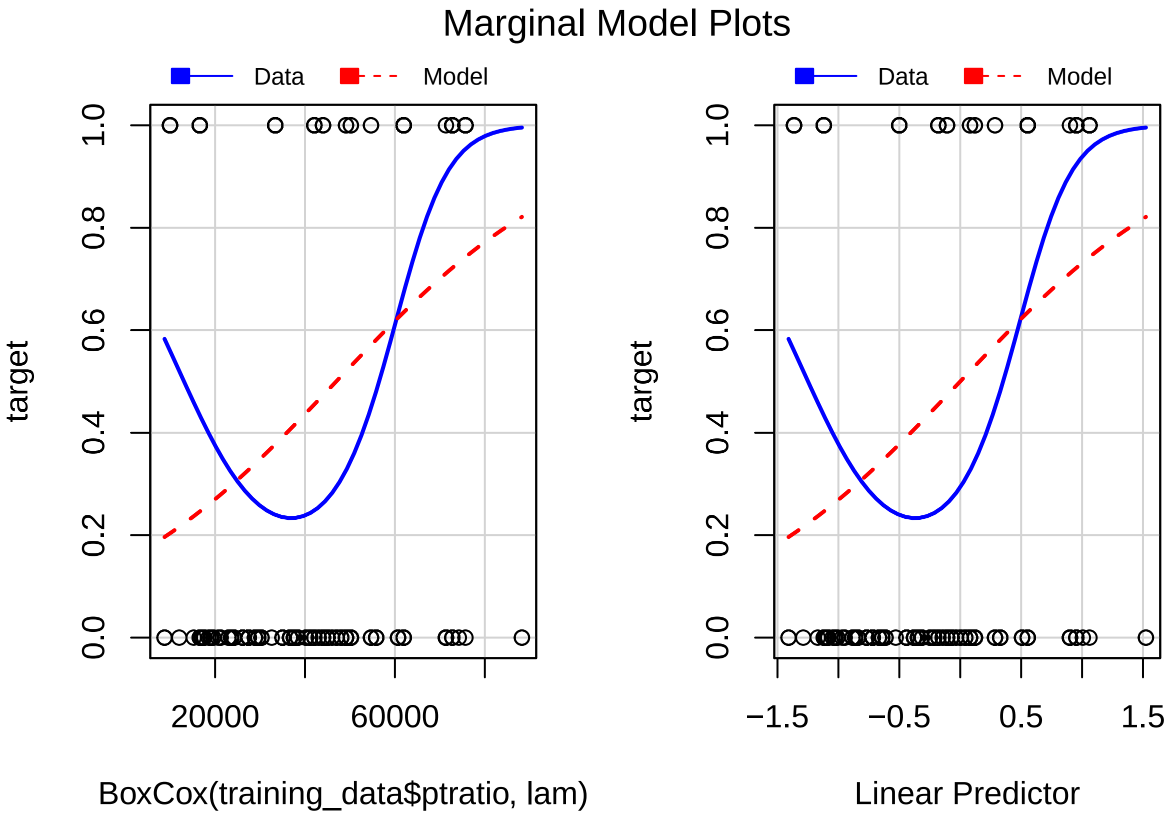

Marginal Model Plots
Data
Model
Data
Model
target
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
2.0
2.5
lam_rad
target
0.0
0.2
0.4
0.6
0.8
1.0
−6
−4
−2
0
2
Linear Predictor

