IDS 572 Assignment 2 – Models for investment decisions in LendingClub loans

Due date: March 2nd

This is a continuation of the previous assignment where you developed tree based models to predict "Fully Paid" vs "Charged Off" loans in the Lending Club platform. This assignment will consider in more detail the question of developing predictive models for effective investment decisions. We will also look into balancing the training data, parameter tuning, and reliable performance estimates through resampling and cross-validation.

1. (a) Develop boosted tree models (using either gbm or xgBoost) to predict loan_status. Experiment with different parameters using a grid of parameter values. Use cross-validation. Explain the rationale for your experimentation. How does performance vary with parameters, and which parameter setting you use for the 'best' model.

Model performance should be evaluated through use of same set of criteria as for the earlier models - confusion matrix based, ROC analyses and AUC, cost-based performance.

Provide a table with comparative evaluation of all the best models from each method (decision trees and random forests from the earlier assignment, and boosted trees); show their ROC curves in a combined plot. Also provide profit-curves and 'best' profit' and associated cutoff values. At the respective best cutoff levels, what are the accuracy values for the different models?

2. (a) Develop linear (glm) models to predict loan_status. Experiment with different parameter values, and identify which gives 'best' performance. Use cross-validation. Describe how you determine 'best' performance.
How do you handle variable selection?
Experiment with Ridge and Lasso, and show how you vary these parameters, and what performance is observed.

(b) For the linear model, what is the loss function, and link function you use ?
(Write the expression for these, and briefly describe).

(c) Compare performance of models with that of random forests (from last assignment) and gradient boosted tree models.

(d) Examine which variables are found to be important by the best models from the different methods, and comment on similarities, difference. What do you conclude?

(e) In developing models above, do you find larger training samples to give better models ? Do you find balancing the training data examples across classes to give better models ?

3. Develop models to identify loans *which provide the best returns*. Explain how you define returns? Does it include Lending Club's service costs?

Develop glm, rf, gbm/xgb models for this. Show how you systematically experiment with different parameters to find the best models. Compare model performance – explain what performance criteria do you use, and why.

4. Considering results from Questions 1 and 2 above – that is, considering the best model for predicting loan-status and that for predicting loan returns -- how would you select loans for investment? There can be multiple approaches for *combining information from the two models* - describe your approach, and show performance. How does performance here compare with use of single models?

5. As seen in data summaries and your work in the first assignment, higher grade loans are less likely to default, but also carry lower interest rates; many lower grad loans are fully paid, and these can yield higher returns. One approach may be to focus on lower grade loans (C and below), and try to identify those which are likely to be paid off. *Develop models from the data on lower grade loans*, and check if this can provide an effective investment approach – for this, you can use one of the methods (glm, rf, or gbm/xgb) which you find to give superior performance from earlier questions.

Can this provide a useful approach for investment? Compare performance with that in Question 4.

6. Considering all your results, which approach(s) would you recommend for investing in LC loans? Explain your rationale.

Please submit a pdf file with answers to the assignment questions, and supporting analyses. Also include a single Rmd file with your R code (note – code needs to be adequately commented and divided into sections in the Rmd file to help readability and ease understanding by others).