**Credit Rating Analysis**

**Mini-Project 1**

A company wants us to help them use analytics to create a model predicting its customers’ ***credit rating*** as a function of information it has about them. The data is in “MiniProject1.csv”. The independent variables are in column B through R. The goal is to predict the credit rating in Column S.

Credit rating can take values 0 (did not finish paying), 1 (may have gotten into default or finished paying late), 2 (sometimes late, maybe finished nearly in time but didn’t do as many payments as they should have), 3 (paid early or on time). When a customer took out multiple loans and had different credit ratings, the cell shows the average of the credit ratings he received by the company.

Because the data set is small, we will not attempt to predict each category. Instead you first have to group the categories together and define what should be 0 and what should be 1, based on the number of observations you have of each.

The objective of this project is to use graphical tools (plots), logistic regression and classification trees to attempt to build a predictive model for this company. The logistic regression and classification tree must be done in R.

**Deliverables:**

* one typed report in Microsoft Word in the following format (imagine that the company executives will read it, focus on what is useful for them):
  + ***Cover page:*** title and names of all students in the team.
  + ***Executive summary*** (on its own page): once you have completed your analysis, summarize what you did and your conclusion in one paragraph.
  + ***Table of contents*** (on its own page, auto-generated by Word [once you call the proper command], make sure to use Headings in your report so that Word can combine everything in one professional-looking table of contents – if you have no clue what I’m talking about, let me know and I’ll demonstrate in class)
  + ***List of tables and list of figures*** (on their own pages, also auto-generated by Word [once you call the proper command] when you use the caption feature in Word for your tables and figures – again, let me know if you have no clue what I’m talking about)
  + ***A Preliminaries section*** describing the problem and basic analysis with study of correlation among independent variables and graphs explaining how you grouped the credit rating categories together to make your predictions (for instance your 1 could be their 3, or their 2& 3, or their 1&2&3).
    - Note: you must have at least one graph of each type we have seen in class (scatterplot, histogram and box plot)
    - You may use this section to argue that some outliers should be removed from the data set before creating the models. If so, explain why.
    - You may use this section to argue that some independent variables should be removed from the analysis.
  + A section implementing ***logistic regression***.
    - You must split your data set between a training set and a testing set.
    - You must (attempt to) refine your model if some of the variables are not significant.
    - Compute accuracy on testing set. Compare with accuracy of baseline method (where we predict most likely outcome for everyone).
    - Explain your results.
  + A section implementing ***classification trees***.
    - You must split your data set between a training and testing set.
    - You must investigate the impact of minbucket on your tree.
    - You must do cross-validation.
    - Compare accuracy on testing set with baseline method.
    - Explain your results.
  + A section called ***Conclusions***, where you select a model among those above and explain why you recommend that one. Or, if you want to argue the company should use logistic regression and classification trees together, explain how they should do that.
  + All pages should be numbered.
  + The report should be double-spaced in Times New Roman 12pt with 1 inch margin all around.
* all software files (R scripts).

***Grading Scheme***: 15 points total

* 3 points for the preliminary analysis
* 5.5 points for the logistic regression
* 5.5 points for the classification trees
* 1 point for the overall quality of the written report and strength of final recommendation.