Predictive Modeling: Project 2

Due: Wednesday, April 27 at 5pm

The data contained in the file `project2.data` is an anonymized dataset corresponding to predicting loan defaults in a bank. The data has been anonymized in two ways. Firstly, true variable names were removed and replaced by generic labels `x1` through `x5`. Secondly, variables were re-scaled to have different means and variances from the original variables. Neither of these actions affect your ability to build a prediction model using the data: you are still able to identify important variables and assess model accuracy using appropriate metrics.

The bank has asked you to look at this dataset and make a recommendation for an appropriate prediction model to use. In the sample, the outcome variable y represents accounts that have defaulted in the last 3 months ($y = 1$) and all other sampled accounts ($y = 0$). Assume that this is a representative sample of accounts held by this bank.
In your project, consider the following guidelines when developing appropriate prediction :

1. Some of the variables contain missing values (`NA` in `R`). Identify the variables that contain missing values. Remove all cases with missing observations from the dataset. How many cases are excluded?

2. After removing all cases with missing values, create a random partitioning your data. Keep 70% of your data as a training set and 30% of your data as a validation set. Be sure to specify a random seed so that your results are reproducible.

3. Using the training data, create data visualizations exploring the relationships between the five numeric input variables and the binary output variable. Also calculate the means and standard deviations of each variable in each sub-group (i.e. separately for successes and failures). Discuss differences and similarities between the groups and comment on how this might translate into effective/ineffective prediction modeling.

4. **Group A1**: Consider two prediction models – a logistic regression model with linear terms only and a logistic regression model with linear and quadratic terms. Use 8-fold cross-validation applied to the training data with the overall mis-classification rate as criterion to choose between the two models.

   Group A2: Consider two prediction models – a logistic regression model with linear terms only and a quadratic discriminant analysis model. Use 8-fold cross-validation applied to the training data with the overall mis-classification rate as criterion to choose between the two models.

   Group B1: Consider two prediction models – a linear discriminant analysis model and a logistic regression model with linear and quadratic terms. Use 12-fold cross-validation applied to the training data with the overall mis-classification rate as criterion to choose between the two models.

Group B2: Consider two prediction models – a linear discriminant analysis model and quadratic discriminant analysis model. Use 12-fold cross-validation applied to the training data with the overall mis-classification rate as criterion to choose between the two models.

All: Report the cross-validation scores for both of the models and make a recommendation as to the preferred model. Discuss why this model appears most appropriate.

5. Fit the selected model to the full training data. Then, using this estimated model, evaluate the predicted values in the validation data. Also calculate the overall misclassification, the false positive rate, and false negative rate of your model in the validation data. Comment on whether the false positive or false negative rates should be of greater concern to the bank.

Your project should be submitted in written report form (3-4 pages including plots) with selected `R` code chunks and analysis output included. Your main report need not include all code/output, but if not you should also submit a supplementary file with all your `R` code. The allowed submission formats for your report are a pdf generated in RMarkdown, a pdf generated in LaTeX, or a Word document saved as a pdf.