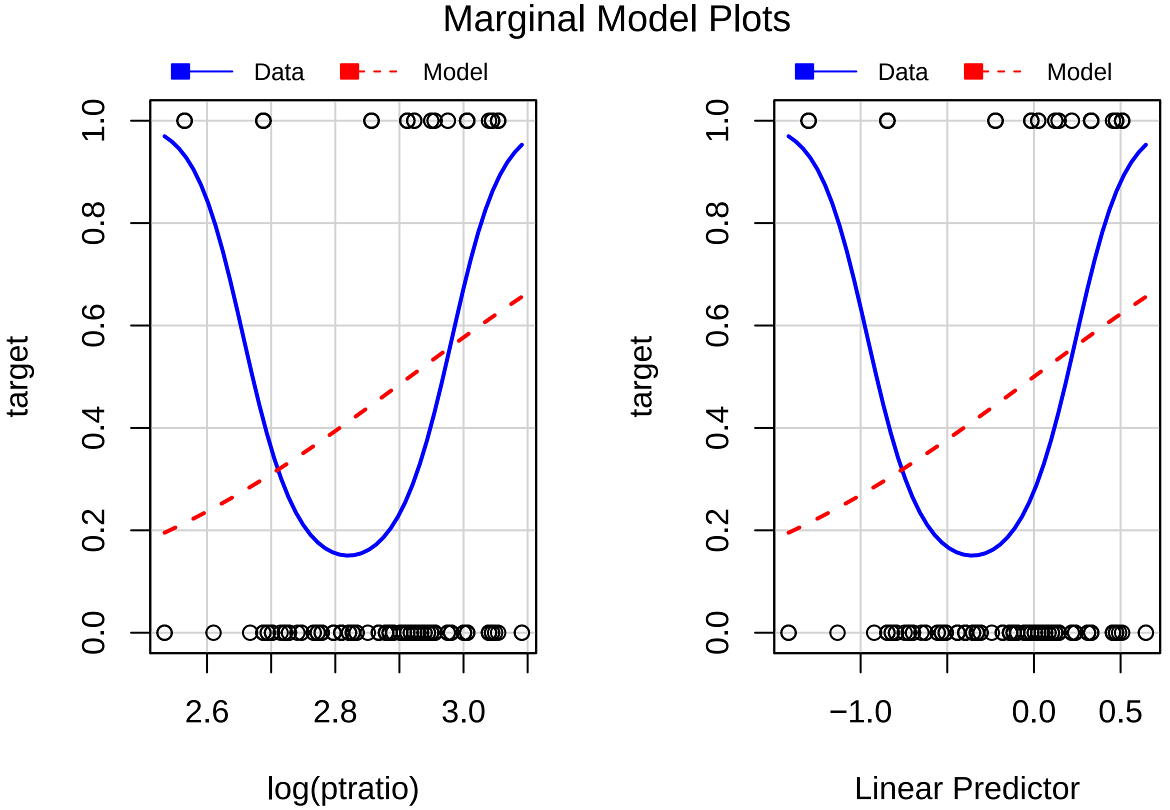

Marginal Model Plots
Data
Model
target
0.0
0.2
0.4
0.6
0.8
1.0
1.84
1.86
1.88
1.90
lam_tax
Data
Model
target
0.0
0.2
0.4
0.6
0.8
1.0
−3
−1
0
1
2
Linear Predictor


Marginal Model Plots
Data
Model
target
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5
lam_lstat
Data
Model
target
0.0
0.2
0.4
0.6
0.8
1.0
−3
−2
−1
0
1
2
Linear Predictor

