Coláiste na Tríonóide, Baile Átha Cliath
Trinity College Dublin
Ollscoil Átha Cliath | The University of Dublin

Faculty of Engineering, Mathematics and Science

School of Computer Science & Statistics

BA (Mod) Management Science & Information Systems Studies **Trinity Term 2017**
Senior Sophister Annual Examination

Data Analytics

Wednesday 17th May 2017 **Sports Centre** **14.00-17.00**

Professor Myra O' Regan

Instructions to Candidates:

Answer three questions. All questions carry equal marks. Each question is marked out of 50.

You may not start this examination until you are instructed to do so by the invigilator.

Non-programmable calculators are permitted for this examination – please indicate the make and model of your calculator on each answer book used.

Q1a: i) How do trees deal with missing values? **(4 marks)**

ii) How do trees deal with outliers? **(4 marks)**

iii) How do trees deal with linear relationships? **(4 marks)**

iv) If two input variables are highly correlated how do trees deal with this situation. **(3 marks)**

v) How do trees deal with a variable where all values of the variable are the same? **(3 marks)**

vi) How does Rulefit incorporate tree output into its analysis?

(8 marks)

Q1b Discuss the role of pruning in the context of growing a single tree model.

(10 marks)

Q1c Compare and contrast trees to logistic regression. **(14 marks)**

Q2a. i) What is an ROC curve? Explain in detail how it is constructed?

(10 marks)

ii) The R package caret prints out the following Specificity, Sensitivity, Accuracy, Kappa for a misclassification table. Explain in detail what each means and how it can be used. **(15 marks)**

Q2b. You have built two models to assess the probability of defaulting on a loan. You need to implement one of the models for use in a bank. Discuss in detail how you would choose between the models. **(10 marks)**

Q2c. Explain how costs and priors can be considered in conjunction with an ROC curve. **(15 marks)**

Q3a. What is an ensemble? Discuss the advantages and disadvantages of ensembles. **(15 marks)**

Q3b. Random Forests, Bagging and Rule Fit are three types of ensembles. Discuss the similarities and differences between them. **(15 marks)**

Q3c. A random forest was run on the Ames housing data. The target variable was a binary variable

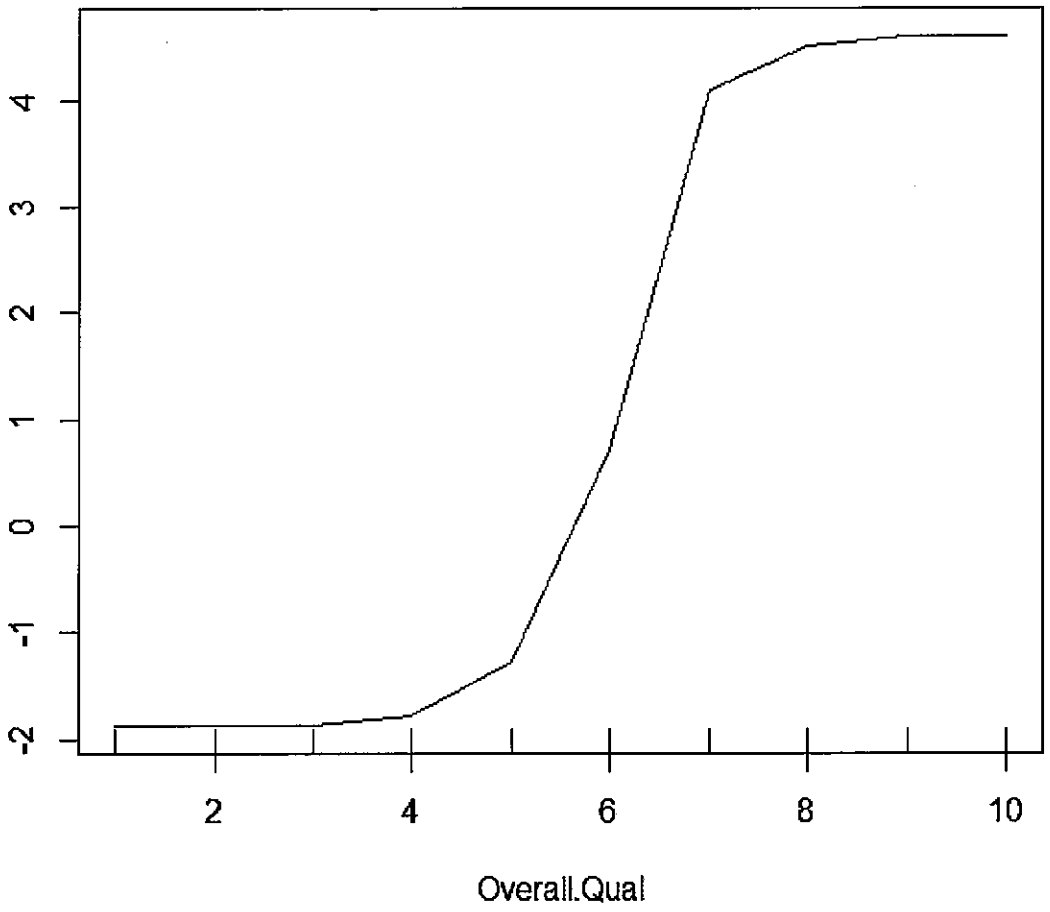
1: High prices houses

0: Low priced houses.

