

# DATA 621 - Homework #3

Claudio, Mauricio

2022-04-11

## // Overview


In this homework assignment, you will explore, analyze and model a data set containing information on crime for various neighborhoods of a major city. Each record has a response variable indicating whether or not the crime rate is above the median crime rate (1) or not (0).










Your objective is to build a binary logistic regression model on the training data set to predict whether the neighborhood will be at risk for high crime levels. You will provide classifications and probabilities for the evaluation data set using your binary logistic regression model. The variables in the dataset are as follows:




- **zn**: proportion of residential land zoned for large lots (over 25000 square feet)
- **indus**: proportion of non-retail business acres per suburb
- **chas**: a factor for whether the suburb borders the Charles River (1) or not (0)
- **nox**: nitrogen oxides concentration (parts per 10 million)
- **rm**: average number of rooms per dwelling
- **age**: proportion of owner-occupied units built prior to 1940
- **dis**: weighted mean of distances to five Boston employment centers
- **rad**: index of accessibility to radial highways
- **tax**: full-value property-tax rate per \$10,000
- **ptratio**: pupil-teacher ratio by town
- **lstat**: lower status of the population (percent)
- **medv**: median value of owner-occupied homes in \$1000s
- **target**: whether the crime rate is above the median crime rate (1) or not (0) (response variable)

## // Data Exploration and Preparation

The dataset consists of 466 observations, with no missing or invalid values. We note the presence of skewed distributions and outliers in a handful of predictor variables. Positively, we note that the binary response variable, *target*, is quite balanced in its two classes.

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	zn [numeric]	Mean (sd) : 11.6 (23.4) min ≤ med ≤ max: 0 ≤ 0 ≤ 100 IQR (CV) : 16.2 (2)	26 distinct values		466 (100.0%)	0 (0.0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
2	indus [numeric]	Mean (sd) : 11.1 (6.8) min ≤ med ≤ max: 0.5 ≤ 9.7 ≤ 27.7 IQR (CV) : 13 (0.6)	73 distinct values		466 (100.0%)	0 (0.0%)
3	chas [factor]	1. 0 2. 1	433 ( 92.9% ) 33 ( 7.1% )		466 (100.0%)	0 (0.0%)
4	nox [numeric]	Mean (sd) : 0.6 (0.1) min ≤ med ≤ max: 0.4 ≤ 0.5 ≤ 0.9 IQR (CV) : 0.2 (0.2)	79 distinct values		466 (100.0%)	0 (0.0%)
5	rm [numeric]	Mean (sd) : 6.3 (0.7) min ≤ med ≤ max: 3.9 ≤ 6.2 ≤ 8.8 IQR (CV) : 0.7 (0.1)	419 distinct values		466 (100.0%)	0 (0.0%)
6	age [numeric]	Mean (sd) : 68.4 (28.3) min ≤ med ≤ max: 2.9 ≤ 77.2 ≤ 100 IQR (CV) : 50.2 (0.4)	333 distinct values		466 (100.0%)	0 (0.0%)
7	dis [numeric]	Mean (sd) : 3.8 (2.1) min ≤ med ≤ max: 1.1 ≤ 3.2 ≤ 12.1 IQR (CV) : 3.1 (0.6)	380 distinct values		466 (100.0%)	0 (0.0%)
8	rad [integer]	Mean (sd) : 9.5 (8.7) min ≤ med ≤ max: 1 ≤ 5 ≤ 24 IQR (CV) : 20 (0.9)	1 : 17 (3.6% ) 2 : 20 (4.3% ) 3 : 36 (7.7% ) 4 : 103 (22.1% ) 5 : 109 (23.4% ) 6 : 25 (5.4% ) 7 : 15 (3.2% ) 8 : 20 (4.3% ) 24 : 121 (26.0% )		466 (100.0%)	0 (0.0%)
9	tax [integer]	Mean (sd) : 409.5 (167.9) min ≤ med ≤ max: 187 ≤ 334.5 ≤ 711 IQR (CV) : 385 (0.4)	63 distinct values		466 (100.0%)	0 (0.0%)
10	ptratio [numeric]	Mean (sd) : 18.4 (2.2) min ≤ med ≤ max: 12.6 ≤ 18.9 ≤ 22 IQR (CV) : 3.3 (0.1)	46 distinct values		466 (100.0%)	0 (0.0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
11	lstat [numeric]	Mean (sd) : 12.6 (7.1) min ≤ med ≤ max: 1.7 ≤ 11.4 ≤ 38 IQR (CV) : 9.9 (0.6)	424 distinct values		466 (100.0%)	0 (0.0%)
12	medv [numeric]	Mean (sd) : 22.6 (9.2) min ≤ med ≤ max: 5 ≤ 21.2 ≤ 50 IQR (CV) : 8 (0.4)	218 distinct values		466 (100.0%)	0 (0.0%)
13	target [integer]	Min : 0 Mean : 0.5 Max : 1	0 : 237 (50.9% ) 1 : 229 (49.1% )		466 (100.0%)	0 (0.0%)