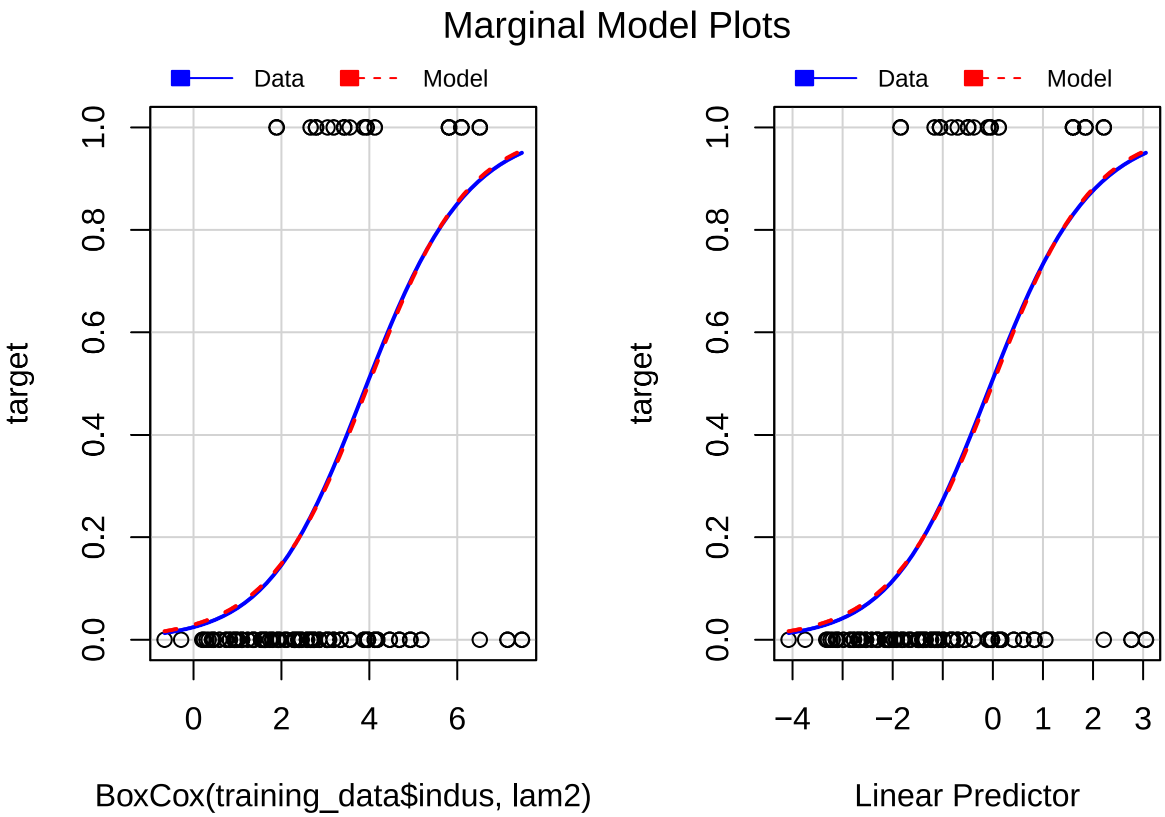

Marginal Model Plots
Data
Model
target
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
lam_indus
Data
Model
target
1.0
0.8
0.6
0.4
0.2
0.0
−4
−2
0
1
2
3
Linear Predictor

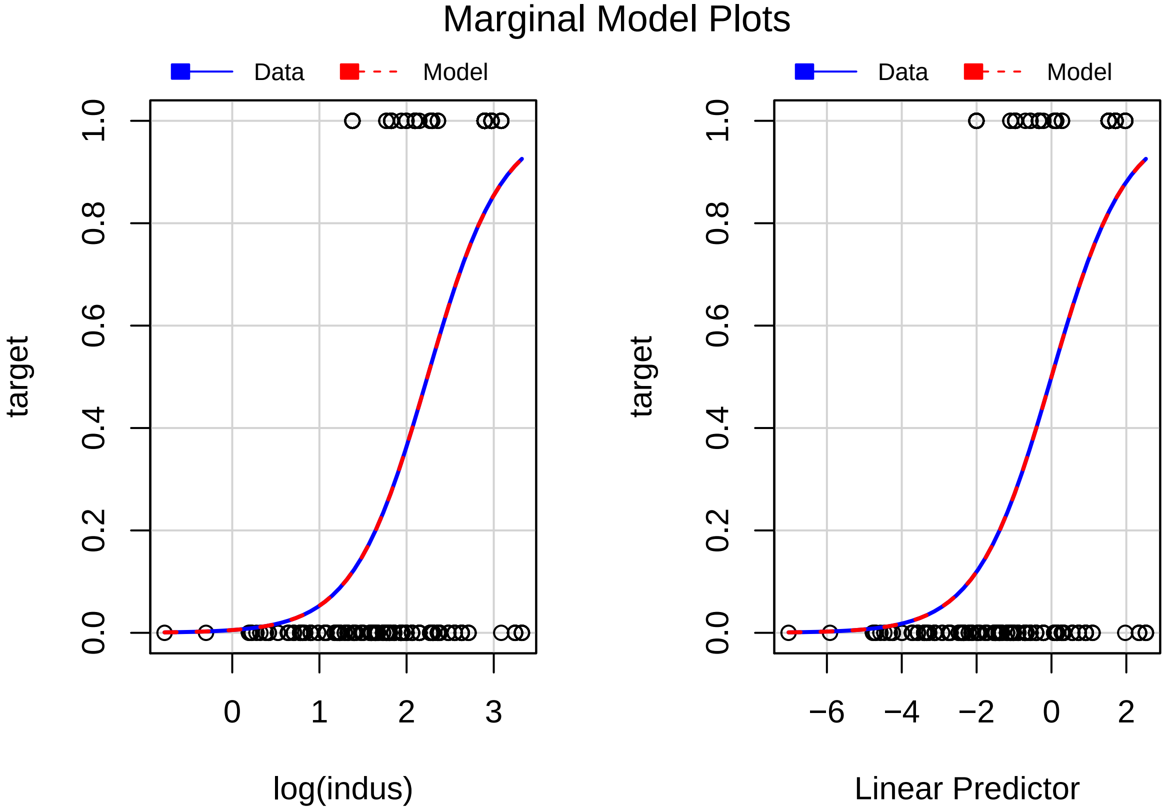
For “ptratio”, we can see below that the log transformation does nothing, while the boxcoxfit only marginally improves the fit.





