Overview

You are meant to carry out substantial data analysis for the final competition. Here are requirements:

- Choose some methods we have covered in the course which you can argue are appropriate for this prediction problem.
- Apply the methods to the prediction problem, including the selection of suitable model complexity on validation data. Assess the performance of the resulting prediction rule(s) and understand their behavior.

Your grade for the final project will be based on a written report. In your report, you should do the following:

- Provide a short background discussion of the contest problem domain as you understand it.
- Provide a more detailed explanation of this particular prediction problem.
- Give an overview of the data set: what is the response, where do the predictor variables come from, and so forth.
- Present the results of your statistical analyses in detail, using text, figures, and tables, and including initial summaries and visualization of the data. The presentation should focus on the prediction rules you obtained, using methods chosen from the course. Explain why you chose those methods over others.
- Provide suitable concluding remarks, such as discussing things you thought of trying but didn't (and why you didn't).

Your report will be graded on the clarity of your exposition, the quality of your analyses, and the extent to which you met the requirements above.

Collaboration

As with the problem sets, you should work on and submit the final report with your group. Feel free to discuss the contest and your approach with other students in the course as well – this is likely to improve your competition standing and your final report. Each person must participate in both the data analysis and the report writing. The report must include an attribution section indicating who did what.

Relevant R packages

Here is a list of some of the methods we have discussed in the course, along with the names of R packages and functions implementing them. You may find other packages/functions implementing these same methods; feel free to use whatever you argue is most appropriate. Some

of the packages mentioned below are distributed with R. Others you would need to install using `install.packages()`.

- *k*-nearest-neighbors: package `class`; function `knn`
- linear regression: function `lm`
- linear/quadratic discriminant analysis: package `MASS`; functions `lda`, `qda`
- logistic regression: function `glm`
- ridge regression: package `MASS`; function `lm.ridge`
- subset-selection regression: package `leaps`; function `leaps`
- the lasso: package `lars`, function `lars`
- additive models using splines: package `gam`; function `gam`
- classification and regression trees: package `rpart`; function `rpart`
- boosted decision trees (classification and regression): package `gbm`, function `gbm`
- support vector machines (classification and regression): package `e1071`, function `svm`; also see package `kernlab`