Critical Thinking Group 4: DATA621 Homework 3

Table of Contents

[TEAM Members: 2](#_Toc36931054)

[**Overview** 2](#_Toc36931055)

[**Deliverables** 3](#_Toc36931056)

[**Data Exploration** 3](#_Toc36931057)

[Missing Values & Data Type Check 4](#_Toc36931058)

[Data Statistics Summary 7](#_Toc36931059)

[Consolidated Data Dictionary 17](#_Toc36931060)

[**Data Preparation** 20](#_Toc36931061)

[Re-scale Data 20](#_Toc36931062)

[**Build Models** 22](#_Toc36931063)

[Model 1: Full Model 22](#_Toc36931064)

[Model 2: Removing Predictors Seemed Unnecessary 23](#_Toc36931065)

[Model 3: Removing Highest VIF Values 24](#_Toc36931066)

[Model 4: Removing Poor Predictors 25](#_Toc36931067)

[Model 5: Stepwise Based on AIC 26](#_Toc36931068)

[Model 6: Stepwise Based on BIC 33](#_Toc36931069)

[Model 7: Best Subset Based on AIC 37](#_Toc36931070)

[Model 8: Best Subset Based on BIC 38](#_Toc36931071)

[**Select Models** 40](#_Toc36931072)

[Fourfold Plots 41](#_Toc36931073)

[Summary Statistics 42](#_Toc36931074)

[ROC / AUC 44](#_Toc36931075)

[R^2, AIC, AICc & BIC 45](#_Toc36931076)

[Model Selection 47](#_Toc36931077)

[Odds Ratio 49](#_Toc36931078)

[**Make Predictions** 50](#_Toc36931079)

[**Appendix** 55](#_Toc36931080)

## TEAM Members:

Rajwant Mishra  
Priya Shaji  
Debabrata Kabiraj  
Isabel Ramesar  
Sin Ying Wong  
Fan Xu

-------------------------------------------------

**Overview**

In this homework assignment, you will explore, analyze and model a dataset 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 dataset to predict whether the neighborhood will be at risk for high crime levels. You will provide classifications and probabilities for the evaluation dataset using your binary logistic regression model. You can only use the variables given to you (or variables that you derive from the variables provided). Below is a short description of the variables of interest in the dataset:

| **Variable Name** | **Definition** | **Variable Type** |
| --- | --- | --- |
| zn | proportion of residential land zoned for large lots (over 25000 square feet) | predictor |
| indus | proportion of non-retail business acres per suburb | predictor |
| chas | a dummy var. for whether the suburb borders the Charles River (1) or not (0) | predictor |
| nox | nitrogen oxides concentration (parts per 10 million) | predictor |
| rm | average number of rooms per dwelling | predictor |
| age | proportion of owner-occupied units built prior to 1940 | predictor |
| dis | weighted mean of distances to five Boston employment centers | predictor |
| rad | index of accessibility to radial highways | predictor |
| tax | full-value property-tax rate per $10,000 | predictor |
| ptratio | pupil-teacher ratio by town | predictor |
| black | 1000(Bk - 0.63)2 where Bk is the proportion of blacks by town | predictor |
| lstat | lower status of the population (percent) | predictor |
| medv | median value of owner-occupied homes in $1000s | predictor |
| target | whether the crime rate is above the median crime rate (1) or not (0) | response |

**Deliverables**

A write-up of your solutions submitted in PDF format. Assigned prediction (probabilities, classifications) for the evaluation dataset. Use 0.5 threshold.

**Data** **Exploration**

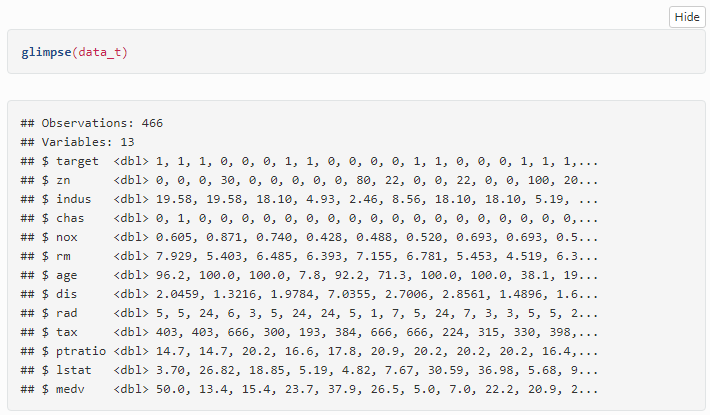
We have two datasets. One is the training set, which includes 12 candidate predictors, 1 response variable, and 466 observations. The other one is the evaluation set, which includes 12 candidate predictors only, 40 observations.

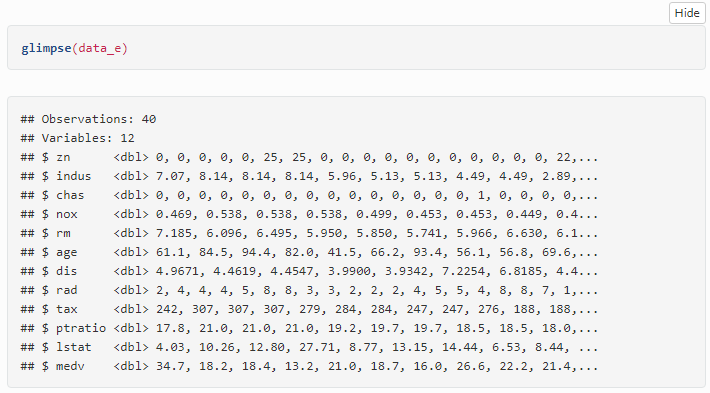
We are going to study their missing values, data types and data statistics.



## Missing Values & Data Type Check

In the training set, there are 12 candidate predictors and 1 response variable with 466 observations. In the evaluation set, there are 12 candidate predictors with 40 observations. Both datasets have no missing values (eg: NA, NULL or ’’). However, the variable black, which is described in the overview section, is not presented in both datasets.

Among the 12 candidate predictors, 1 is categorical (chas), the other 11 are continuous numerical. The response variable target is categorical.





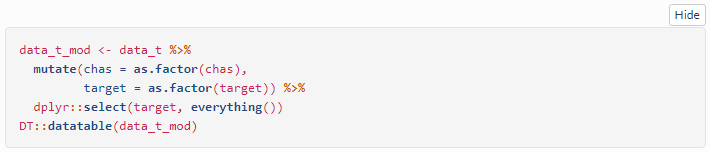
Below is the summary of the datasets and some inference of it.

1. It seems there are no Null values in the predictor and response variables.
2. Each variables are in different scale.
3. Categorical variables are chas and target.
4. There are a total of 466 observations and 12 predictor variables and 1 response variable.

## Data Statistics Summary

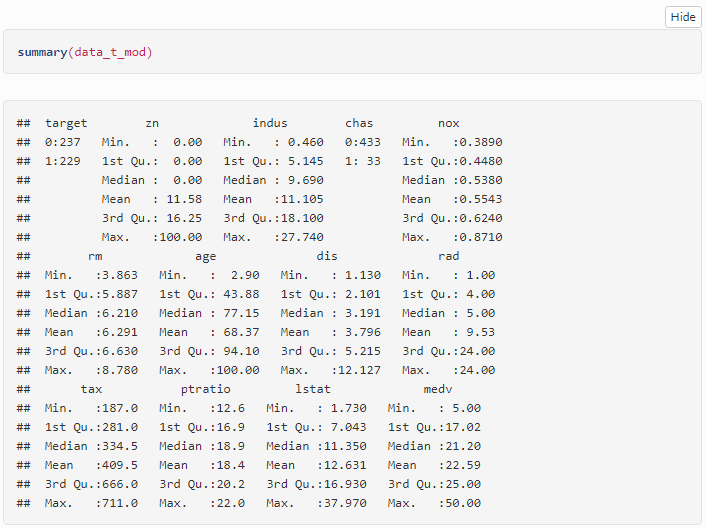
A binary logistic regression model is built using the training set, therefore the training set is used for the following data exploration.