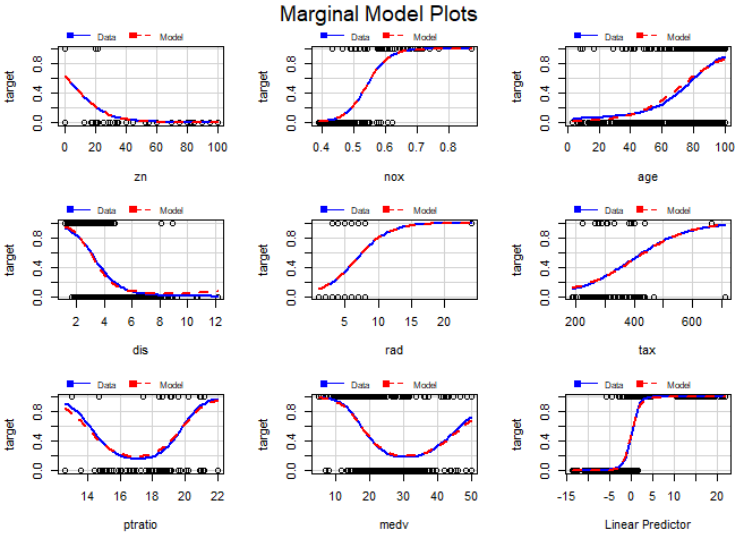
Generated by [summarytools](#) 1.0.0 (R version 4.1.0)  
2022-04-11

No data preparation, pruning or transformation are warranted at this stage so none are performed. Therefore, we proceed directly to build our binary logistic regression models.

## // Model Building and Selection

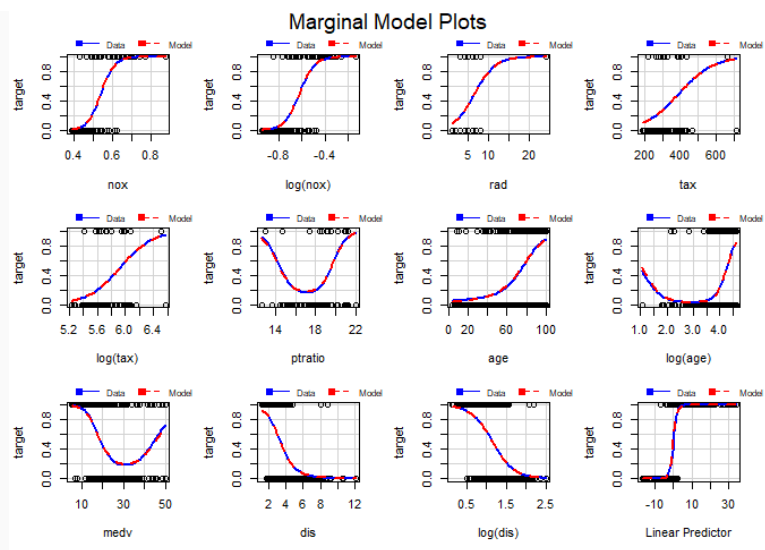
We build our model through backward and forward selection using the MASS library's `stepAIC()` function on the training data. Both methods, backward and forward selection, converge on the following model:

