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

Due date: Oct 11th

This is a continuation of the previous assignment where you developed decision tree based models to predict "Fully Paid" vs "Charged Off" loans in the Lending Club platform. In this second assignment, you will develop additional models – GBM, GLM (XGB) - to predict which loans are likely to be paid off and which will default. The previous assignment ended with the question on effective investment decisions based on your predictive models – we will examine this in more detail in the second assignment. We will also focus on parameter tuning, and reliable performance estimates through resampling and cross-validation.

1. (a1) Develop gradient boosted models to predict loan_status. Experiment with different parameter values, and identify which gives 'best' performance. How do you determine 'best' performance?

(a2) For the gbm model you develop, what is the loss function, and corresponding gradient in the method you use? (Write the expression for these, and briefly describe).

(b1) Develop linear (glm) models to predict loan_status. Experiment with different parameter values, and identify which gives 'best' performance. How do 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.

(b2) 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 (which you did in your last assignment).

(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?

2. 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. Do you find larger training sets to give better models ?

3. 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?

4. 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. Compare performance of models from different methods (glm, gbm, rf).

Can this provide a useful approach for investment? Compare performance with that in Question 3.

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).