The data types in the raw dataset are all ‘doubles’, however the candidate predictor chas and the response variable target are categorical, therefore, we update the data types of these two variables to ‘factors’.

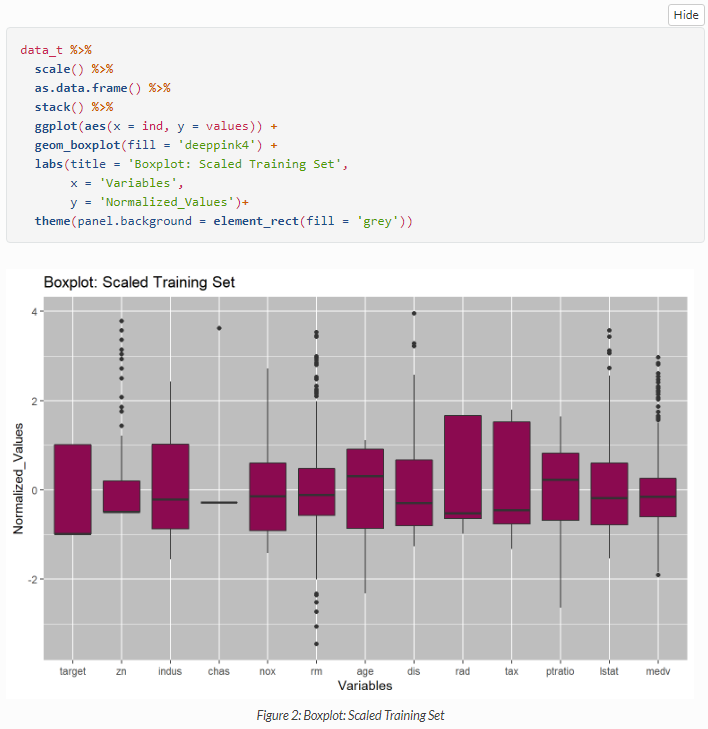




The statistics of all variables are list below:



The box plot below shows that outliners exist in variables zn, rm, dis, istat, medv. We use scaled training set to draw the box plot to show the corresponding outliers by ratio.



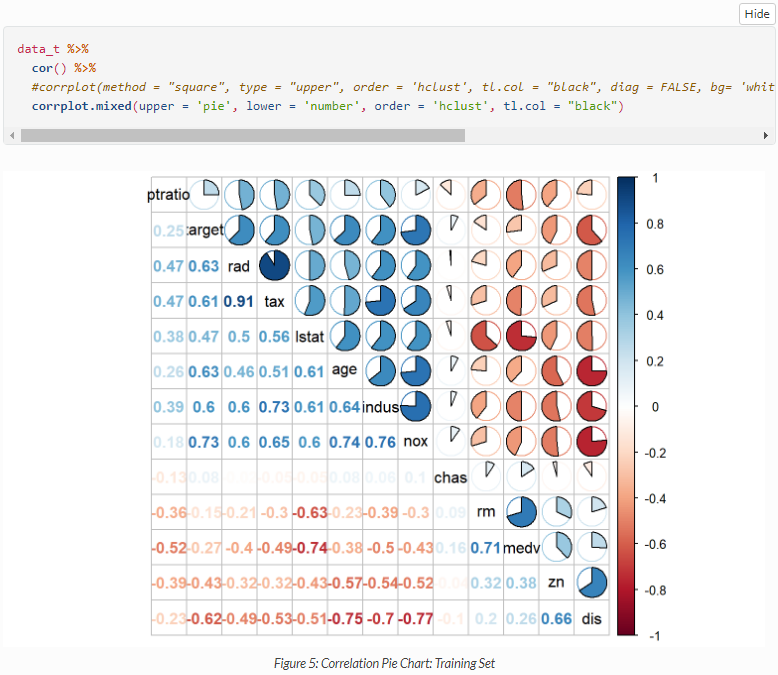
The scaled histogram and density plot show that variables zn,nox, dis, lstat, medv are right skewed; age, ptratio are left skewed; rad, tax are bimodal; target, chas are categorical however target is close to unbias while chas is highly biased; the rest are close to normal.





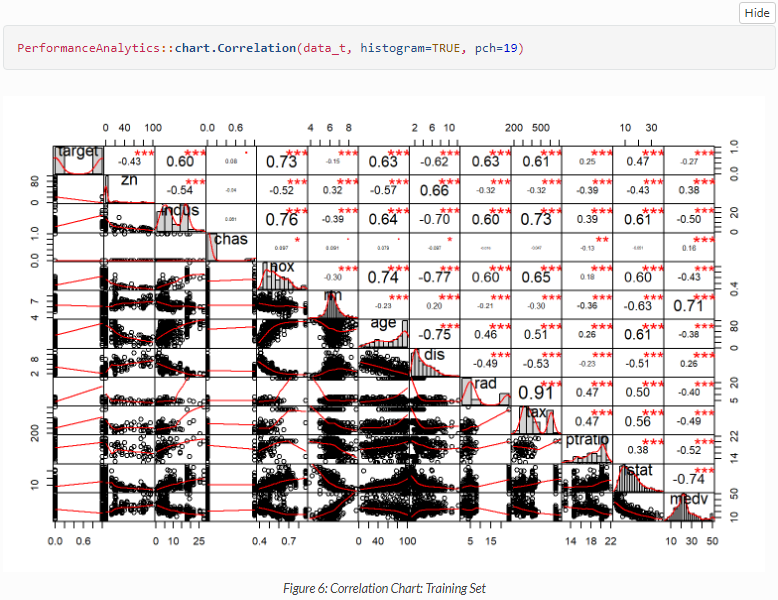
The correlation matrix below shows that the response variable target has strong positive relationship (>=0.6) with variables rad,tax,age,indus,nox, and strong negative relationship (<=-0.6) with variable dis.

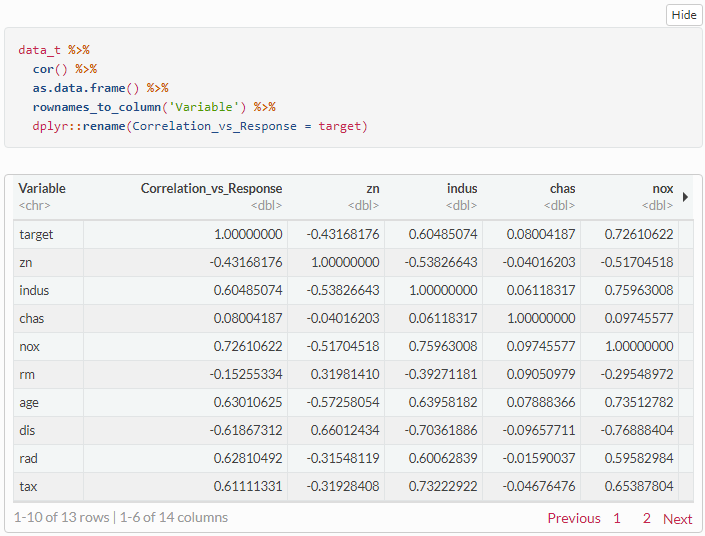
Meanwhile, it worths notice that some pairs of candidate predictors have strong correlationship, such as rad and tax (0.92), indus and nox (0.76), nox and dis (-0.77), etc.



