# Lab03: ROC and Tree-based Methods

**Handed out:** Thursday, April 8, 2021

**Return date:** Sunday, April 18, 2021 by midnight in the *eLearning* assignment folder **Lab03Submit**.

**Grades:** Lab03 counts 13 % towards your final grade.

**Format of answers:** Your answers (statistical figures and verbal description) should be submitted ***electronically*** as Word-files. Add a running title with the following information: Lab03, your name and page numbers. You may use this document as template. Copy the requested statistical figures into your document. Trial and error answers will lead to a deduction of points. Label each answer properly with the highlighted task and sub-task headings. You are expected to hand in professionally formatted answers: use a fixed pitch font, like **Courier New**, for any  code and output (use single line-spacing). You may reduce the font size for scripts and output. Copy and paste figures into your document. Make sure that each figure has a proper ***caption*** describing its content.

## Part 1: ROC

### Task 1: Receiver Operating Curve [3 points]

[a] Calculate the *sensitivity* and *specificity* at cut-off values for predicted *positive* probabilities Also draw the associated ROC diagram.The predicted probabilities are given in the table below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | True Observed | Predicted |  | True Observed | Predicted |
| 1: | *negative* | 0.00 | **11:** | *positive* | 0.55 |
| 2: | *negative* | 0.05 | **12:** | *positive* | 0.60 |
| 3: | *negative* | 0.10 | **13:** | *positive* | 0.65 |
| 4: | *negative* | 0.15 | **14:** | *positive* | 0.70 |
| 5: | *negative* | 0.20 | **15:** | *positive* | 0.75 |
| 6: | *negative* | 0.25 | **16:** | *positive* | 0.80 |
| 7: | *negative* | 0.30 | **17:** | *positive* | 0.85 |
| 8: | *negative* | 0.35 | **18:** | *positive* | 0.90 |
| 9: | *negative* | 0.40 | **19:** | *positive* | 0.95 |
| 10: | *negative* | 0.45 | **20:** | *positive* | 1.00 |

Table 1

[b] Calculate the *sensitivity* and *specificity* at cut-off values for predicted *positive* probabilities . Also draw the associated ROC diagram. The predicted probabilities are given in the table below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | True Observed | Predicted |  | True Observed | Predicted |
| 1: | *negative* | 0.55 | **11:** | *positive* | 0.00 |
| 2: | *negative* | 0.05 | **12:** | *positive* | 0.60 |
| 3: | *negative* | 0.65 | **13:** | *positive* | 0.10 |
| 4: | *negative* | 0.15 | **14:** | *positive* | 0.70 |
| 5: | *negative* | 0.75 | **15:** | *positive* | 0.20 |
| 6: | *negative* | 0.25 | **16:** | *positive* | 0.80 |
| 7: | *negative* | 0.85 | **17:** | *positive* | 0.30 |
| 8: | *negative* | 0.35 | **18:** | *positive* | 0.90 |
| 9: | *negative* | 0.95 | **19:** | *positive* | 0.40 |
| 10: | *negative* | 0.45 | **20:** | *positive* | 1.00 |

Table 2

[c] Interpret and compare both ROC diagrams with respect to their underlying data in tasks 1 [a] and [b].

## Part 2: Trees

For the following tasks continue working with the **credit.csv** data set to predict the default probabilities. Split the data into a stratified training data set with 70% of the observations and a test data set with the remaining 30% of the observations.

For all tasks also show your properly formatted  scripts. Apply the functionality used in Boehmke et al. Chapters 9 to 12.

### Task 2: Standard Tree [2 points]

Build tree with the control parameters **mincut=5** and **minsize=10** based of the training data set. Plot the tree. Evaluate the predictive quality of this tree using the test data set with respect to the **confusionMatrix**, **auc** and the **roc** functions. Evaluate the importance of the individual features and whether particular features were not included in building the tree.

### Task 3: Pruned Tree [2 points]

Identify an optimally pruned the tree using cross-evaluation based of the training data set. Plot the final tree Evaluate the predictive quality of this pruned tree using the test data set with respect to the **confusionMatrix**, **auc** and the **roc** functions. Evaluate the importance of the individual features.

### Task 4: Bootstrapped (Bagged) Tree [2 points]

Identify an optimally bootstrapped tree based of the training data set. Evaluate the predictive quality of this bootstrapped tree using the test data set with respect to the **confusionMatrix**, **auc** and the **roc** functions. Evaluate the importance of the individual features.

### Task 5: Random Forest [2 points]

Identify an optimal random forest using the library **ranger** based of the training data set. Evaluate the predictive quality of this random forest using the test data set with respect to the **confusionMatrix**, **auc** and the **roc** functions. Evaluate the importance of the individual features.

### Task 6: Boosted Tree [2 points]

Identify an optimally boosted tree using the library **gbm** based of the training data set. Evaluate the predictive quality of this bootstrapped tree using the test data set with respect to the **confusionMatrix**, **auc** and the **roc** functions. Evaluate the importance of the individual features.