Classification model analysis and selection for Boston's MASS per capita crime rate by town*

Manuel Vallejo Sabadell - 100383186
Universidad Carlos III de Madrid - Escuela de Postgrado
Subject: Data Analytics in CI4.0
Second Assignment

21 November 2022

Universidad
Carlos III de Madrid

*With automated and generalized R code to improve flexibility. Tried to be done in LaTeX I hope general structure is alright.

Contents

List of Figures

1 Data preprocessing - Exercises 1 and 2

As indicated by the exercise, the crime per camita rate variable, from now on called *crim*, should be converted into a binomial or binary variable. In order to do so, we need to take a cutoff and a probability for assigning either a 0 or 1. As instructed,the median will be taken as the cutoff and for both a probability of 70%.

As instructed Boston's data set will be separated into a *Train data set* (Ω_{Train}) which is equivalent to a sample of 80% of the original observations and a *Test data set* (Ω_{Test}), which will take the rest of observations

The random seed has been adjusted and kept at 100383186 (NIA) in order to maintain results throught the analysis.

2 Logistic Regression - Exercises 3 and 4

For now on, every model will be trained on Ω_{Train} and applied on Ω_{Test}. Moreover, each of the methods will be compared using the following measurements:

- Confusion Matrix
- Accuracy
- Specificity
- Sensitivity

Logistic regression is the first classification model applied to the data set. It models the probability that *crim* belongs to a particular category (either 1 or 0 - crime or not crime) assuming a logistic function.

Taking *crim* as output and the rest of the variables as predictors, the model is computed. After computing the model, it will be used to make a prediction into the test data set. That prediction will compute the probability that a specific observation tends to 1 (crime). In order to select probabilities with a significative interval, those higher than 60% of probability will be counted as *crime* and the rest as not *crime*.

Finally, using this data the confusion matrix and accuracy, sensitivity and specificity measurements are calculated, giving the following results:

LR **Confusion Matrix**

	0	**1**
0	True Negative 44	False Positive 17
1	False Negative 21	True Positive 20
Accuracy 62.7%	Specificity 67.7%	Sensitivity 54.1%

2.1 P-Values and variable selection

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.5343624  2.7571864  -0.556    0.578
zn           0.0055543  0.0069812   0.796    0.426
indus       -0.0117366  0.0302563  -0.388    0.698
chas        -0.3191094  0.4163511  -0.766    0.443
nox          1.9825658  1.9762522   1.003    0.316
rm          -0.2347867  0.2592150  -0.906    0.365
age          0.0282479  0.0070817   3.989 6.64e-05 ***
dis          0.0217060  0.1021635   0.212    0.832
rad          0.0333919  0.0311312   1.073    0.283
tax         -0.0003935  0.0018372  -0.214    0.830
ptratio      0.0284801  0.0704729   0.404    0.686
black       -0.0013445  0.0014404  -0.933    0.351
lstat       -0.0532131  0.0292668  -1.818    0.069 .
medv         0.0218728  0.0244293   0.895    0.371
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 1: Summary for Logical Regression Model

As observed in the results from Figure 1, the only variable within the requested 1% significance level is *age*, which represents the proportion of owner-occupied units built prior to 1940.

The logical regression model experiment is repeated but only taking into account the *age* predictor giving the measurements seen in the table at the start of the next page. Accuracy in predicting crime is lower, although it should be noted that a lot less data is needed for a model with an accuracy $< 10\%$ of the original

LR − Age **Confusion Matrix**

	0	**1**
0	True Negative 38	False Positive 23
1	False Negative 21	True Positive 20

Accuracy 56.9%

3 Linear Discriminant Analysis - Exercise 5

LDA **Confusion Matrix**

	0	**1**
0	True Negative 31	False Positive 30
1	False Negative 15	True Positive 26
Accuracy 55.9%	Specificity 67.4%	Sensitivity 46.4%

Linear Discriminant Analysis is a classification technique normally used for categorical variables with more than two classes. The idea is to try to find the class for which the probability that an observation belongs to that class is maximum, assuming that the data within each class are normally distributed.

As could be seen, observations are less accurate than in both LR models

Computations in R, without taking into account the theoretical depth of the matter, outside of the explicit explanation of the code are straitforward: an lda model is computed using *crim* as output and the rest of variables as predictors and then a prediction is made based on the previous model. It should be noted however that there is no necessity for cutoff, as the interpretation is made by the model.

4 K-Nearest Neighbours - Exercises 6 and 7

In K-Nearest Neighbours model, a positive integer K — which in this part of the assignment will be iterated in order to find the most accurate variation — is used to identify the K number of points in the training data closest to each observation and then estimating the conditional probability for each of the classes as the fraction of observations in previously determined subset of neighbours which response value is equal to the class.