We implement a correlation matrix to better understand the correlation between variables in the dataset. The below matrix is the results and we noticed a few interesting correlations.

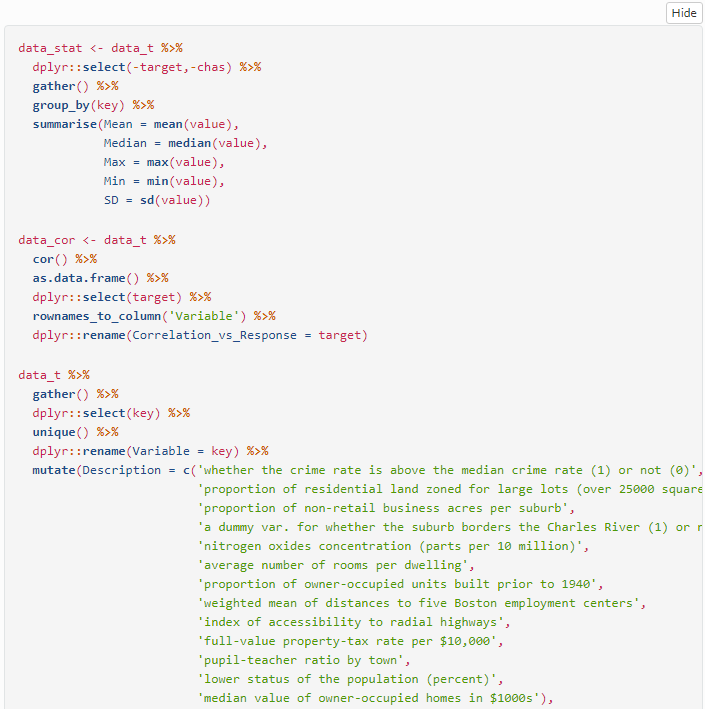
* nox : High nitrogen oxides concentration (parts per 10 million) (“nox”) is positively correlated with higher than median crime rates. As defined by the EPA - “NOx pollution is emitted by automobiles, trucks and various non-road vehicles (e.g., construction equipment, boats, etc.) as well as industrial sources such as power plants, industrial boilers, cement kilns, and turbines”. It is clear to see that nox is concentrated in areas of high road traffic and possible high industrial use which would be neighborhoods of low value and may attract crime.
* dis: The weighted mean of distances is negatively correlated with a city with higher than median crime rate. This is intuitive in that employment centers would be more closely located in cities of high crime due to high unemployment being positively correlated with higher crimes rates.
* tax : It is also counterintuitive how the crime rate has a positive correlation with the property tax. It would be anticipated that if the property tax increases, the crime rate would decrease due to the money that home occupants and owners would spend on “promised” security systems. However, when the crime rate starts to increase, the housing prices would decrease due to the fact that the home occupants and owners would not want to risk their safety.

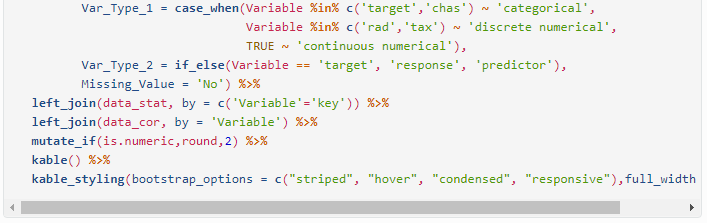




## Consolidated Data Dictionary

As a summary of the data exploration process, a data dictionary is created below:



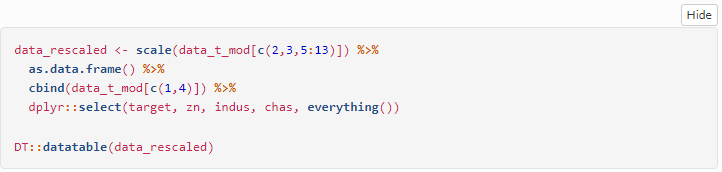


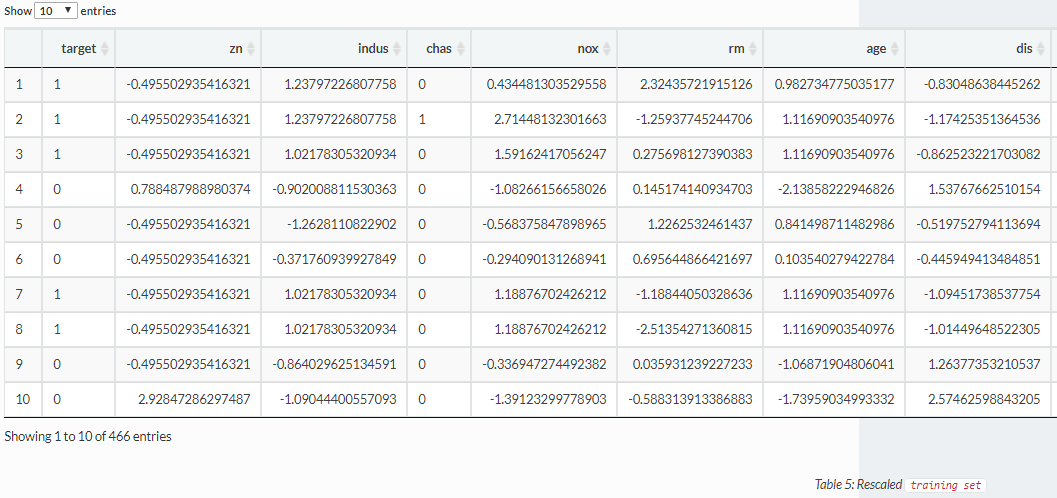
| **Variable** | **Description** | **Var\_Type\_1** | **Var\_Type\_2** | **Missing\_Value** | **Mean** | **Median** | **Max** | **Min** | **SD** | **Correlation\_vs**  **Response** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| target | whether the crime rate is above the median crime rate (1) or not (0) | categorical | response | No | NA | NA | NA | NA | NA | 1.00 |
| zn | proportion of residential land zoned for large lots (over 25000 square feet) | continuous numerical | predictor | No | 11.58 | 0.00 | 100.00 | 0.00 | 23.36 | -0.43 |
| indus | proportion of non-retail business acres per suburb | continuous numerical | predictor | No | 11.11 | 9.69 | 27.74 | 0.46 | 6.85 | 0.60 |
| chas | a dummy var. for whether the suburb borders the Charles River (1) or not (0) | categorical | predictor | No | NA | NA | NA | NA | NA | 0.08 |
| nox | nitrogen oxides concentration (parts per 10 million) | continuous numerical | predictor | No | 0.55 | 0.54 | 0.87 | 0.39 | 0.12 | 0.73 |
| rm | average number of rooms per dwelling | continuous numerical | predictor | No | 6.29 | 6.21 | 8.78 | 3.86 | 0.70 | -0.15 |
| age | proportion of owner-occupied units built prior to 1940 | continuous numerical | predictor | No | 68.37 | 77.15 | 100.00 | 2.90 | 28.32 | 0.63 |
| dis | weighted mean of distances to five Boston employment centers | continuous numerical | predictor | No | 3.80 | 3.19 | 12.13 | 1.13 | 2.11 | -0.62 |
| rad | index of accessibility to radial highways | discrete numerical | predictor | No | 9.53 | 5.00 | 24.00 | 1.00 | 8.69 | 0.63 |
| tax | full-value property-tax rate per $10,000 | discrete numerical | predictor | No | 409.50 | 334.50 | 711.00 | 187.00 | 167.90 | 0.61 |
| ptratio | pupil-teacher ratio by town | continuous numerical | predictor | No | 18.40 | 18.90 | 22.00 | 12.60 | 2.20 | 0.25 |
| lstat | lower status of the population (percent) | continuous numerical | predictor | No | 12.63 | 11.35 | 37.97 | 1.73 | 7.10 | 0.47 |
| medv | median value of owner-occupied homes in $1000s | continuous numerical | predictor | No | 22.59 | 21.20 | 50.00 | 5.00 | 9.24 | -0.27 |

**Data Preparati****on**

## Re-scale Data

The dataset contains variables of different measurements, such as percentage, distance, money values, etc. To put all the predictors and the response on a comparable scale, they are all normalized with mean = 0 and SD = 1.