For “indus”, both Box Croft and log improves the fit





Building Models

Following the data preparation phase, we brainstormed how best to construct an appropriate model design process. Given that the dataset we are working with is fairly small, we used a K-Fold Cross Validation technique to train the models. Additionally, we split our data into an additional training and test set in order to use 80% of it in the models and then evaluate their performance with the predictions against the remaining 20%.

Using the partition, we constructed a saturated regression model which contained all variables of the dataset. This gave us a starting point to analyze the statistical significance of each variable and their associated correlations to the dependent variable.

Our first model includes all predictor variables including the transformations we created earlier.

```
Call:
NULL

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.50018  -0.05738   0.00000   0.03745   3.12663

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.3659     1.4164   0.258 0.796158
zn           -2.4880     1.8454  -1.348 0.177582
indus         0.9335     4.8221   0.194 0.846504
nox          17.7576     8.9506   1.984 0.047261 *
rm            0.3254     0.5698   0.571 0.567986
age          -4.7912     5.3873  -0.889 0.373820
dis          -3.8139     2.5548  -1.493 0.135483
rad          15.7083     4.5995   3.415 0.000637 ***
tax         -33.2607    10.0865  -3.298 0.000975 ***
ptratio      -4.0435     2.1905  -1.846 0.064897 .
lstat         4.4050     2.1342   2.064 0.039014 *
lam_pt        4.8982     2.0681   2.368 0.017864 *
lam_nox     -11.2107     7.4148  -1.512 0.130550
lam_age       5.4740     5.4231   1.009 0.312786
lam_dis       3.9859     2.7604   1.444 0.148759
lam_rad       0.4032     1.3009   0.310 0.756597
lam_tax      18.8870     6.4592   2.924 0.003455 **
lam_lstat    -4.3826     2.2100  -1.983 0.047363 *
lam_indus     1.3084     5.2186   0.251 0.802027
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 518.378  on 373  degrees of freedom
Residual deviance:  94.317  on 355  degrees of freedom
AIC: 132.32

Number of Fisher Scoring iterations: 10
```

For our second model, we have chosen the variables with high collinearity between the response and predictor variables and take the log on some variables:

```
Call:
NULL

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.96532  -0.25678  -0.01360   0.00402   2.73937

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    2.6332     0.7149   3.683 0.000230 ***
nox            3.4315     0.6872   4.993 5.94e-07 ***
rad            6.6791     1.3915   4.800 1.59e-06 ***
age            0.6609     0.3145   2.101 0.035609 *
tax           -2.0229     0.5753  -3.516 0.000438 ***
`log(indus)`   0.6406     0.4478   1.431 0.152546
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 518.38  on 373  degrees of freedom
Residual deviance: 162.64  on 368  degrees of freedom
AIC: 174.64

Number of Fisher Scoring iterations: 8
```