As it happened in LDA, computation in R is determined in the code, for which a KNN model will be made using Ω_{Train} and Ω_{Test} for each K from 1 to 20 and the usual values computed (confusion matrix, accuracy ...)

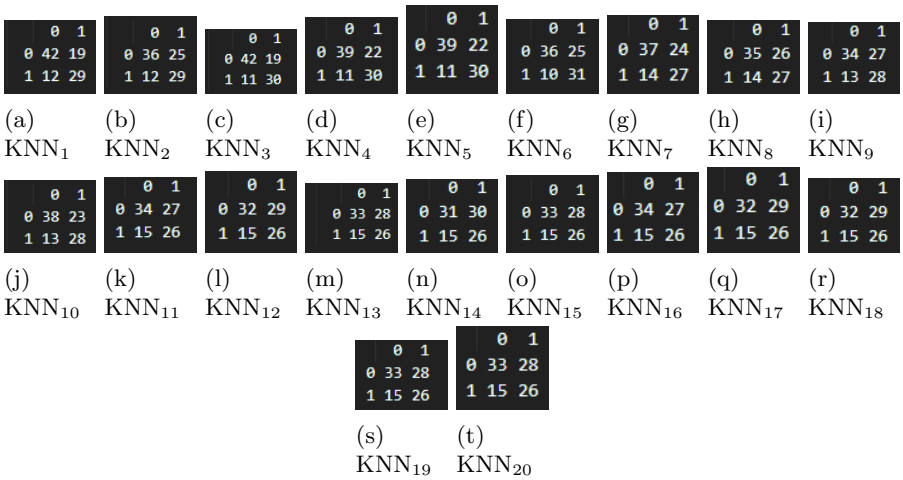

Figure 2: Confusion matrices for different values of K

From those we can infer (from left to right and from top to bottom):

$$\vec{Accuracy} = \begin{bmatrix} 0.696 & 0.637 & 0.706 & 0.676 & 0.676 & 0.657 & 0.627 & 0.608 & 0.608 & 0.647 \\ 0.588 & 0.569 & 0.578 & 0.559 & 0.578 & 0.588 & 0.569 & 0.569 & 0.578 & 0.578 \end{bmatrix} \tag{1}$$

$$\vec{Specificity} = \begin{bmatrix} 0.788 & 0.750 & 0.792 & 0.780 & 0.783 & 0.725 & 0.714 & 0.723 & 0.745 & 0.694 \\ 0.681 & 0.688 & 0.674 & 0.688 & 0.694 & 0.681 & 0.681 & 0.681 & 0.688 & 0.688 \end{bmatrix} \tag{2}$$

$$\vec{Sensitivity} = \begin{bmatrix} 0.604 & 0.537 & 0.612 & 0.577 & 0.577 & 0.554 & 0.529 & 0.509 & 0.509 & 0.549 \\ 0.491 & 0.473 & 0.481 & 0.464 & 0.481 & 0.491 & 0.473 & 0.473 & 0.481 & 0.481 \end{bmatrix} \tag{3}$$

By a quick general inspection, it is observed that the model with $K = 3$ neighbours is the most accurate and less susceptible to false positives or false negatives. Therefore being the best model. Therefore $Accuracy_{max} = 0.706$, $Specificity_{max} = 0.792$ and $Sensitivity_{max} = 0.612$ It can also be represented for easiness of inspection by plots:

Figure 3: Measurement plots for different values of K

5 Random Forest - Exercise 8

Random Forest referes to the use of a collection of Classification Trees. Forest trees are made dividing the set of possible values into distinct and non-overlapping regions which are then assigned to a class — new observations are assigned to a region. This is plotted graphically and determined numerically with each node being the dataset (root node), the regions (leaf nodes) or intermediate auxiliary nodes, and the branches are the condition for going into a node or another.

As classification trees may be unstable, Random Forest are a more reliable option that allows us to not only model the data set but also to recognize the most important predictors in the model by means of Accuracy.

