Final Report for the Final Project in OIS 6850

1 Introduction

The final report is a required element of the Final Project. Your objective should be to fit a full model to the data, based on what you learned fitting the parsimonious, five predictor model for the interim report, and communicate your results in a thorough report. You should introduce the problem, describe the data and your cleaning procedures, along with any variable creation/feature engineering you did, and then describe your model and its performance. Along with our report please submit a .csv file of your predictions for every row in the test set. Please also submit your .csv file to Kaggle, and report your returned RMSE and leaderboard ranking.

2 The data

https://www.kaggle.com/c/house-prices-advanced-regression-techniques

The data has been split into 50% train and 50% test sets at the Kaggle website (with 1460 and 1459 observations, respectively). The test set contains all the predictor variables found in the train set, but is missing the outcome variable, SalePrice.

3 Details

The model you use for the final report should predict SalePrice using the predictors you deem optimal—"optimal," of course, defined not as in-sample performance but out-of-sample performance. The challenge is that you will not be able to measure out of sample performance directly but must rather estimate it using the train set.

The length of your final report will likely depend on how many plots and tables you choose to include. However, your report should be at least 5 pages long.

You should strive to write a report that is polished and client-ready—consisting in a carefully formatted PDF document—but which is also reproducible. So, along with your PDF, turn in the .Rmd file that produced the PDF. Your plots and tables should be accompanied by explanatory captions. Take care not to use the report as simply a dump for all the plots and tables you can think to include. Plots should have labelled axes, and it should be clear why you've included any given plot or table and how it advances the main point you are making. Review the slides for class 9 for best practices in report writing. You should, for example, think about how to use white space to increase the readability of your document.

Please include the names of your team members on the final report.

You may structure your report any way you like. The rubric I'll be using for assessment is the same one I used for the interim report; use that to guide your efforts.

- **Due date: Thursday, November 3 by midnight.** However, note that I can be flexible on this due date. If you will be later than November 3, please let me know.

- **Grading: 80% for the final report (or 24% of the overall grade).**

4 Grading rubric

You will be scored on the following elements, with possible scores of poor, fair, good and excellent.

- *Introduction.* What is the problem you are working on? Describe the Kaggle train/test format, and your ultimate objective in the project.

- *Data modeling /Cleaning.* What sort of data modeling and cleaning decisions did you make? Did you create any variables? Did you refactor any variables? How did you deal with NAs?

- *Model /Model development.* Describe your model. Which variables did you end up using in your model, and why? How did you choose them? Which statistical method(s) did you choose (e.g., linear model, random forest, etc.), and why?

- *Model performance.* How did your model perform on the train set? What is your estimate of the model's performance on the test set? If you have submitted your predictions to Kaggle, what is the returned RMSE of your submission? What is your leaderboard ranking?

- *Statistical communication.* Are your figures labelled clearly, accompanied by explanatory captions, and referenced in the text? Likewise, are your tables titled and/or captioned appropriately and referenced in the text? Please do include a table presenting your model coefficients/variable importance but also translate coefficients into quantities that are easily interpretable.

- *Overall quality.* Are you writing in full sentences? Are there many grammatical or spelling errors? Are you using the .Rmd or .Rnw formats correctly so that your final document is attractively formatted? Does the report look sharp, or are there issues with the compile from .Rmd such that there is lots of junk code getting printed to the screen?

- *Leaderboard rating.* How good is your model compared to those that others reported. Please report the RMSE returned by Kaggle when you submit your .csv file.