target			
Predictors	Log-Odds	CI	p
(Intercept)	-37.42	-50.25 – -26.45	<b>5.653e-10</b>
age	0.03	0.01 – 0.06	<b>2.622e-03</b>
dis	0.65	0.25 – 1.10	<b>2.217e-03</b>
medv	0.11	0.04 – 0.18	<b>1.829e-03</b>
nox	42.81	30.58 – 56.89	<b>1.459e-10</b>
ptratio	0.32	0.11 – 0.55	<b>3.668e-03</b>
rad	0.73	0.45 – 1.04	<b>1.293e-06</b>
tax	-0.01	-0.01 – -0.00	<b>3.459e-03</b>
zn	-0.07	-0.14 – -0.01	<b>3.203e-02</b>
Observations	466		
R <sup>2</sup> Tjur	0.733		



Seeking to reduce the AIC score and achieve better fit with predictor variables `age`, `nox`, `tax` and `dis`, we modify the model by adding log terms to these four variables, eliminating in the process variable `zn` due to non significance at 95% confidence. We then arrive at the following model with an AIC of 173.

target			
Predictors	Log-Odds	CI	p
(Intercept)	-431.27	-653.24 – -227.22	<b>6.197e-05</b>
age	0.11	0.05 – 0.16	<b>6.540e-05</b>
dis	-2.73	-4.43 – -1.13	<b>1.001e-03</b>
log(age)	-3.51	-5.61 – -1.25	<b>1.038e-03</b>
log(dis)	12.96	5.98 – 20.54	<b>4.254e-04</b>
log(nox)	-129.22	-234.08 – -23.34	<b>1.496e-02</b>
log(tax)	36.30	19.81 – 56.01	<b>6.250e-05</b>
medv	0.12	0.04 – 0.22	<b>7.144e-03</b>
nox	305.29	101.80 – 513.70	<b>3.329e-03</b>
ptratio	0.57	0.31 – 0.84	<b>2.424e-05</b>
rad	0.96	0.63 – 1.36	<b>1.575e-07</b>
tax	-0.12	-0.19 – -0.07	<b>4.184e-05</b>
Observations	466		
R <sup>2</sup> Tjur	0.815		



Behind the scenes we have manually attempted to add log terms to the other variables and/or eliminate variables in an effort to improve the fit and AIC of the model. None of those efforts reveal a model significant at 95% confidence with better fit and increased AIC.

Therefore, we select this model, as summarized and plotted above, based on its superior i.) AIC score, ii.) statistical significance and ii.) fit on the marginal model plots relative to alternative models.

## // Model Validation

We validate our selected model by splitting the training data set into 50% train and a 50% test sets by random sampling. We then fit the model with the train split and calculate its performance with the test split. We do this  $\sqrt{n} = 466$  or 22 times, each time with a different, randomly sampled train/test split. For replicability, we specify `set.seed(v*10)` where `v` is the v-th validation. The model confusion matrix on the 22nd validation, average performance metrics across the 22 validations and ROC curve are shown below:

### | Confusion Matrix

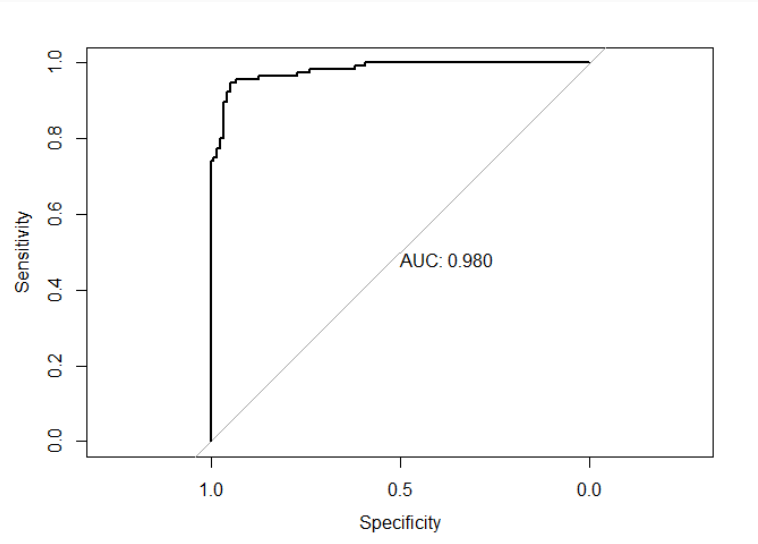
TargetPrediction			
Target	0	1	Total
0	112	6	118
1	7	108	115
Total	119	114	233

Generated by [summarytools](#) 1.0.0 (R version 4.1.0)  
2022-04-11

### | Validated Model Performance

Accuracy	Sensitivity	Specificity	Error	F1	Precision	validations
0.93	0.95	0.91	0.07	0.93	0.91	22

### | ROC graph



Testing shows that the model is robust and well-performing, with Accuracy and F1 Scores of 93%. Specificity and Precision are its least strong metrics at around 91%. The area under the ROC curve is quite high, showing good model performance once again. Based on this iterative validation, we can be reasonably confident of our selected model's predictive performance on the evaluation dataset.

## // Prediction

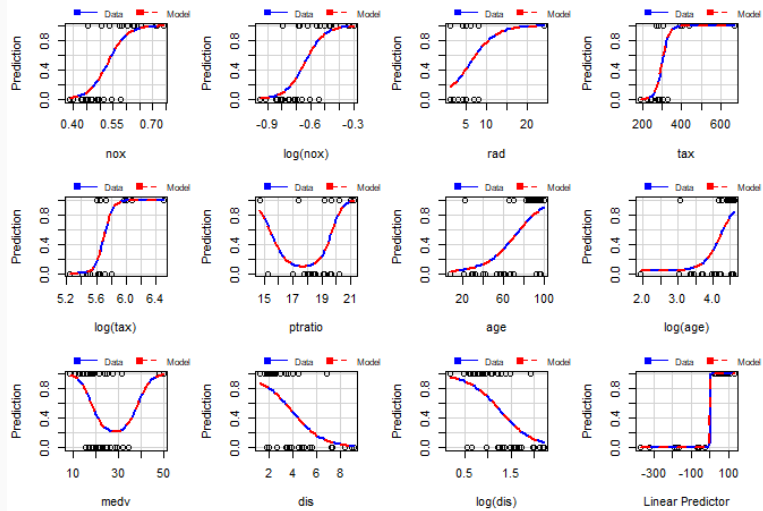
The predicted probabilities and predictions for the 40 observations in the evaluation data set are given below.

### | Evaluation dataset: predictions summary

Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid
Prediction [numeric]	Min : 0	0 : 18 ( 45.0% )	<input type="text"/>	40 (100.0%)
	Mean : 0.6	1 : 22 ( 55.0% )	<input type="text"/>	
	Max : 1			

Generated by [summarytools](#) 1.0.0 (R version 4.1.0)  
2022-04-11

### | Evaluation dataset: predicted probabilities



#### | Evaluation dataset: predictions

Observation	Prediction
1	0
2	1
3	1
4	1
5	0
6	0
7	1
8	0
9	0
10	0
11	0
12	0
13	1
14	1
15	1

Observation	Prediction
16	0
17	1
18	1
19	0
20	0
21	0
22	0
23	0
24	0
25	0
26	0
27	0
28	1
29	1
30	1
31	1
32	1
33	1
34	1
35	1
36	1
37	1
38	1
39	1
40	1