**Build Models**

Because we have a small number of observations to train over, we will use k-fold Cross Validation to train, with k = 10. We’ll hold out 15% of the data for validation while doing the initial modeling, but once we select our model, we will retrain over the full training set.

Each of our logistic regression models will use binomial regression with a logit link function.

## Model 1: Full Model

The first model includes all the variables. A review of the VIF output of the model suggests some points that are highly colinear and a number of variables that may not be necessary. Model 1 uses the formula:

**target ~ .**



|  | **x** |
| --- | --- |
| zn | 1.775536 |
| indus | 2.615682 |
| chas1 | 1.289891 |
| nox | 4.090926 |
| rm | 6.680172 |
| age | 2.408913 |
| dis | 3.574289 |
| rad | 2.078134 |
| tax | 2.209580 |
| ptratio | 2.433736 |
| lstat | 2.735861 |
| medv | 9.246747 |

## Model 2: Removing Predictors Seemed Unnecessary

Our second model ignores the colinear issues but removes models that seemed unnecessary in Model #1. Model 2 uses the formula:

**target ~ zn + nox + age + dis + rad + ptratio + medv**

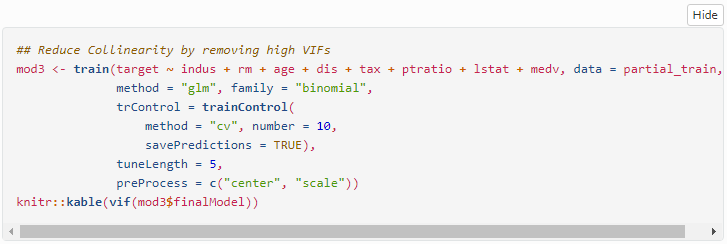


|  | **x** |
| --- | --- |
| zn | 1.801287 |
| nox | 3.049522 |
| age | 1.685178 |
| dis | 3.659469 |
| rad | 1.235992 |
| ptratio | 1.826575 |
| medv | 2.094548 |

## Model 3: Removing Highest VIF Values

Model #3 removes the variables with the 2 highest VIF values from model1. The model formula is:

**target ~ indus + rm + age + dis + tax + ptratio + lstat + medv**

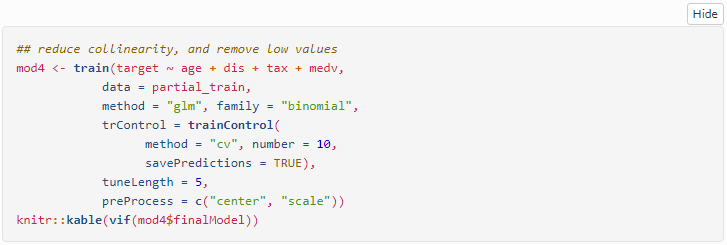


|  | **x** |
| --- | --- |
| indus | 2.190206 |
| rm | 4.462813 |
| age | 2.097140 |
| dis | 1.956005 |
| tax | 1.749705 |
| ptratio | 1.423980 |
| lstat | 2.765737 |
| medv | 5.782926 |

## Model 4: Removing Poor Predictors

Model #4 takes the advances in model #3 and removes those values shown to be poor predictors.

**target ~ age + dis + tax + medv**

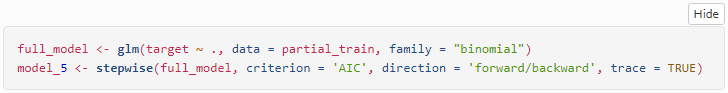


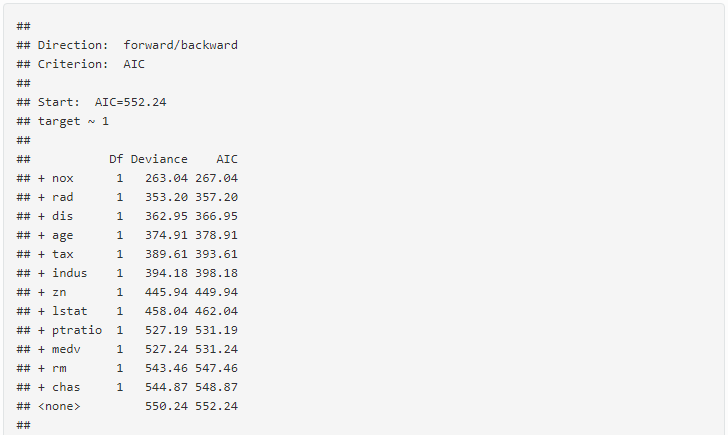
|  | **x** |
| --- | --- |
| age | 1.733106 |
| dis | 1.715677 |
| tax | 1.386751 |
| medv | 1.413739 |