The analysis included the following variables Lot.Area, Lot.Config, Bldg.Type, Overall.Condition, Overall.Quality, Central.Air, TotRms.AbvGrd and TotalARea of house. Overall Quality was coded on a scale of 1 to 10 with 10 representing the highest quality. (It is not necessary to know how the remaining independent variables were coded to answer this part of the question). A partial dependency plot was produced for the variable Overall Quality and is given below:

Continued on next page.....

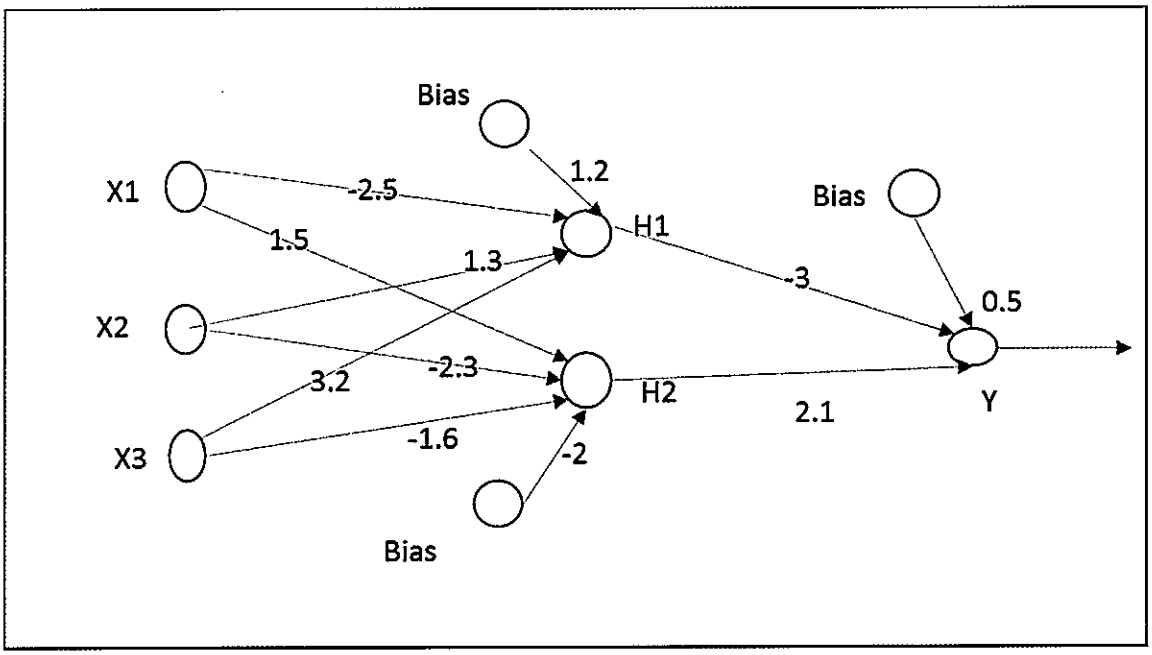
For the high priced houses



i.) Explain how this diagram was produced discussing any potential problems. **(15 marks)**

ii.) Explain what is depicted in diagram. **(5 marks)**

Q4: The following neural network was built to predict an outcome variable Y from three independent variables X1, X2, and X3. Two hidden nodes were used H1 and H2. The activation function g(a)=$\frac{1}{1+e^{-a}}$ was employed at the hidden nodes. At the output node a linear function was used. The following neural network was built



a) Show how to use the above network to predict a new value when X1=0, X2=2 and X3=-1. **(10 marks)**

b) What is the difference between the above neural net and the multiple regression model Y=$\alpha + \beta_1 X1 + \beta_2 X2 + \beta_3 X3 + \varepsilon$? **(10 marks)**

c) In what situations would you expect a neural net to outperform the multiple regression model? **(10 marks)**

d) You have been given a very unbalanced dataset with regard to the binary dependent variable. Discuss some methods of dealing with this situation. **(10 marks)**

e) Discuss the role of background knowledge of a problem in building a model. **(10 marks)**