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

MSc Computer Science – Data Science
Annual Examination

Hilary Term 2018

Data Analytics

Tuesday 9th January 2018 Exam Hall 14:00-17:00

Professor Myra O'Regan

Answer all questions.

Materials permitted for this examination:

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

Question 1:

a) What is a regression tree? 5 marks

A regression tree was fit to the Ames housing data. Some of the results are reproduced below:

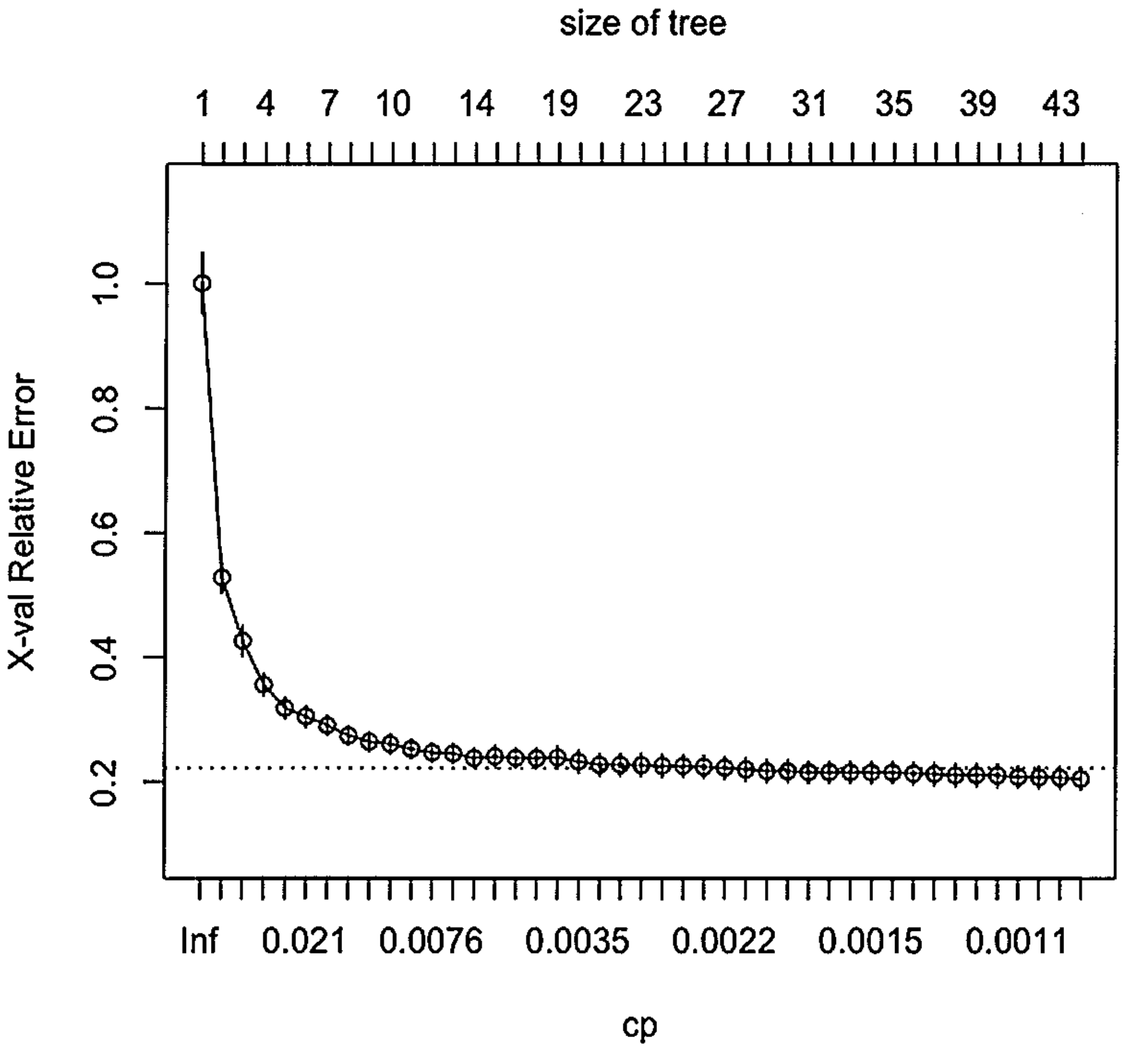
```
Node number 1: 2930 observations,    complexity param=0.4728876
  mean=180796.1, MSE=6.379705e+09
  left son=2 (2442 obs) right son=3 (488 obs)
  Primary splits:
     OverallQual < 7.5     to the left,  improve=0.4728876, (0 missing)
     TotSF       < 1490.5  to the left,  improve=0.3279457, (0 missing)
     TotalBsmtSF < 1388.5  to the left,  improve=0.3226935, (1 missing)

  Surrogate splits:
     GarageArea  < 690.5   to the left,  agree=0.880, adj=0.279, (0 split)
     TotalBsmtSF < 1562.5  to the left,  agree=0.867, adj=0.203, (0 split)
```

OverallQual:	Overall Quality of the house
TotSF:	Total Square Feet
TotalBsmtSF	Total Basement Square Feet
Garage Area	Area of Garage in square feet

b) Explain the following terms from the above output: complexity parameter, MSE, improve, agree and adj. 15 marks