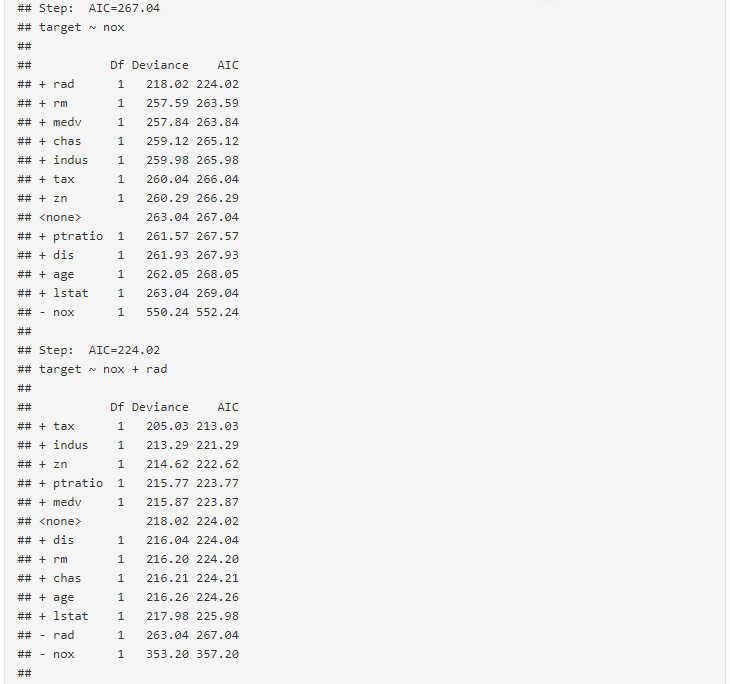
## Model 5: Stepwise Based on AIC

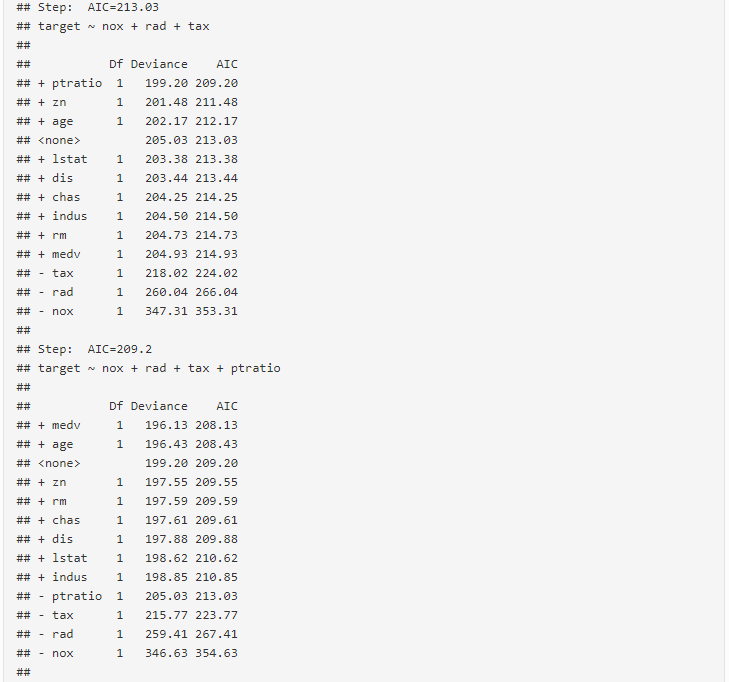
Model #5: We use stepwise function based on AIC criterion in both direction and get Model #5 in 10 steps.

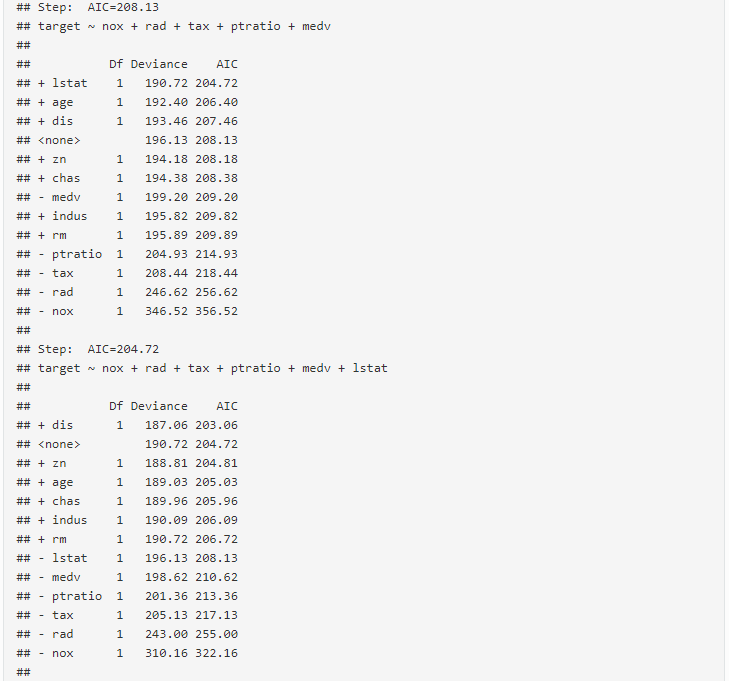
**target ~ nox + rad + tax + ptratio + medv + lstat + dis + zn + age**

