Your task is to build a model to predict the most important (and significant) habitat factors affecting the presence or absence of the great crested newt.

Question

The data you will use comes from a survey of an amphibian (great crested newt) in southern England (gcn dataset).

The first column, presence, gives the presence or absence of newts as a binary variable (0 = absence, 1= presence). The researcher used standard survey methods to detect the presence (or otherwise) of newts at 200 ponds. The other variables are habitat factors:

- `area` – the pond area in square metres.
- `dry` – pond seasonality (1-4 with 1 being non-seasonal and 4 being most seasonal).
- `water` – a subjective measure of water quality (1-4, 1 = bad).
- `shade` – the shadiness of the pond as a %.
- `bird` – presence of waterfowl (1-3, 1 = absent).
- `fish` – presence of fish (1-4, 4 = absent).
- `ponds` – number of other ponds within 1km.
- `land` – terrestrial habitat quality (1-4, 4 = good).
- `macro` – cover of macrophytes as a %.
- `HSI` – habitat suitability index (0-1). This is a standard measure compiled from other habitat measures.

In this kind of survey, the various habitat factors are converted to an index, the indices are combined to make a final HSI (habitat suitability index). The HSI is used to make it easier to assess waterbodies for their potential to support populations of the great crested newt and to give a measure of reproducibility to surveys.

Develop a model to predict the most important (and significant) habitat factors affecting the presence or absence of the great crested newt. You should also consider how well the model fits the observed data.

Assessment must contain the following parts (weight of each part between brackets) :

- `Data summary` (20%)
- `Model selection` (20%)
- `Model evaluation/prediction` (10%)
- `ROC analysis` (10%)
- `Script` (40%)