Solutions Online Test 4

Instructions: The test is worth up to 5 points. Each correct response is worth 1 point. There is a bonus question. There are no R files associated with this assignment.

1. In order to develop a data mining process, after we decompose the business problem, one piece requires the estimation of conditional probabilities in a binary classification problem. Which of the following data mining methods and performance metric combination would you recommend?
   a) Cluster and Accuracy
   b) Quantile regression and Profits
   c) Linear regression and Mean squared error
   d) ✓Logistic regression and Binomial deviance

Option d) is correct. Since the estimation that is required is of conditional probabilities of a binary response variable, logistic regression is the appropriate choice. Binomial deviance is an appropriate criterion for probability estimation.

Clustering does not estimate probabilities. Similarly, quantile and linear regression do not estimate conditional probabilities.

2. Consider the following statements regarding the use of baselines in data science projects:
   (i) To anchor the audience to allow a better communication of performance.
   (ii) To help the data scientist understand the magnitude of improvements in performance.
   (iii) To communicate to stakeholders that mining the data has added value.

   Which of the following contains all the correct statements?
   a) None of the statements apply to baselines.
   b) (i) and (ii) only.
   c) (ii) and (iii) only.
   d) ✓ (i), (ii) and (iii).

Option (i) is true is a key use of baselines. The audience might not be able to grasp the quality of the (raw) performance of the proposed method. A baseline is an understandable policy which the audience is typically familiar with. Thus improving upon it provides a clear message.
Option (ii) is true as in many complex settings it is not clear at all what is achievable. Beating baselines provides confidence that we are not overfitting the data or falling into some pitfalls.
Option (iii) is true for similar reasons to the option (i).

3. In online advertisement the base rate of response to an advertisement may be very small. In fact one in a million is not unusual. Modelers typically would prefer to avoid handling data sets with millions of non-responders for every responder. Therefore they down-sample the non-responders to create a more balanced dataset for modeling and evaluation. Which of the following performance metrics will be more appropriate?
   a) Profit curve
   b) ✓ROC
   c) Cumulative response curve
   d) Lift curve

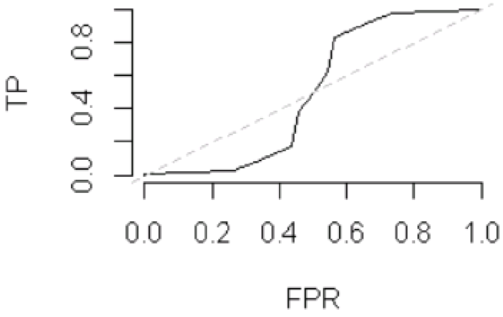
   Because the imbalance between positive and negative responses in the population, it is important to report metrics that do not depend on the population. ROC relies on TPR and FPR which are insensitive to the population. In contrast, all the others depend highly on the balance of the population (and since we down-sample non-responders, the sample we have is not appropriate).

4. Consider the Churn problem in which we can compute reliable estimates of the CLV of customers, of the probability of churn, and the promotion costs. Further, the class priors for available data is believed to be similar to the class priors where we intend to apply the data mining solutions. Which of the following is visualization is more appropriated for stakeholders?
   a) ✓Profit Curve
   b) ROC
   c) Cumulative Response curve
   d) Lift Curve

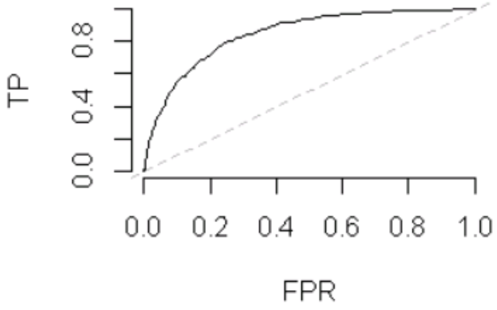
Because the class priors for the available data is the same to the data, using metrics that are affected by the class priors is appropriate. This is the case of Profit curves, cumulative response curve and lift curve. However Profit curve is likely to the curve that is more appropriate for other stakeholders (that might not be as familiar as you with data mining tools).

5. Which of the ROC curves has the worst performance? (All four are different.)

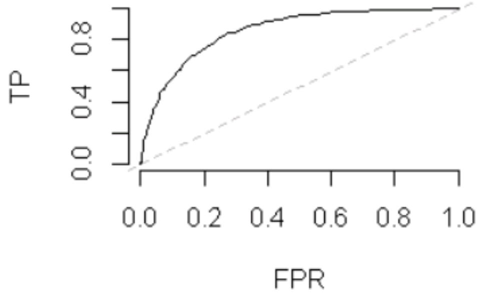
(a) ✓Classification tree



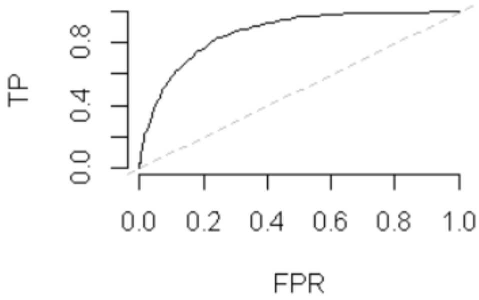
(b)



(c)



(d)



The worst performance is of the curve with lower TPR (true positive rate) and larger FPR (false positive rate). If a curve is below the other it is worse. We see that curve (a) is below the others so (a) is the correct response.

6. Area under the curve of the ROC curve provides an overall measure of performance of the classifier. Which of the following is not a property of the AUC measure?
   a) It is a simple summary of performance between 0 and 1.
   b) It is independent of class prior (i.e. balance of positive and negative).
   c) It is independent of cost-benefits.
   d) ✓It is at least .5 since that is the performance of random guessing.

Note that a method that is worse than random guessing (for example if you overfit the data) the AUC can be less than 0.5. Thus letter (d) is not a property of the AUC measure.