Question continued overleaf.........

c) The following graph was also given. Explain how it is built and used in growing a regression tree. 10 marks



d) Discuss the differences between regression trees and classical multiple linear regression. 10 marks

e) Explain how trees are employed in the RuleFit ensemble method. 10 marks

Question 2:

a) What is a ROC curve? 5 marks

b) Draw a ROC curve for the data in the following table. 15 marks

Predicted probability of an event	Target variable (1==event)
0.5	0
0.0	1
0.3	1
0.0	0
0.3	0
0.0	0
0.8	1
0.5	0
0.3	0
0.5	1

c) How can costs and priors be viewed in conjunction with ROC curves? 15 marks

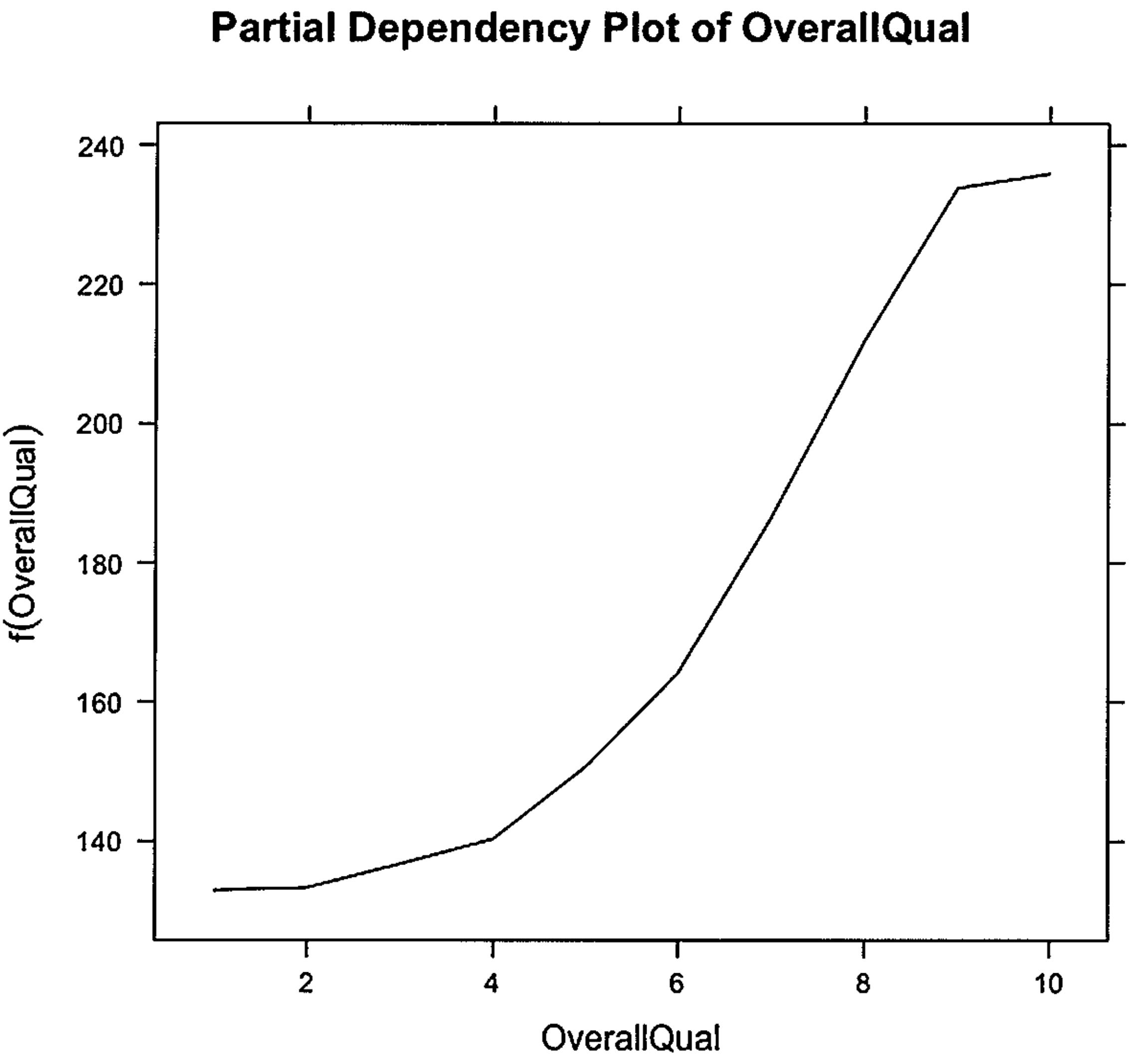
d) You have been given a task to build a model to predict whether a new customer is likely to default on a loan. You have built two models for this purpose. Describe how you would choose between your two models. 15 marks

Question 3:

a) What is an ensemble? 5 marks

b) Explain in detail the differences and similarities of the following ensemble techniques: Random Forests, Bagging and Stacking. 15 marks

c) The following is a partial dependency plot using the Ames Housing data and a random forest model.



i) Explain what is depicted in the plot.

ii) Explain how the plot was constructed?

iii) Explain the difference between partial dependency plots and ICE plots. 15 marks

Question continued overleaf........

d) Explain how the gradient boosting method works. What parameters need to be set?

15 marks