STAT 444 Spring 2019
Final Project

Timeline

- For **graduate students**, the final project will be an individual project. For **undergraduate students**, groups of three students should be formed before the reading week. An initial topic should be suggested by June 26 and be approved by the instructor by July 2.
- **Until 11:59 pm on July 31th:** to upload your final project report in a .pdf format and other relevant files together in a zip file to Learn Dropbox Project folder. It should be called "Teamname_project.zip". For complete information about files, see the last section of this document. The title page of your project report should contain the course name, the current term, topic of your project, team name, and names of the students in the team.

  Only team representatives should upload the files.

  If you make more than one submission before the deadline, the most recent one will be considered. No (re)submissions, under any circumstances, will be accepted after the deadline.

Written Report

Page limit 10–40 pages, excluding title page, appendix and code.
Your report should have but not limited to the following sections and information:

- **Motivation and introduction of the problem:** Start with the problem. Describe the general aspects of your project, and why it matters to do what you have done in your project.
- **Data:** Introduce your data, and where you have gotten it from, or how you have collected it (in the case that you have collected your own data). Perform some descriptive analysis on your data including, but not limited to, the plot of the data. You can provide tables, or graphs to familiarize the reader with your data.
- **Preprocessing (optional):** This is the place to discuss missing data, outliers, and/or any problems existing in your data. If you preform any feature reduction and/or transformation and/or imputation, detail your solution/steps here. Some of the aspects, for example, outlier detection and handling can be postponed till later modeling steps.
- **Smoothing methods:** Use smoothing methods (e.g. spline or local regression) to model your data. Describe how you test/model interaction effects if appropriate, and report your prediction error, which could be based on leave-one-out CV, GCV, or 5-fold CV.
- **Radom Forests:** Use random forests models on your data, report the importance of variables and your prediction error. If you use CV to estimate prediction error, use 5-fold.
- **Boosting:** Use boosting methods on your data, report the importance of variables and your prediction error. If you use CV to estimate prediction error, use 5-fold.

- **Statistical Conclusions:** Compare your smoothing, random forests and boosting models or any other method you have attempted, and pick the best candidate with respect to some criteria (prediction error, ease of use, computation time, etc.). This is where you provide your statistical conclusion on the models.
- **Conclusions in the context of the problem:** Provide your findings in the context of the project. For example, highlight the implications/interpretations that you have come to during your modelling/prediction process in the context of the project. You should avoid technical (mathematical/statistical) language as much as possible in this part of the project. This is the section in reports which is usually read carefully by managers, who may not have any statistical background.
- **Future work:** Any aspect of the project you wish could/have been done better. Any weakness of the current methods you tried that you wish there are improvements. Any other statistical/machine learning methods you wish to learn/try in the future. Any data you wish have been collected or to collect in the future to strengthen your model's prediction accuracy for the problem.
- **Contribution:** explain which team members were responsible for which tasks in the project. The tasks can include but not limited to: data collection, data cleaning, analysis, report writing.

Alternative Options for Written Report:

If you are interested in ideas in journal papers related to the course, you are also allowed to write a review paper. Additionally, for students who have original ideas on statistical learning, a research essay is acceptable as well. Make sure that the paper or the research topic chosen by your team is approved by me.

Grading of the Projects

General clarity and correctness of writing are required for all three options.

- For the first option, we will grade your reports on the statement of problem and data, statistical analysis performed, and conclusions.
- For the second option, we will be more concerned about whether you can make critical comments on the paper you choose to review.
- For the second option, we will focus on originality of your ideas in your research essay.

Appendix

The appendix should be in the same pdf document as the main report but not counted in the page limit. The appendix will be read by the grading team only if it is necessary, but the Literature section is required if relevant. Other parts are strongly recommended. It could contain

- **Data:** More detailed descriptions about source, how data are collected if applicable, important variables and known relationship among variables, etc.
- **Literature:** i.e., what has been done by others on your dataset prior to your analysis. If the data are download from Kaggle or other repository, there should/may exist analyses done by other teams. Please summarize past results and describe in detail the most relevant ones. You are only responsible for relevant literature up to the end of this February.
- **Modeling details:** It is unlikely all models you have fitted are useful/interesting to report. Some models are only of intermediate values, and some models may need to be checked but turns out to fare worse than the current one. You can detail your model building process here to justify the models you presented in the main text, please cross-reference in the report.

Additional Information

- **R Markdown:** You are encouraged to use R Markdown to write your report. There will be a grading criterion about how reproducible your results are, and using R Markdown is most likely to score full mark on that part after verification. The alternative is to comment your R code diligently to indicate which part corresponds to which model and/or produce which graph, etc.
- **Cleaning data:** You can use Python to clean the data before analysis. You need to explain what cleaning steps have been done to the raw data, also provide the Python code and the raw data to reproduce your cleaning process. This applies to the raw data downloaded directly from internet, say, Kaggle. If you collect your own data, the data may be merged from multiple sources, and the raw data is considered as the starting dataset when you begin to perform preprocessing (transformation, impute missing data) and statistical analysis.
- **Latex:** The R Markdown will create a latex file besides the pdf file if you set "keep_tex: yes" under "output: pdf_document" section. You can modify the latex file for better control of the layout and margin of the document if you prefer.

Project Submission

Again, your team representative should submit the final project report in a .pdf format and other relevant files together in a zip file to Learn Dropbox Project folder. It should be called "Teamname_project.zip". The files should include:

- **Report:** 10–40 pages report plus the appendix called "Teamname_report.pdf".
- **Data** If you collect your own data or made changes to the raw data file before loading into R, please provide the data file(s). Otherwise, link to the data provided in the report or appendix would be enough.
- **Source Code:** If your report and appendix are generated using R Markdown, you can submit the Rmd file(s). Otherwise, R codes for analysis with detailed comments are expected. The R codes can be organized into several files, for example, loading/preprocessing, smoothing, etc. In case of multiple files, write a readme file to describe the function of each file and the

order they should be executed. Other relevant codes should also be included, for example, if you used Python to clean the data.

- **Others:** Any files needed to reproduce your result and report.