RF **Confusion Matrix**

		0	**1**
	0	True Negative 35	False Positive 26
	1	False Negative 14	True Positive 27
Accuracy 60.8%		Specificity 71.4%	Sensitivity 50.9%

As mentioned, the most important predictors can also be obtained by RF with the following results:

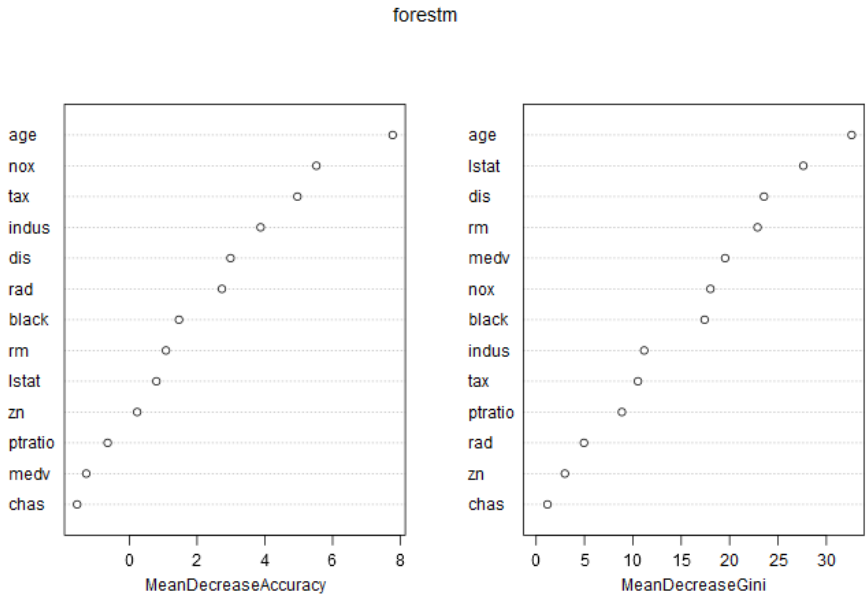

Figure 4: Random forest importance analysis

6 Forward model selection - Exercises 9, 10, 11

Finally, in order to determine which set of predictors for each model (Classification trees, logical regression, linear discriminant analysis and K-nearest neighbours) preserve the most accuracy possible, a forward model selection will be conducted in the order of the random forest's mean decrease in accuracy

Forward model selection consists in taking into account progresively each subset of variables where the first iteration it will consist of just the most relevant predictor, the second will have the second most relevant and the first and so on and so forth.

All the typical measurement values in this report are calculated for each and every iteration on each and every model. It is not going to be printed here as the extense of the data is quite big, however it is available in the R code once compiled. Note that accuracy, specificity and sensitivity are defined in vectors while the list of confussion matrices and model's are stored in lists, which allows non-unidimensional variables to be listed. The result was, based on accuracy:

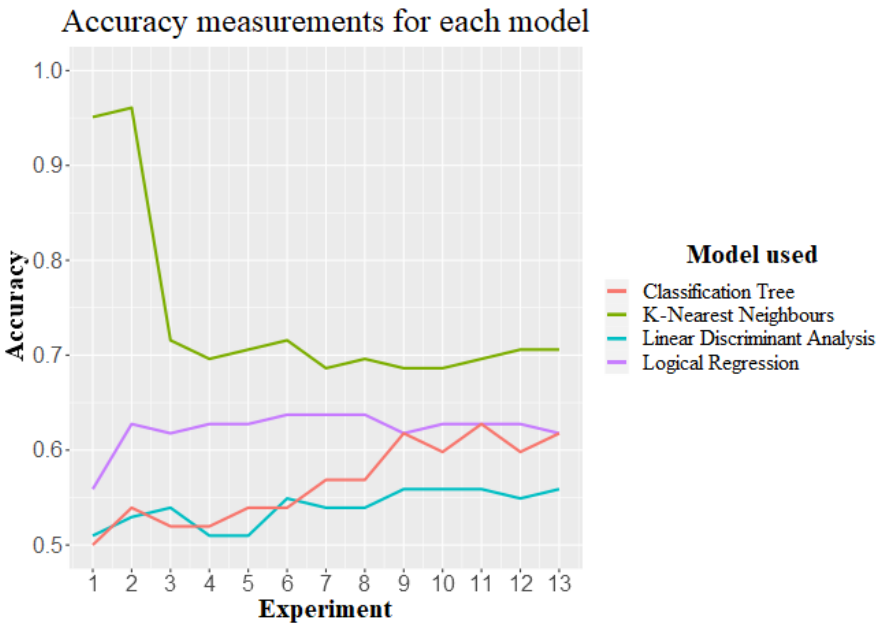

Figure 5: Final model selection

It is clearly visible that, following what was obtained in the previous sections of this report, the K-Nearest Neighbours model with $K = 3$ neighbours and predictors *age*, *nox* — the nitrogen oxides concentration in parts per 10 million — presents the most accuracy of all the models tested in this fashion with a value of:

$$Accuracy_{final} = 96.1\% \tag{4}$$

Finally, it is also remarcable that the K-Nearest Neighbours approach was also the best out of the 4 models analysed previous to this section.