Assignment 2: Imputation, Logistic Regression & Random Forests

Total Marks: /100

Due Date: Monday, 22nd August, 2pm

This assignment uses the hepatitis dataset (**hepatitis.data**). Details for this data are summarised in *hepatitis.pdf*. Note that some explanatory variables are strings but have been converted to numbers (1 or 2) – these columns will need to be turned into factors before any modelling is performed. The response variable is called **Class –** and represents one of two possible outcomes due to hepatitis. The dataset also contains NAs (missing values).

There are a number of published papers that are relevant to this assignment:

Breiman_2001_Statistical Modelling.pdf

ScientificAmerical_Bootstrap_Hepatitis.pdf

Assistant_86 Cestnik Hepatitis Decision Tree.pdf

Examine the following supplied R scripts and ensure you understand the functions.

mice.R

make.imputations(data,m=5, printFlag=FALSE) by default returns a list object with 5 imputed datasets using the mice package.

get.imp.train.test(imp,perc.train=0.9) takes the result from make.imputations(..) and returns a list with a training and test dataset. By default 90% of the data is used for training.

logistic.R

test.logistic(imp,formula,perc.train=0.9) uses imputed datasets to builds a logistic model with perc.train training data, and calculates the accuracy of the model with the remaining test data. The single value of accuracy is returned.

glm.coefficients(imp,formula,perc.train=0.9) uses imputed datasets to build a logistic model and returns the coefficients for this model. Used to evaluate variable importance.

rf.R

test.rf (imp,perc.change=0.9, formula=Class ~ .,maxdepth=20, ntrees=10) uses the imputed dataset to build a forest of decision trees and do a single test for accuracy.

rf.predictions(rf.trees,test) is used internally by *test.rf(..)* to predict the test data using the random forest decision trees.

utils.R

data.factorise(data,factor.cols) takes a dataset (data) and a vector of column numbers and creates a new dataset where the columns have been turned into factors. This new dataset is returned. For example, *newdata <- data.factorise(data,factor.cols=c(1,3,4:8))* would make a new dataset where columns 1,3,4,5,6, 7 and 8 are now defined as factors. This function just saves having to explicitly set every column as a factor.

QUESTION ONE (10 MARKS)

Visualise the patterns of missing data for **hepatitis.data** and briefly comment on the patterns you observe.

QUESTION TWO (10 MARKS)

Justify why the explanatory variable *"Sex"* should be removed from the data prior to modelling.

QUESTION THREE (45 MARKS) - Logistic Regression

Perform the following steps

3.1 Load in the hepatitis data.
3.2 Turn the response and appropriate explanatories into factors and remove the "Sex" column.
3.3 Create an imputed dataset for the hepatitis data using the default values.
3.4 Estimate the accuracy of a logistic model using all explanatories (Class ~ .) by running *test.logistic(…)* 100 times – **present the result as a boxplot**. Hint: Use *replicate(…)*.

Comment on the accuracy versus that stated in the Breiman and Diaconis/Efron paper. **Why might your result differ** from the published examples? **(10 marks)**

3.5 Breiman suggests that variable importance can be based on examining the absolute value of the coefficients of a variable divided by their standard deviation (over many runs of the model with different training sets). Using *glm.coefficients(imp,formula,perc.train=0.9)* do 100 runs of this function, produce the variable importance measure stated above, and **create a barplot showing variable importance for each explanatory**.

How stable is this measure? Rerun the model several times and comment on your observations. What might this suggest about model stability and logistic regression for the hepatitis dataset? Would you be confident in using this approach to assess which variables are important? **(10 marks)**

3.6 Implement **variable importance** using **permutation** for the logistic model and examine the stability using this approach. HINT: The function definition *variable.importance(…)* has been provided but is empty. You need to write this function, which uses permutation of one explanatory column to estimate the percentage increase in error when the variable relationship in the data has been removed. There is a function provided (*collect.var.imp(..)*) that collects up the results for all explanatories.

In your writeup include your function code for *variable.importance(…),* the resulting visualisation of the model, evidence of stability (or lack of it), and a discussion that addresses:

- Is the stability improved over the approach in part 3.5?
- Which variables appear to be important? How do these compare with the provided literature? NOTE: Be aware that you have removed one column so the Breiman paper (which just uses column numbers) needs to be interpreted with care.

(25 marks)

QUESTION FOUR (20 Marks) – Decision Trees and Random Forests

4.1 Implement a decision tree (using the default parameters) to predict Class using the imputed data (note that you want to predict the response as a "class") and compare this to the logistic model. Include in your writeup the function (code) that does the decision tree and a boxplot showing the resulting error (x100 runs) and compare this to the logistic regression result. You should also compare the 2 distributions and the mean/sd as further evidence of their similarity/difference.

(10 marks)

4.2 Examine the provided code for a random forest in rf.R– This is a simple implementation: the *predict.rf(..)* function takes an imputed list of datasets, and builds a random forest and does the prediction and returns the accuracy measure (for direct comparison with previous models).

Give a brief description of how the random forest is created and tested.

Run the supplied code and compare the resulting accuracy against the logistic model and single decision tree. Use replicate to run *predict.rf(…)* 100 times with ntrees=1, 5, 10 and 20. Leave the other parameters at their default value. Does the performance improve? Briefly explain why this model seems superior (does it?) to logistic regression and a single decision tree.

(10 marks)

QUESTION FIVE (15 MARKS) – Write approx. 300 words

Assume you are given a dataset from a loans company that describes the following characteristics of customers including whether they were found to be a good or bad credit risk (the response):

Age of customer;
Experience: professional experience in years;
Income of customer;
Average monthly credit card spending;
Mortgage: size of mortgage;
Credit Card: No/Yes;
Educational level: three categories (undergraduate, graduate, professional);
Credit Risk: Good/Bad

You are ultimately interested in building a model that predicts Credit Risk given the other independent (explanatory) variables.

Explain the steps that you would follow to understand the patterns in the dataset, PRIOR to building a model for prediction. Ensure in your description that you include comments on what information each method or visualisation gives, and why this may be important prior to modelling the data.

Submit a pdf with the associated code, figures and supporting discussion to blackboard by the due date.