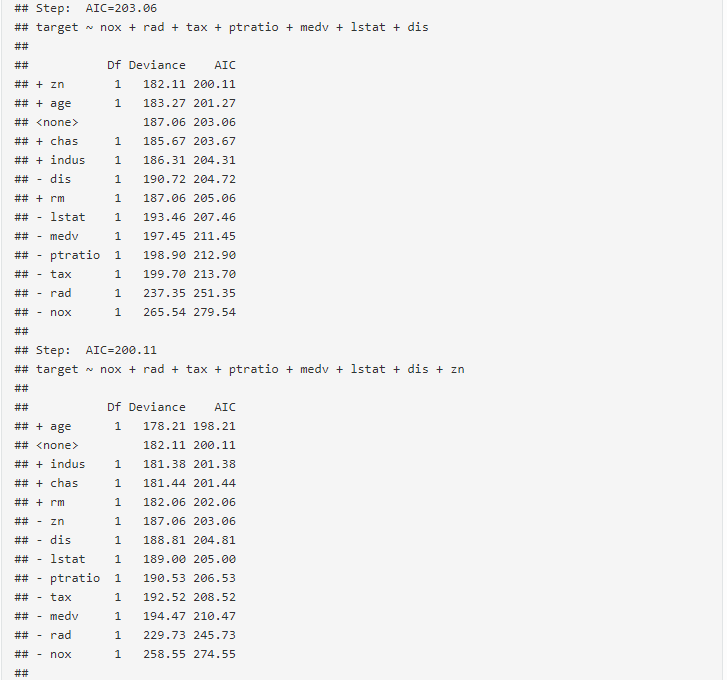


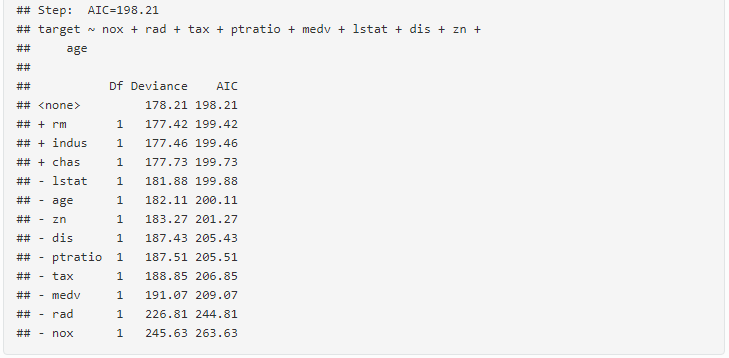


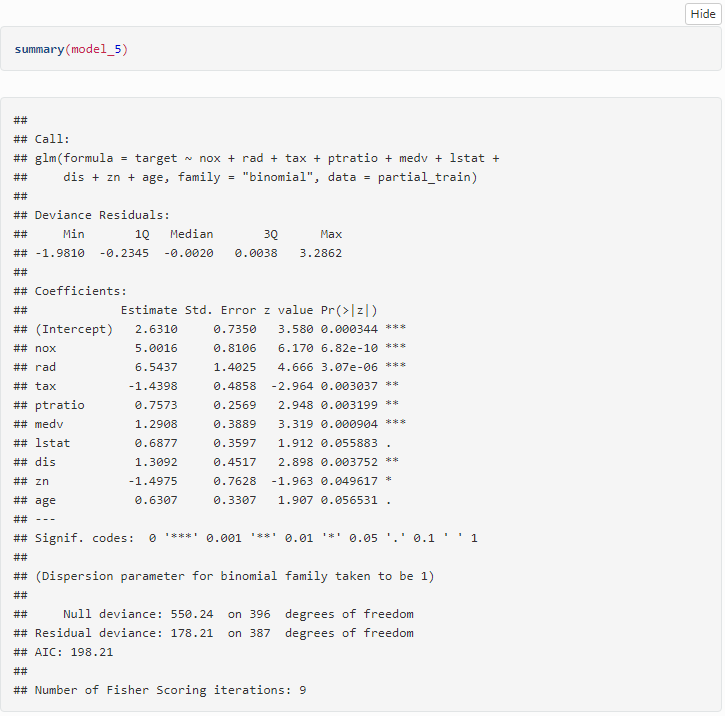








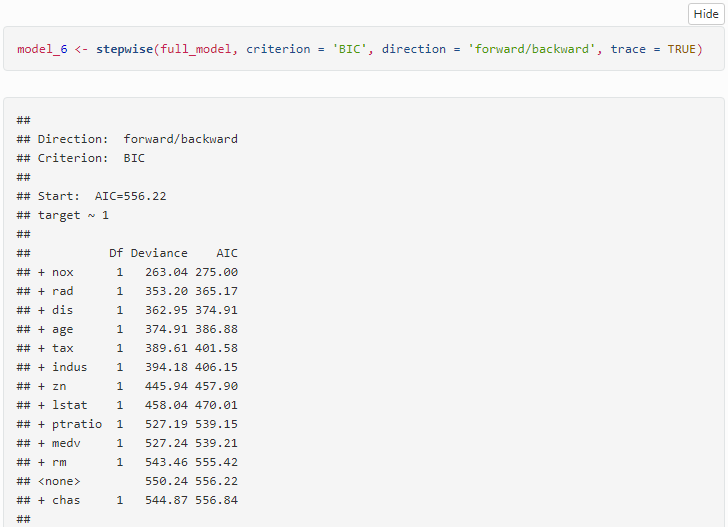


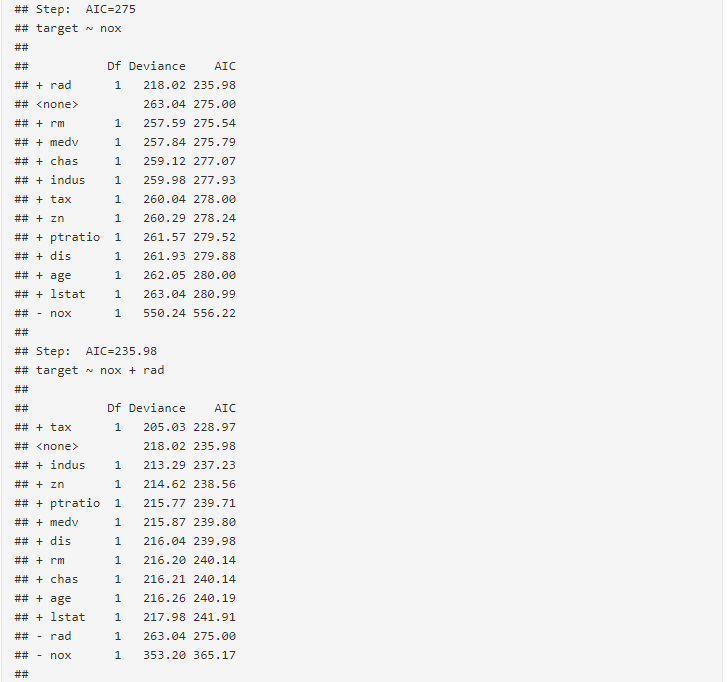


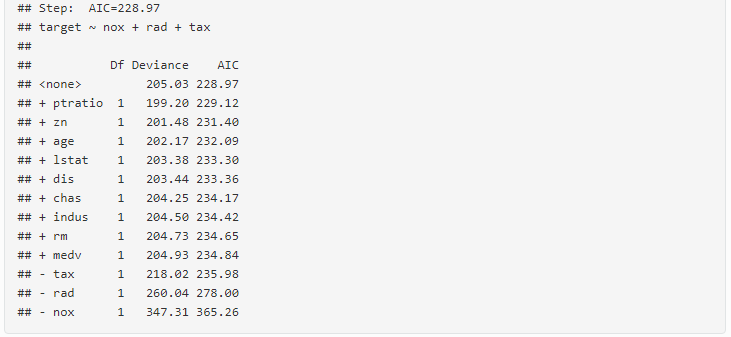
## Model 6: Stepwise Based on BIC

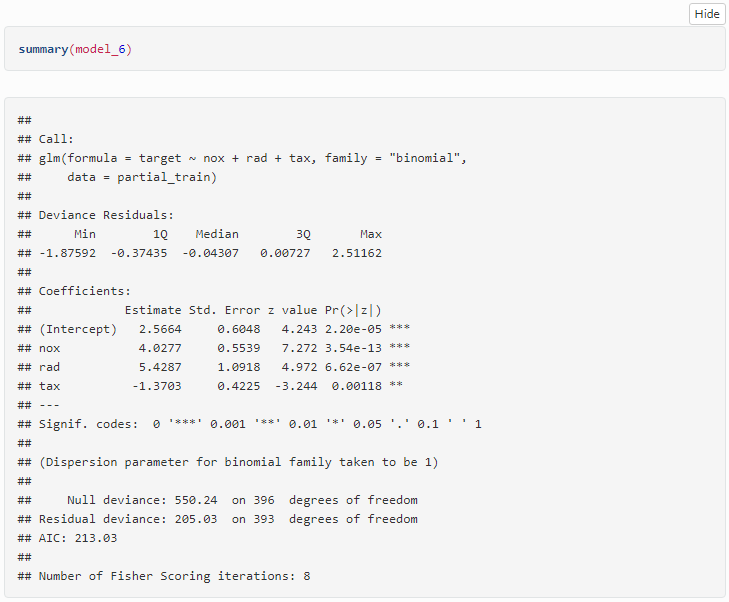
Model #6: We use stepwise function based on BIC criterion in both direction and get Model #6 in 4 steps.

**target ~ nox + rad + tax**









## Model 7: Best Subset Based on AIC

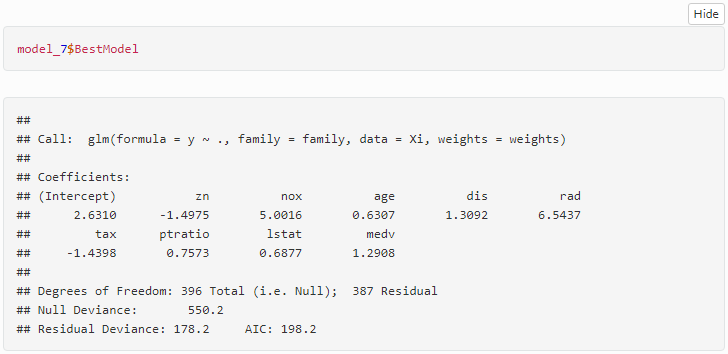
Model #7: We use best subset method based on AIC criterion to find Model #7.

**target ~ zn + nox + age + dis + rad + tax + ptratio + lstat + medv** (Same as Model 5)



| **model\_rank** | **zn** | **indus** | **chas** | **nox** | **rm** | **age** | **dis** | **rad** | **tax** | **ptratio** | **lstat** | **medv** | **Criterion** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | TRUE | FALSE | FALSE | TRUE | FALSE | TRUE | TRUE | TRUE | TRUE | TRUE | TRUE | TRUE | 196.2099 |
| 2 | TRUE | FALSE | FALSE | TRUE | TRUE | TRUE | TRUE | TRUE | TRUE | TRUE | TRUE | TRUE | 197.4200 |
| 3 | TRUE | TRUE | FALSE | TRUE | FALSE | TRUE | TRUE | TRUE | TRUE | TRUE | TRUE | TRUE | 197.4567 |
| 4 | TRUE | FALSE | FALSE | TRUE | TRUE | TRUE | TRUE | TRUE | TRUE | TRUE | FALSE | TRUE | 197.5250 |
| 5 | TRUE | FALSE | TRUE | TRUE | FALSE | TRUE | TRUE | TRUE | TRUE | TRUE | TRUE | TRUE | 197.7333 |

The rank 1 model is selected as model 7.



## Model 8: Best Subset Based on BIC

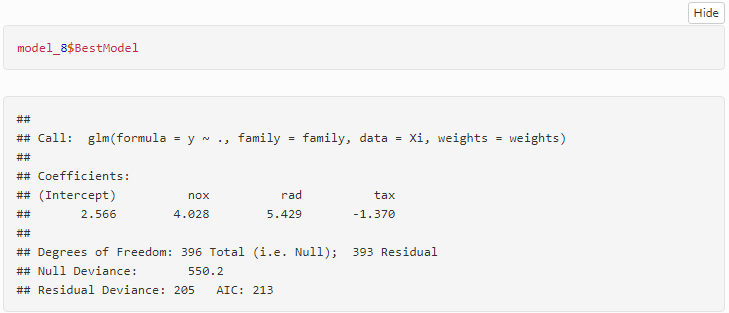
Model #8: We use best subset method based on BIC criterion to find Model #8.