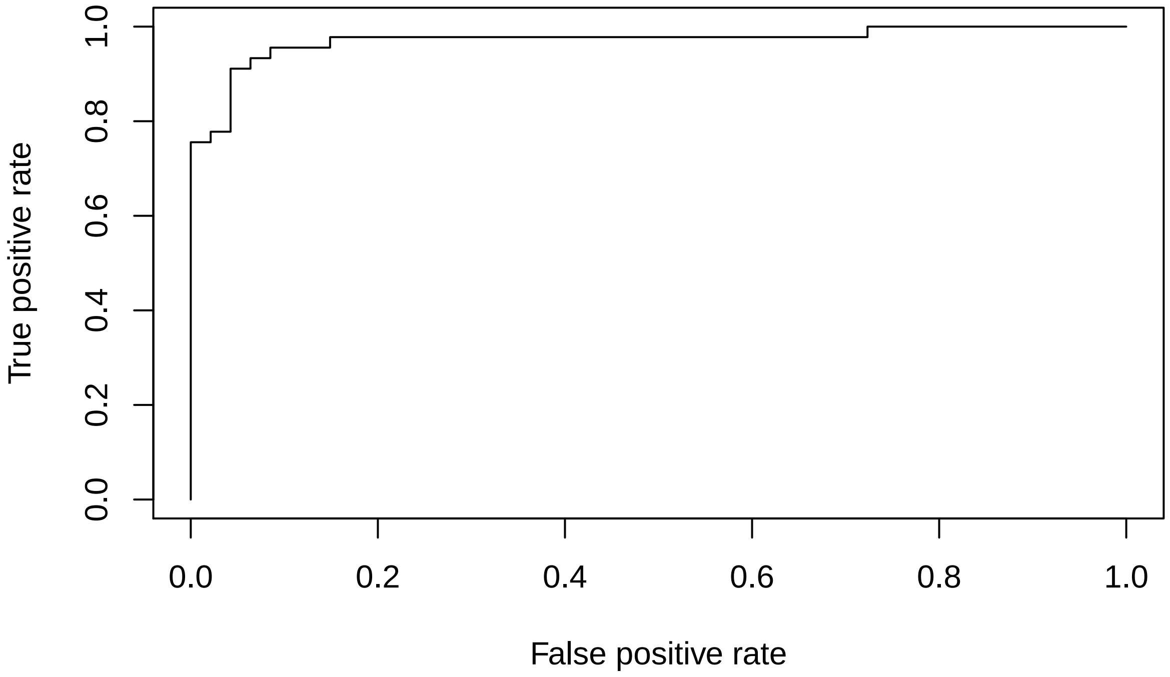
In our final mode, we mix and match transformations, and include some normal values as well. We determined the values below by process of elimination, removing the intercepts with low significant values while watching the AIC score change. The best combination was found below.

```
Call:
NULL

Deviance Residuals: 
   Min      1Q  Median      3Q     Max  
-2.511   0.000   0.000   0.000   2.548  

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)   
(Intercept)     5.267e+01  2.242e+01   2.349  0.01880 * 
indus           1.049e+01  3.472e+00   3.020  0.00253 **
tax            -1.989e+02  6.713e+01  -2.963  0.00305 **
ptratio        -2.138e+04  7.385e+03  -2.895  0.00379 **
lstat           6.437e+00  2.505e+00   2.570  0.01017 * 
lam_pt          2.046e+05  7.046e+04   2.904  0.00368 **
lam_tax         1.278e+02  4.313e+01   2.964  0.00304 **
lam_lstat      -6.144e+00  2.537e+00  -2.421  0.01546 * 
`I(zn^2)`      -9.030e+01  3.155e+01  -2.862  0.00421 **
`I(rad^2)`      2.241e+02  7.793e+01   2.876  0.00403 **
`I(ptratio^2)`  4.634e+04  1.599e+04   2.898  0.00375 **
`I(ptratio^4)` -2.296e+05  7.908e+04  -2.904  0.00368 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 518.378  on 373  degrees of freedom
Residual deviance:  53.215  on 362  degrees of freedom
AIC: 77.215

Number of Fisher Scoring iterations: 15
```

Selecting a Model

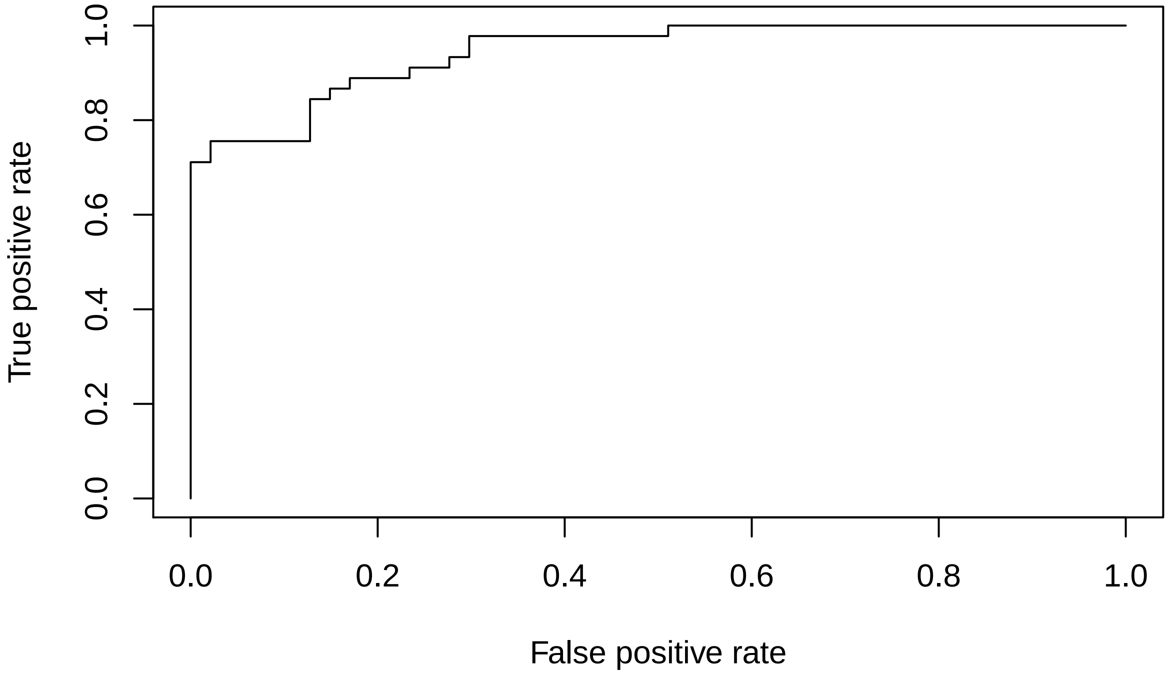
In selecting the best model, first we need to measure performance of the models prior to selection. We can do so by looking at the confusion matrix and AUC curve for our models. For the first model we have:

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 45  4
         1  2 41

               Accuracy : 0.9348
                 95% CI : (0.8634, 0.9757)
    No Information Rate : 0.5109
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8694

 Mcnemar's Test P-Value : 0.6831

            Sensitivity : 0.9574
            Specificity : 0.9111
         Pos Pred Value : 0.9184
         Neg Pred Value : 0.9535
             Prevalence : 0.5109
         Detection Rate : 0.4891
   Detection Prevalence : 0.5326
      Balanced Accuracy : 0.9343

       'Positive' Class : 0
```



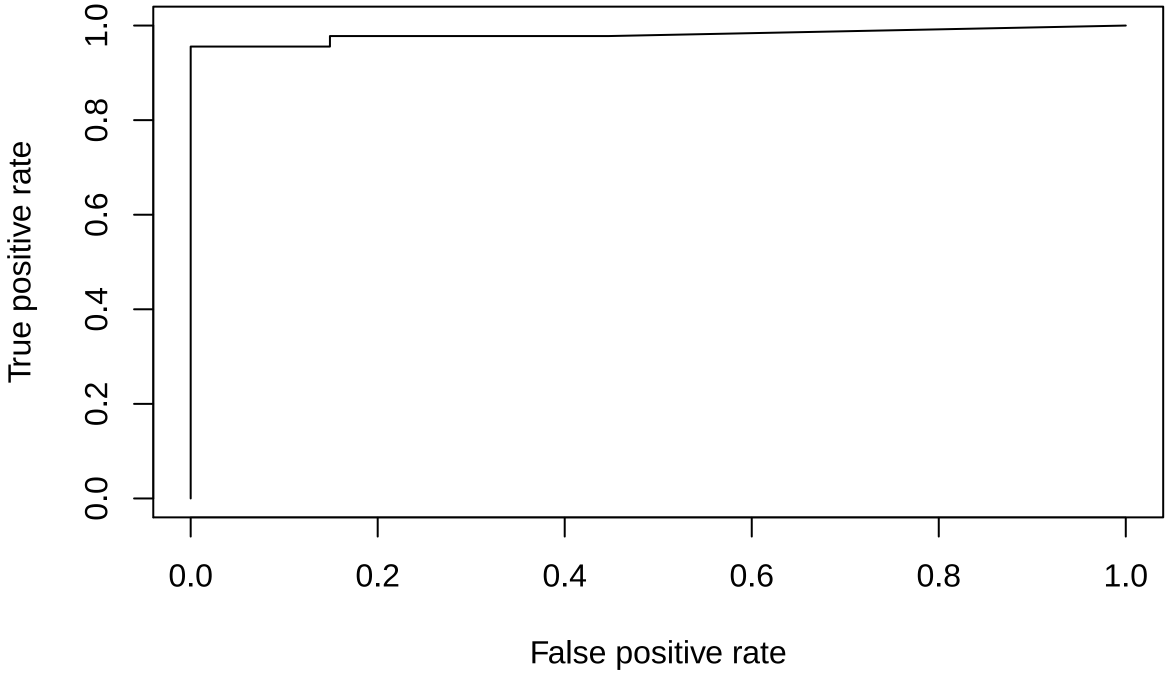
Model Two

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 42 11
         1  5 34

               Accuracy : 0.8261
                 95% CI : (0.733, 0.8972)
    No Information Rate : 0.5109
    P-Value [Acc > NIR] : 2.917e-10

                  Kappa : 0.651

 Mcnemar's Test P-Value : 0.2113

            Sensitivity : 0.8936
            Specificity : 0.7556
         Pos Pred Value : 0.7925
         Neg Pred Value : 0.8718
             Prevalence : 0.5109
         Detection Rate : 0.4565
   Detection Prevalence : 0.5761
      Balanced Accuracy : 0.8246

       'Positive' Class : 0
```



Model Three:

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 47  2
         1  0 43

               Accuracy : 0.9783
                 95% CI : (0.9237, 0.9974)
    No Information Rate : 0.5109
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9565

