What: You'll work in a group of 2-3 other people to prepare an R markdown document on your investigation of an interesting dataset of your choice. Your group will work to build a predictive model for classification or regression.

1. Submit report as an R markdown document. Should be professional, containing all sections listed below. 5+ pages including plots.
2. Class presentation: 5-10 min slides or R markdown presentation

How: Once you form your group, find out if you have any common interests, academic or otherwise—those are a good place to start! A few places to find interesting data sets are listed on the course webpage. If you need some ideas, please feel free to stop by office hours and talk to me! To give ML algorithms the best chance, pick a dataset with at least 10 variables and 500 observations, including both numerical and categorical predictors, and is large enough to perform cross-validation. Do not reuse datasets used in examples / homework in the class.

The lab report should be professional, and include only the plots/code you need to demonstrate your results (no warnings or error messages, don't display the data, no extra plots). This should read like a paper. The lab report should contain the following sections:

1. Abstract: one stand-alone short paragraph indicating the problem of interest, and what you found. For someone to quickly find out what you did and your conclusions without reading the paper.
2. Introduction: A place to introduce the research questions you are exploring, provide any relevant background knowledge on the topic. You should also comment on the data and cite your sources. In practice, research questions are developed first and data to answer them are then acquired by finding an existing dataset or doing an experiment. For a class exercise, they can blend together, as you have to find a dataset that meets our criteria.
3. Exploratory Data Analysis: Here you should provide a statistical overview of the data set including both numerical descriptive statistics and visualizations about the variables of interest. Give information about the data: How many observations? Outliers? Correlations? Anything missing? Do they show anything interesting that is informing your modeling? You will probably make many plots as you work through this, but in your lab report you only need to include the most interesting stats and most informative plots (no more than 5).
4. Methods: This should be a fairly brief section where you describe your model building process: which models you chose to evaluate (there should be at least 3 that you try). Discuss how your models were tuned, and what criteria you chose to pick the best performing one. You should discuss any preprocessing techniques you tried. You should include performance metrics and visualizations (if appropriate) in this section.
5. Results: Here you will discuss the results of the methods you described in the previous part.
6. Conclusion: Summarize what you did, discuss any limitations of the work, and describe questions for future research.

When:

1. A first report proposal will be due on March 17th. This is an R markdown file that, at a minimum, demonstrates that you have access to a dataset, and you have some idea about what you're going to do with it.

2. A rough draft will be due March 27th. The rough draft should have all the parts of the final report, but does not need to be polished or professional. It may have lots of extraneous code and plots, but should have the main idea of what you want to try.
3. Project presentations will be on **Apr 5th**. If you'd like to make changes to your final draft report after getting feedback from the presentation, you will have until **April 6th** to turn in a final draft.

Grading Rubric [100pts]:

1. Report Proposal [5pts]
2. Rough Draft [10pts]
3. Professionalism [30pts]:
   - The lab report contains all sections listed in the outline.
   - The lab report reads like a paper and is neat, organized, and free of spelling mistakes
   - All figures are properly labeled and legible.
   - No extra code or plots (except those needed for your narrative)
   - All deadlines met.
   - **All sources are appropriately cited.**
4. Modeling [30pts]:
   - EDA includes informative numerical and visual descriptions of data
   - Necessary preprocessing is clearly described
   - Hyperparameter tuning is clearly described, including metrics for choosing best model
   - Justification is given for the metrics used to choose the best model.
   - Results are neatly presented
5. Presentation [20pts]:
   - Timing (10-15min)
   - Slides are organized
   -
6. Team Work [10pts]:
   - Did all group members contribute to the project?