STAT 696: Statistical Communication in Data Science Professor Levine

Final Project

Paper, due Thursday December 20; submit on Blackboard

Consider the `Hitters` data set in the `ISLR` R package where we are trying to predict baseball player salaries off a variety of explanatory variables. Consider three statistical learning approaches to accomplishing this task: regression, lasso, and bagging (bootstrap aggregation). You are encouraged to resort to the course e-textbook and browse the web to find background on the data set and these statistical learning methods.

Do note that the deliverable is a data analysis report. Make sure then to present exploratory data analysis to justify how you apply each learner (particularly the regression model), a brief methodological comparison of the statistical learning methods, and an evaluation of predictions made relevant to the sports analytics problem at hand. As always, your report is restricted to 7 pages including all tables and figures, but excluding appendices and references. The report must be compiled in LaTeX using the `documentclass{article}` with 11 point font, one inch margins, and single spacing. Please see the course Blackboard site for the LaTeX preface. Please see the document "Generic Outline for Data Analysis Report" on the course Blackboard site for organization of the report. The audience of your paper are statisticians and statistically-savvy scientists. Statistical results should be appropriately interpreted and put within scientific context throughout. Do not present *R* output dumps; instead present results in well formatted tables and figures that allow you to effectively communicate your findings and aid the reader in interpreting your text. All figures and tables presented in the main document must be discussed in the main text.

Your report must include an appendix of code and supplemental material. The code should allow another student in the class to repeat the analysis you performed. The code must then include appropriate documentation and output for the user to understand your programming and perform quality control checks. The supplemental material may contain well labeled, concise, and neatly presented EDA and analysis results (i.e., tables and figures) that did not fit in the main document. Though the supplemental material should not be a critical part of the discussion and argument in the main text, it may support analysis decisions, findings, and conclusions presented in the main text.

The course Blackboard site presents the grading rubric for the writing project.