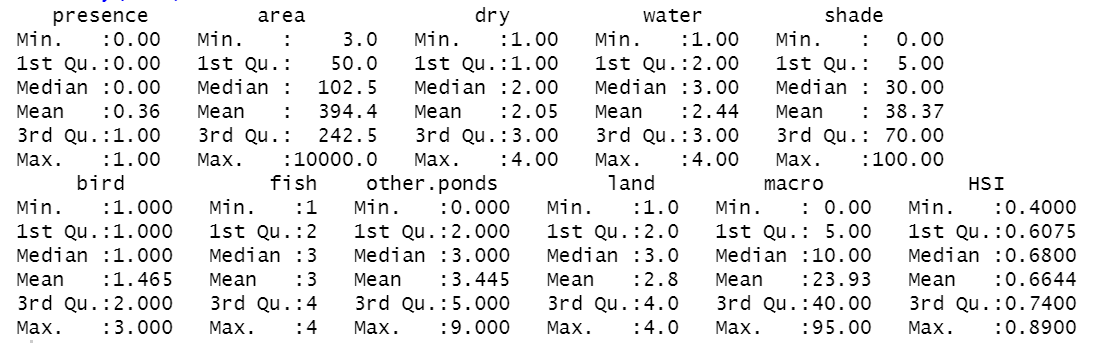
**Submission Document**

This document has everything about the problem, dataset and the solution of the great newt problem. We will be starting off from the data summary, then we will discuss the problem and the model which is most suitable for the solution. Finally, we will talk about validation/accuracy of the model and the ROC analysis.

**1) Summary of Dataset:**

Dataset has 11 columns and 200 rows. The target column is named presence. Its values have binary data type. Further, dataset has 9 other attributes which represent specific features of the environment. The last feature is HSI, this feature encapsulates essence of every other non-target attribute and provides a numeric count.

R language provides great built-in functions for data analysis. The following figure is comprehensive insight in the great newt dataset.



**2) Model Selection:** When it comes to finding if there is a relationship between variables, linear regression stands out. Linear regression finds a perfect line on the graph of comparison variables. This line with the slope tries to cover maximum plotted data. This learned line illustrates the relationship between both variables. If the line is going from down to up, it shows a positive correlation. Whereas if line is going from up to down, a negative correlation lies there.

These features of linear regression make it a perfect model for recognizing the attribute which has a positive or negative impact on the presence of newts.

**3) Model Evaluation/Prediction:**

Here I have used built-in prediction method for model evaluation and predictions. The three methods used for evaluation are coefficient of determination, mean squared error and root mean squared error. Initially I divided the dataset into two parts, namely train and test for the performance evaluation.

First of all, about the dynamics of error. A high R-square of above 0.60 is required for studies in the 'pure science' field because the behavior of molecules and/or particles can be reasonably predicted to some degree of accuracy in science research. while an R-square as low as 0.10 is generally accepted for studies in the field of arts, humanities and social sciences because human\animal behavior cannot be accurately predicted. Furthermore, if you are modeling binary (0 and 1) outcomes it is exceedingly difficult to achieve high R2 values as the predicted probability values are not very likely to be exactly 1 and 0.

Newts are unpredictable and heterogenous in their behavior as shown by the coefficient of determination 0.238, Mean squared is 0.175 and Root mean square error is 0.41 which is well below 0.5 which is considered good. This was about the model evaluation and its fitness with the provided data. Now we know model is fit enough.

Now moving towards which factor has a connection and relationship with presence of newt. For this, first of all we calculated correlations between every factor and the presence. Here c(entity1, entity2) represents correlation between entity 1 and entity 2,

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| C(area, presence) | C(dry, presence) | C(water, presence) | C(shade, presence) | C(bird, presence) | C(fish, presence) | C(ponds, presence) | C(land, presence) | C(macro, presence) | C(HSI, presence) |
| -0.09970 | -0.01644 | 0.1753102 | -0.248128 | -0.040908 | 0.1430013 | 0.2164184 | 0.0531512 | 0.2744544 | 0.3822992 |

It is clear from the correlation that HSI has the greatest positive correlation with presence (0.38). Whereas macro is the non-indexed attribute having greatest positive correlation with presence (0.27). After macro the biggest correlation is a negative correlation. Where shade and presence are corelated with magnitude of –0.248.

Now, coming towards the accuracy of linear regression models trained in one-to-one manner with each attribute. The environment attributes which best predicted the presence attribute are as follows in order.

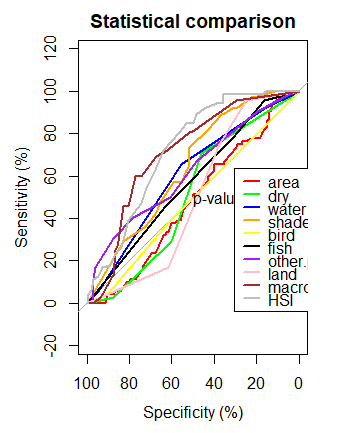
The accuracy of linear regression model with presence and HSI attribute can be judged by the mean squared error which is 0.1967264 and root mean squared error of 0.4435385.

The accuracy of linear regression model with presence and macro attribute can be judged by the mean squared error which is 0.2130451 and root mean squared error of 0.4615681.

The accuracy of linear regression model with presence and shade attribute is demonstrated by the mean squared error which is 0.2162148 and root mean squared error of 0.464989.

**3) ROC Analysis:**

Finally, ROC analysis is done which shows true positives(sensitivity) vs false positives(specificity). In a nutshell the curve having bigger area under it is statistically superior and valuable than others. Here the grey curve of HSI has biggest area under the curve making it way more superior than others. Other than the clustered index HSI, the curve with the greatest area under curve is macro. After macro, shade has the better roc. So, we can say HSI and after that macro and then shade has a better connection with the presence.



**3) Conclusion:** We used multiple different analysis/validation techniques for recognition of perfect relationship attributes and our findings conclude almost same significant attributes repeatedly. Correlation points out HSI, macro then shade as the leading significant attributes. Accuracies show HSI, macro and then shade in order of accuracies (accuracies represent a pattern and relationship). At last, ROC analysis plots the same trend having HSI (grey)with the best curve, then macro (brown) and then shade (orange). In a nutshell HSI is the most significant attribute, but if we talk about standalone attributes then macro and shade are on the lead.