P8106 Data Science Homework 2

Emil Hafeez, eh2928

Contents

Parameters for Report

Your report should have the following sections (you can add other sections if you want) and should be no more than 3 pages, excluding figures and tables. The total number of figures and tables should not exceed 6.

EMIL CONSIDERING USING THE ENDFLOAT PACKAGE IN LATEX TO GET ALL THE FIGURES INTO THE APPENDIX https://tex.stackexchange.com/questions/164140/how-to-tell-latex-to-place-all-figures-at-the-end-of-pdf-file

Section 1: Introduction

Describe your data set. Provide proper motivation for your work.

Load the dataset and describe the outcome and predictors. Provide summary statistics; examine its structure (and unit of analysis), as well as its cleanliness. Go into more detail about the outcome to provide motivation, and consider a citation.

This dataset contains 5110 observations, with de-identified individual patients as the unit of analysis. Each patient is characterized by their id, and a set of numerical and factor variables detailing basic demographic and biomedical information, like their gender, age, marriage status, and glucose level. The outcome of interest is a binary indicator of stroke.

While limited information is provided by the data source, this dataset is thought to consider xxx or yyy type of stroke as stroke outcome. major burden of disease. preventative would be good. COMEBACKTOTHISCOMEBACKTOTHISCOMEBACKTOTHISCOMEBACKTOTHISCOMEBACKTOTHISCOMEBACKTOTHIS.CITE.

What questions are you trying to answer?

This is a public health example; trying to most effectively predict stroke (as a binary classification). Simple dataset. Relevant predictors, diagnostic criteria. Thus, this is a classification question.

How to allocate resource and to promote special precautions is hard. decision support system. thus, evaluate the feasibility of several predictive methods and determine if sufficient accuracy is available from a simple dataset like this. ideally, also get insight into which factors may be most relevant.

How did you prepare and clean the data?

The data required minimal steps in order to prepare and clean it. Essentially, after importing the data, variables were transmuted into more appropriate data types for analysis (for example, character variables into factor variables). Then, the whole dataset was analyzed for missing data, and tidyverse "data tidying" best practices were implemented to rename and reorder variables. The caret package was used to create testing and training subsets. Straightforward mise en place.

Section 2: Exploratory analysis/visualization

Is there any interesting structure present in the data?

Do the plots; the one that creates pairwise scatter plots, and maybe the one from the knitted lecture 10 rmd. Comment on findings.

What were your findings?

Describe the summary statistics and talk about the potential relevance; anything else?

Here you can use any techniques as long as they are adequately explained. If you cannot find anything interesting, then describe what you tried and show that there isn't much visible structure. Data science is NOT manipulating the data in some way until you get an answer.

Section 3: Models (Biggest section!!)

What predictor variables did you include?

All of them! Name them. ID is not a predictor.

What technique did you use? What assumptions, if any, are being made by using this technique?

Logistic, Penalized logistic, GAM, MARS, LDA, QDA, naive bayes' available.

Looks like logistic regression, linear discriminant analysis (useful when classes are well-separated), and maybe use QDA and maybe KNN. Decide.

If there were tuning parameters, how did you pick their values?

CV; describe function(s) using the tuning parameter grid.

Discuss the training/test performance if you have a test data set.

Make a box plot and discuss the winner in terms of mean RMSE or diagnostic criteria/ROC/AUC.

Which variables play important roles in predicting the response?

What are the limitations of the models you used (if there are any)? Are the models flexible enough to capture the underlying truth?

Discuss limitations of selected models via ther lecture notes.

Section 4: Conclusions

What were your findings? Are they what you expect? What insights into the data can you make?

#Section 666: OPEN QUESTIONS How to deal with this imbalanced data? COnsider adding to data cleaning steps. What is the influence of so many predictors here?