Requirements in Assignment 5 Question 2

[10% pts can be deducted for poor presentation in the report]

There are two parts will be considered in the submitted report:

A: Answers to questions

- Without all details on what you did.
- Summary of results, as necessary to justify your response to specific questions.

Reference to detailed results in part B, as needed.

B: Details of what you did (details on experiments and results)

- Tables can be very useful in showing, for example, different experimental settings and results. Also, graphs on performance with different parameters. Make sure graphs are labeled. Tables need proper headings.
- Only graph/table is pointless – these are there to help make a point. So you need to explain/describe.
- Confusion matrix should be clear on "actual/predicted"
- Settings of the used R models should be clear.

Questions:

1. Explore the data: What is the proportion of "Good" to "Bad" cases? Obtain descriptions of the predictor (independent) variables – mean, standard deviations, etc. for real-values attributes, frequencies of different category values. Anything noteworthy in the data?

- How did you address this question? What did you do?
- Meta-data description: mean, stdev, range; counts in categories. Any missing values?
- For some (?) variables: Frequency of different classes (independent / target/ label variable).
- Number of categorical and numeric variables.
- Does #-of-categories or range vary widely?

- Anything noteworthy? Eg. Specific category in Var x holds large proportion of +ve class?
- Did distribution of values in certain variable(s) catch your attention? If so, why?
- Anything else comes to your mind?

2. We will first focus on a descriptive model – i.e. assume we are not interested in prediction. Develop a decision tree on the full data. Which variables are used to differentiate "good" from "bad" cases? What levels of accuracy/error are obtained? What is the accuracy for the "good" and "bad" cases? Do you think this is a reliable (robust?) description?

- What parameters did you use for building the tree? (Why?)
- Variables that are useful to differentiating classes.
- Accuracy of model? Acc. of different classes? False positives? … Confusion matrix.
- Also, provide your interpretation of these results.
- Reliable model? Why?

3. Next consider developing a model for prediction. For this, we should divide the data into Training and Validation sets.
Consider a partition of the data into 50% for Training and 50% for Test. What model performance do you obtain? Is the model reliable (why or why not)?
Consider partitions of the data into 70% for Training and 30% for Test, and 80% for Training and 20% for Test and report on model and performance comparisons. Feel free to experiment with other size partitions on the data. Is there any specific model you would prefer for implementation?

- In developing the models above, change some of the decision tree options and see if and how they affect performance (for example, the minimum number of cases at a leaf node, the split criteria). Also, does pruning give a better model please explain why or why not?
- What did you do?
- What parameters? (Why?)
- Summarize performance. [Table?]

(Only key results to justify your response to questions here, details in part B) Which is your preferred model, why? Try out different Decision Tree nodes – CART, C5.0 etc.)

4. In implementation, the model will be used to identify "good" vs. "bad" credit applicants. It is conceivable that certain applications will not carry reliable data on the different attributes (data on some attributes may be missing for the case you want to predict) used in the model. Will the decision tree model still be usable? How do you propose to use the model in such cases?

- Justify your answer.

5. Use the misclassification costs in obtaining a model. Do you observe any changes in the model and /or performance? Are there any benefits from specifying misclassification costs?

- Justify your answer.
- Please explain what you did. [Table?]

6. Change some of the decision tree options and try the above. Also, see what happens when Pruning is set on.

- Explain all your settings.
- Summarize your observations.
- Do you see any differences? Explain.
- Accuracy?

7. What are the best nodes for classifying "Good" applicants? Output rules corresponding to these.

- Explain your choice.
- Write down all the rules.
- How do you evaluate each rule?
- Which rule is strongest in your opinion?

Report

We have individual styles and preferences – rather than a prescribed 'best' way to present, each team should submit a report that they think helps answer the questions (would you like to receive such a report? would you submit it to your biz 'boss'?)