ECN431
Lab session 5
Predicting fish quality

Aud 12
Wednesday, March 21$^{\text{th}}$ 2018

Before the Lab

Obtain the datasets and the example R notebook from the GitHub repository of the course.

Purpose of this lab

Get familiar with a typical prediction task using Norwegian fisheries data.

Learning goals

- Get to know some of the basic building blocks of a typical prediction problem.
- Learn how to use logistic regression for classification tasks.

Practical information

The example R notebook (which you either pulled to your fork from the main repository or downloaded from GitHub) is opened in RStudio. This file contains several commands that you will need in order to answer the questions in these exercises, in addition to information on new commands. Remember: If/when you are uncertain about the purpose, functionality and extensions to/options for some of the commands, remember that you can get documentation for the command in RStudio by pressing F1. You should also try to google your particular problem.

Prediction of fish quality

In 2015, there were about 6700 active fishing vessels in Norway. The fishing vessels delivered fish to about 600 receiver facilities along the coast. There were over 200,000 deliveries with a combined value of about 16 billion NOK.

Since fishing is regulated by quotas, every delivery is reported to the fisheries authorities. Each observation in the data set we will use today corresponds to one reported delivery. We will use data from 2013-2015, and focus on one particular species, namely the haddock:

Figure 1: Haddock

Haddock is a very popular food fish, and is sold fresh, smoked, frozen or dried. It is also one of the most popular fish used in fish and chips.

Scenario

Imagine that you have been employed by a firm with several receiver facilities along the coast. The firm is interested in buying haddock, but only of the highest quality. For the firm, a problem has been that many of the deliveries contain haddock of only normal quality (or worse).

When deciding whether to receive a given fishing vessel, a receiver facility can ask a number of questions: where was the fish caught, which fishing tool was used, etc. Asking directly about the quality of the fish has proved ineffective (the fishing vessels often claim that the quality is higher than what turns out to be the case).

You have been given the task of coming up with a model that can be used to predict whether a given catch contains fish of the highest quality. The model will be used by the local receiving facilities when deciding whether to accept to receive a given fishing vessel or not.

You have access to reported deliveries of haddock to Norwegian fisheries in 2013-2015 (this is public information). Each report contains information about the quality of the fish (as agreed upon at the receiver facility) and a number of other variables (listed below).

Data

Please use *fisheries.dta*. Some information has been removed from the original delivery reports, either because of confidentiality (e.g., the ID of the fisher and the sales price) or because of irrelevance.

Variable list for *fisheries.dta*:

- age – number of days between the fish was caught and delivered to the receiver facility
- quality – the quality (code) of the fish (a low number corresponds to high quality)
- tool – the fishing tool used
- condition – the condition of the fish (e.g., whether it is round or gutted)
- preservation – the method used for preservation on the boat.
- area – the area in which the fish was caught

Tasks

Descriptive statistics

1. Tabulate `quality`. Create a variable `superior` that equals 1 if the quality is 10 or 12 (i.e., the two top categories), 0 otherwise.[1]

2. Try to get an impression of the relationship between `superior` and the other variables by making tables or graphs.

Divide data into test and training sets

3. We want to divide the data into a training set that we will use for analyses and model fitting, and a test set that we reserve for testing the models.

   (a) Decide on the fraction of the data you want to reserve for testing.

   (b) Split the dataset into a training and testing set in accordance with the fractions you have decided upon. Observations should be allocated to the test and training set on a random basis.

Predictive analyses

4. What is the proportion of fish in the training set that is of superior quality? This proportion is an essential benchmark when it comes to evaluation our prediction models.

5. In the following analyses, we will use logistic regression to estimate the probability of superior quality, and then use the estimated probabilities to predict whether a given catch contains superior quality. In this task you should only use `condition` as your explanatory variable.

   (a) Run a logistic regression of `superior` on `condition`.

   (b) Create a variable `yhat` that predicts the probability of superior quality for each observation.[2]

   (c) Generate a variable `class` that classifies the observations as superior or not (i.e., that predicts the label of each observation), based on the predicted probability.[3]

   (d) Create a two-way frequency table with `condition` and `yhat`.

   (e) Compare with a two-way frequency table with `condition` and `superior`.

   (f) Create a two-way frequency table with `condition` and `class`. Can you state the classification rule in words?

   (g) Generate a variable `success` that equals 1 if the classification is correct, that is if `class` is equal to `superior`, and 0 otherwise.

   (h) The *accuracy* of a model is the proportion of observations that are correctly classified. What is the accuracy of the model in the training set? What is the accuracy of the model in the test set?

[1] The function `ifelse` might be useful here

[2] Use the `predict` command.

[3] The functions `round` or `ifelse` can be used here

6. Repeat task 5 with different models, adding and removing explanatory variables as you like. Are any of your models overfitting the data?

   Extra

7. Sometimes, the cost of a false positive is not equal to the cost of a false negative. When it is very important to avoid false positives (think about falsely classifying a toxic mushroom as edible), it can make sense to evaluate a model based on its *specificity*, which is the proportion of truly negative observations that are classified as negative.[4]. Other times, it may be crucial to avoid false negatives. Then *sensitivity*, which is the proportion of the truly positive observations that are classified as positive, can be an important metric.[5] Calculate the sensitivity and specificity of (some of) your models.[6] Would you choose the same model based on the three different evaluation metrics (accuracy, sensitivity and specificity)?

[4] When specificity equals 1, all true negatives are classified as negatives and we have no false positives

[5] When sensitivity equals 1, all truly positive observations are classified as positive, and we have no false negatives

[6] The `confusionMatrix` function from the caret package can be useful here, it gives a lot of information about the predictive performance of a model