 Mcnemar's Test P-Value : 0.4795

            Sensitivity : 1.0000
            Specificity : 0.9556
         Pos Pred Value : 0.9592
         Neg Pred Value : 1.0000
             Prevalence : 0.5109
         Detection Rate : 0.5109
   Detection Prevalence : 0.5326
      Balanced Accuracy : 0.9778

       'Positive' Class : 0
```

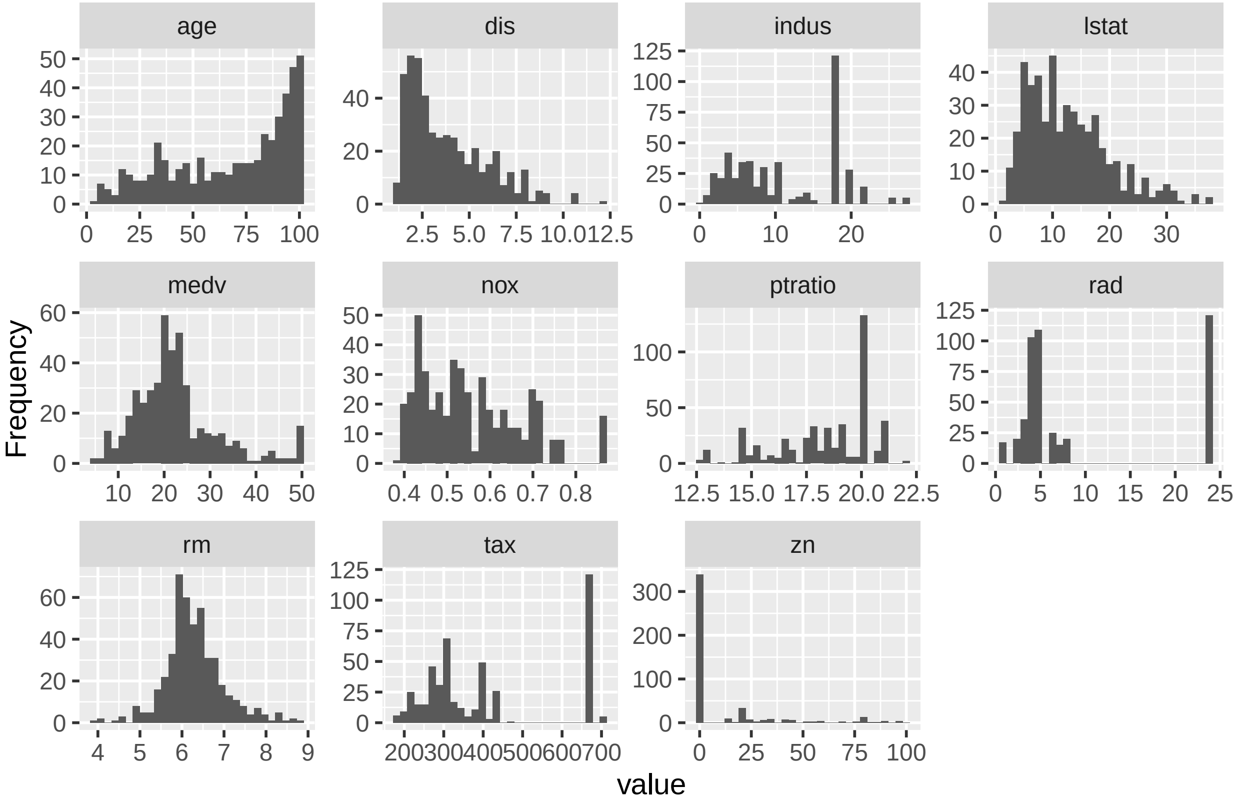


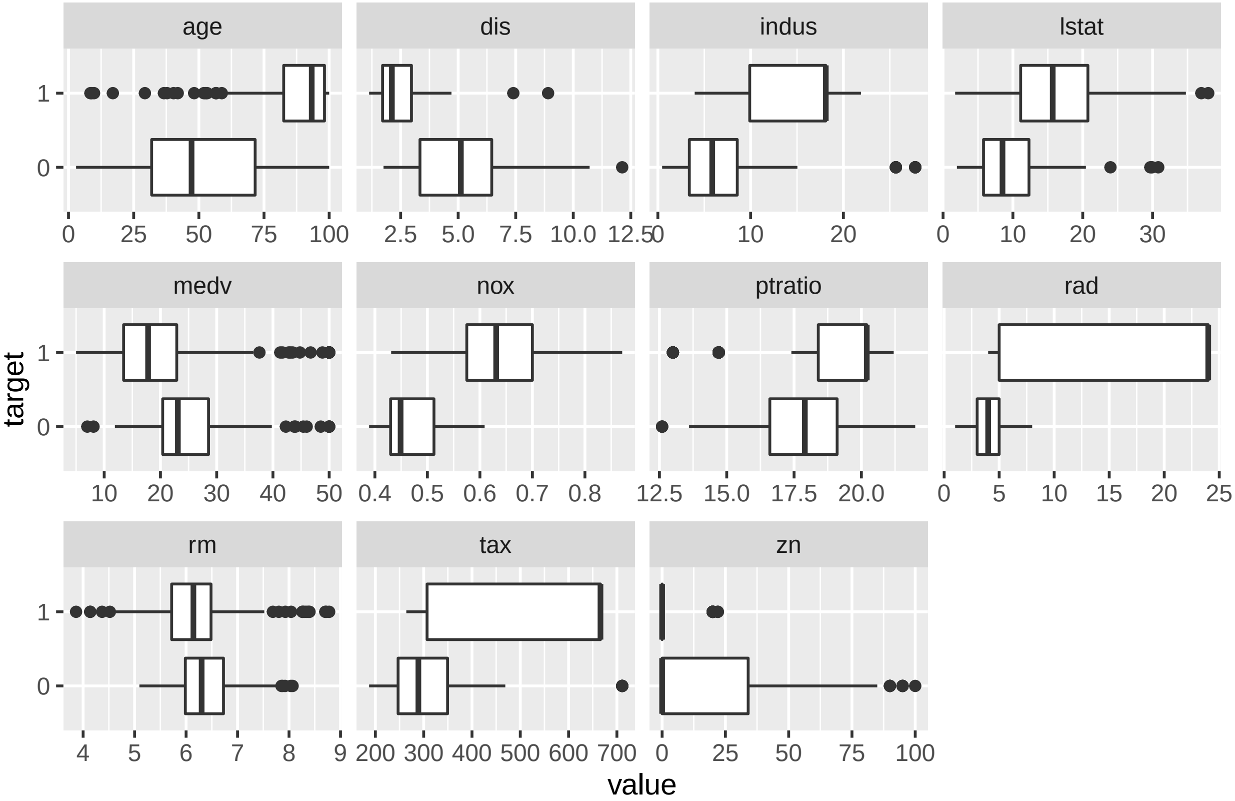
Evaluation

Analyzing the three model results, we determined that model three has the best predictive power and represents the strongest relationship to underlying data. The applied data transformations helped adjust for underlying skews and multicollinearity in the data. It also has near perfect AUC representing the strong predictive nature of the model.

We will now use that model on our evaluation data and create predictions with it. Shown below are the results of doing so. We note that our results closely resemble the distributions found in our training data.

	vars	n	mean	sd	min	max	range	se
zn	1	466	11.5772532	23.3646511	0.0000	100.0000	100.0000	1.0823466
indus	2	466	11.1050215	6.8458549	0.4600	27.7400	27.2800	0.3171281
chas*	3	466	1.0708155	0.2567920	1.0000	2.0000	1.0000	0.0118957
nox	4	466	0.5543105	0.1166667	0.3890	0.8710	0.4820	0.0054045