**target ~ nox + rad + tax** (Same as Model 6)



| **model\_rank** | **zn** | **indus** | **chas** | **nox** | **rm** | **age** | **dis** | **rad** | **tax** | **ptratio** | **lstat** | **medv** | **Criterion** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE | FALSE | FALSE | FALSE | 222.9816 |
| 2 | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE | TRUE | FALSE | FALSE | 223.1394 |
| 3 | TRUE | FALSE | FALSE | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE | FALSE | FALSE | FALSE | 225.4175 |
| 4 | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE | FALSE | TRUE | TRUE | TRUE | FALSE | TRUE | 226.0475 |
| 5 | FALSE | FALSE | FALSE | TRUE | FALSE | TRUE | FALSE | TRUE | TRUE | FALSE | FALSE | FALSE | 226.1033 |

The rank 1 model is selected as model 8.

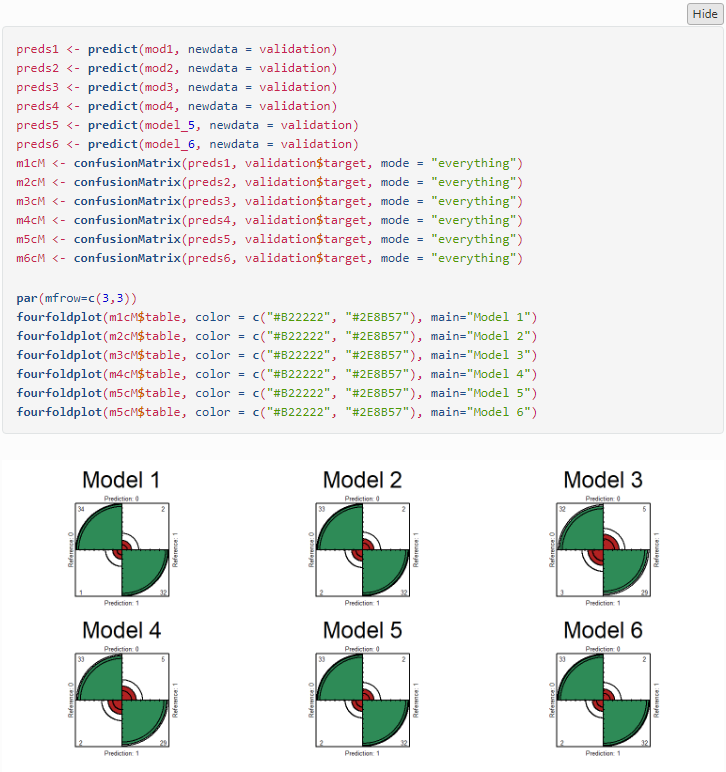




**Select Models**

To help aid in model selection, we will review their accuracy by making predictions on our holdout validation set, and comparing their performance using a variety of confusion matrix adjacent functions like fourfold plots, summary statistics, and ROC / AUC plots.

## Fourfold Plots



## Summary Statistics

Model 1, Model 2 and Model 5 have best performance in at least one category.



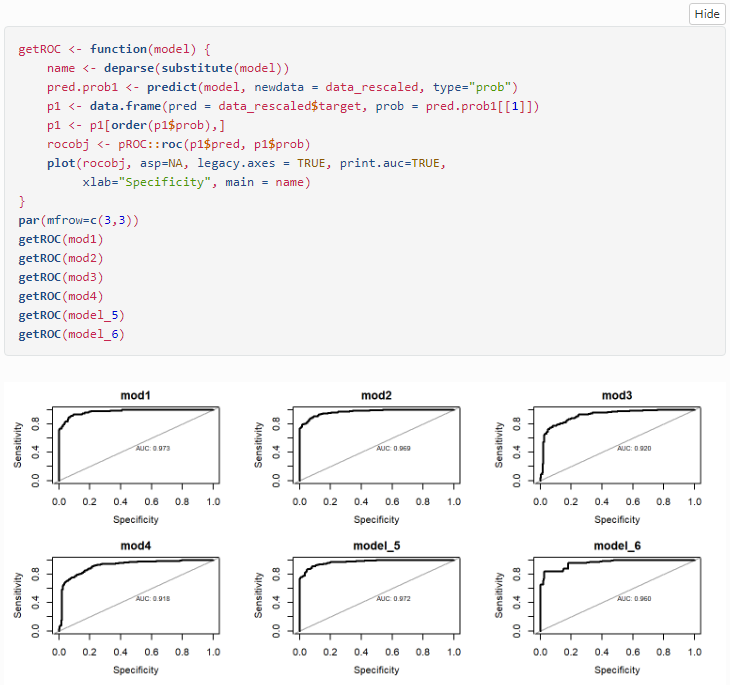


| **Model** | **Accuracy** | **Classification\_Error\_Rate** | **Precision** | **Sensitivity** | **Specificity** | **F1** |
| --- | --- | --- | --- | --- | --- | --- |
| **Model 1** | **0.957** | **0.043** | **0.944** | **0.971** | **0.941** | **0.958** |
| **Model 2** | 0.942 | 0.058 | 0.943 | 0.943 | **0.941** | 0.943 |
| Model 3 | 0.884 | 0.116 | 0.865 | 0.914 | 0.853 | 0.889 |
| Model 4 | 0.899 | 0.101 | 0.868 | 0.943 | 0.853 | 0.904 |
| **Model 5** | 0.942 | 0.058 | 0.943 | 0.943 | **0.941** | 0.943 |
| Model 6 | 0.928 | 0.072 | 0.917 | 0.943 | 0.912 | 0.93 |
| Model 7 (Same as Model 5) | NA | NA | NA | NA | NA | NA |
| Model 8 (Same as Model 6) | NA | NA | NA | NA | NA | NA |

## ROC / AUC

The larger the area under the curve, the better the model.

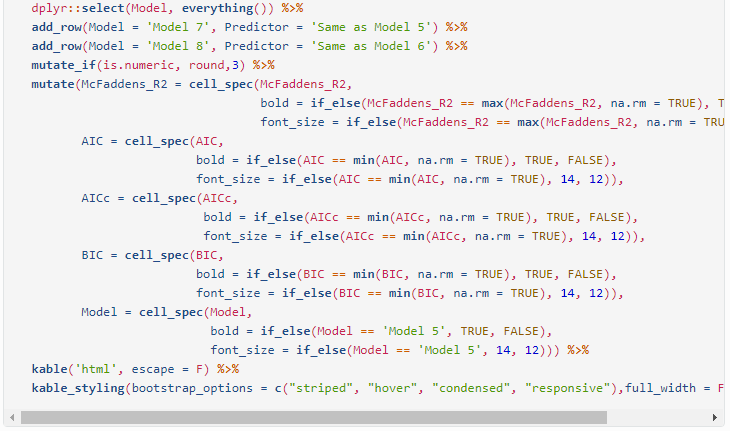
AUC: model 1 > model 5 > model 2 > model 6 > model 3 > model 4



## R^2, AIC, AICc & BIC

Although Model 1 has the largest R^2, Model 5 has the smallest AIC and AICc, and the second largest R^2.

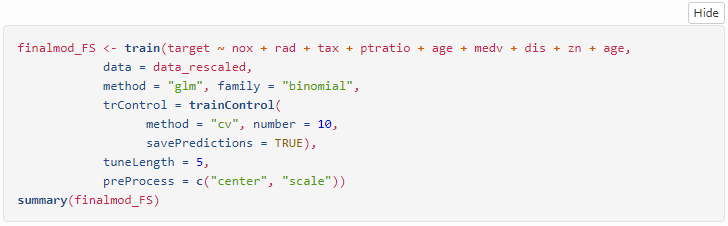


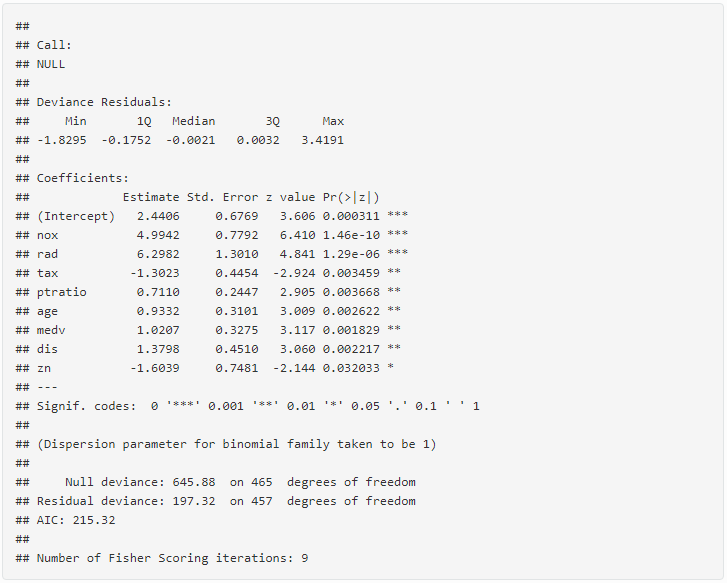


| **Model** | **Predictor** | **McFaddens\_R2** | **AIC** | **AICc** | **BIC** |
| --- | --- | --- | --- | --- | --- |
| Model 1 | zn,indus,chas1,nox,rm,age,dis,rad,tax,ptratio,lstat,medv | **0.68** | 201.853 | 202.804 | 253.644 |
| Model 2 | zn,nox,age,dis,rad,ptratio,medv | 0.653 | 207.183 | 207.555 | 239.055 |
| Model 3 | indus,rm,age,dis,tax,ptratio,lstat,medv | 0.463 | 313.336 | 313.801 | 349.191 |
| Model 4 | age,dis,tax,medv | 0.455 | 310.012 | 310.165 | 329.932 |
| **Model 5** | nox,rad,tax,ptratio,age,medv,dis,zn | 0.669 | **199.883** | **200.348** | 235.739 |
| Model 6 | nox,rad,tax | 0.627 | 213.03 | 213.132 | **228.966** |
| Model 7 | Same as Model 5 | NA | NA | NA | NA |
| Model 8 | Same as Model 6 | NA | NA | NA | NA |

## Model Selection

From the model selection process above, we know that Model 1 suffers from co-linearity issues, the rest of the models tried to eliminate these issues but also to achieve best prediction performance. Among them, Model 5 has 1) the highest Specificity, 2) second highest accuracy, precision, sensitivity, F1 Score, AUC and McFadden’s R squared proceed by model1, 3) lowest AIC and AICc. Therefore Model 5 is selected to be the final model.





## Odds Ratio

We will also create a table of the Odds Ratio for our final model beside the 95% confidence interval of those boundaries. Odd Ratio (OR) is a measure of association between exposure and an outcome. The OR represents the odds that an outcome will occur given a particular exposure, compared to the odds of the outcome occurring in the absence of that exposure.

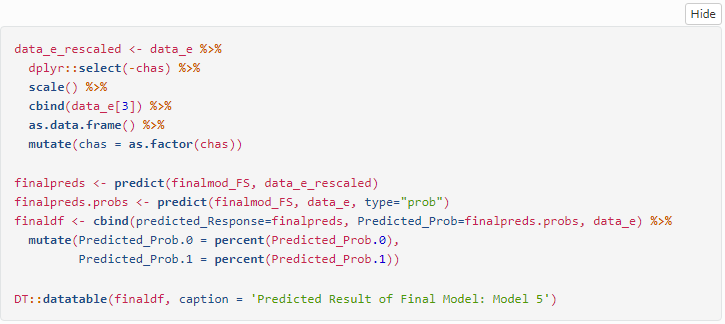


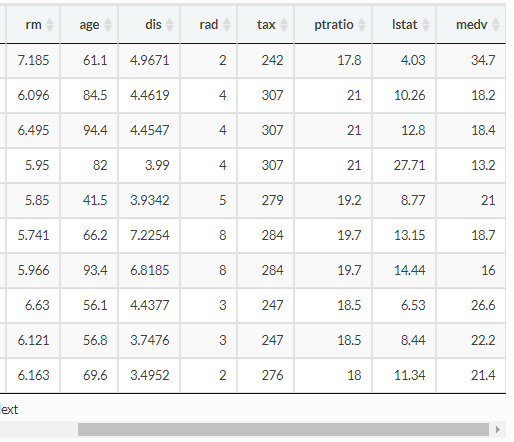
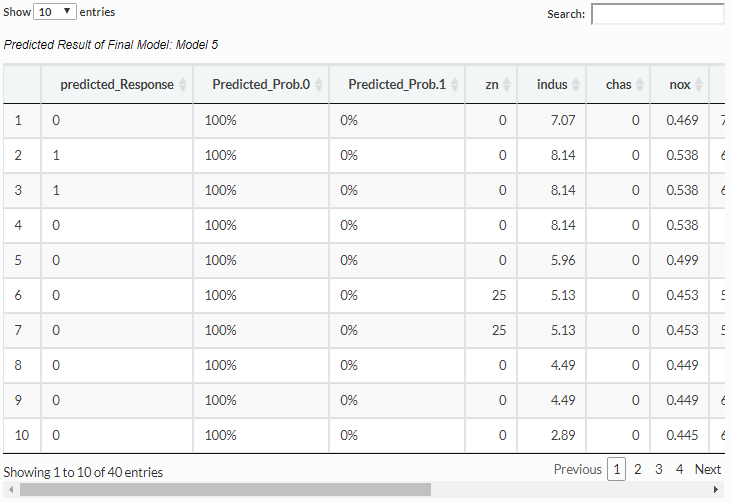
|  | **OddsRatio** | **2.5 %** | **97.5 %** |
| --- | --- | --- | --- |
| (Intercept) | 11.479 | 3.161 | 45.685 |
| nox | 147.561 | 35.436 | 763.373 |
| rad | 543.616 | 49.797 | 8423.905 |
| tax | 0.272 | 0.104 | 0.615 |
| ptratio | 2.036 | 1.274 | 3.343 |
| age | 2.543 | 1.408 | 4.780 |
| medv | 2.775 | 1.503 | 5.441 |
| dis | 3.974 | 1.693 | 10.049 |
| zn | 0.201 | 0.040 | 0.747 |

So we can now say that with a one unit increase in the scaled age variable, the odds of the neighborhood being below the median crime rate increase by 2.543%.

All that is left is to use our final model to make predictions over the evaluation dataset.

**Make Predictions**

We make our final predictions, create a dataframe with the prediction and the predicted probabilities along with the `evaluation set`. The data set is rescaled in as well. The result shows that among the 40 observations, 23 are predicted to have crime rate below median (`0`), 17 are predicted to be above median (`1`).  






**Appendix**

<https://github.com/Rajwantmishra/DATA621_CR4/blob/master/HW3/Homework3_Final.Rmd>

**Thank you**