rm	5	466	6.2906738	0.7048513	3.8630	8.7800	4.9170	0.0326516
age	6	466	68.3675966	28.3213784	2.9000	100.0000	97.1000	1.3119625
dis	7	466	3.7956929	2.1069496	1.1296	12.1265	10.9969	0.0976026
rad	8	466	9.5300429	8.6859272	1.0000	24.0000	23.0000	0.4023678
tax	9	466	409.5021459	167.9000887	187.0000	711.0000	524.0000	7.7778214
ptratio	10	466	18.3984979	2.1968447	12.6000	22.0000	9.4000	0.1017669
lstat	11	466	12.6314592	7.1018907	1.7300	37.9700	36.2400	0.3289887
medv	12	466	22.5892704	9.2396814	5.0000	50.0000	45.0000	0.4280200
target*	13	466	1.4871245	0.5003714	1.0000	2.0000	1.0000	0.0231793





Conclusion

The underlying nature of this dataset was simple yet complex. In the way of modifications, there was not a great need to use dummy variables or transform the underlying data structure for this analysis. However, there was a large focus on transforming our variables to smooth out the distributions and reduce multicollinearity. After processing the data and transforming the necessary variables we were able to determine that our third model performed the best even if there may be some slight overfitting. It most accurately interpreted the multi-dimensional nature of the data and seemed best poised to deal with tails.