Statistical Modeling Project
Math 32, Fall 2019, Prof. Bhat (hbhat@ucmerced.edu)

As described in the full syllabus and in lecture, one of the main components of this course will be this statistical modeling project.

Real Data. One aspect of the project is to work with real data. I have uploaded to CatCourses four real data sets that are formatted to work with R. Detailed descriptions of the data sets are available on our CatCourses site under "projectdata/files/projectdata.pdf".

If you want to use a different data set, i.e., something that is not one of the four choices on CatCourses, then you must receive prior approval from me. I will want to see that you have the entire data set in an R-readable format, that the data presents you with a clear statistical modeling task (regression or classification) such that there are clear metrics to gauge model accuracy, that the data does not lend itself to a trivial model, and that the data does not have a number of pathologies that would tend to cause you trouble and/or distraction.

We have chosen the four data sets on CatCourses because they have great educational value—you can jump in and start building statistical models. Raw data sets that one encounters in the wild often require a ton of *data cleaning* to reshape/transform the raw data into a data set that is suitable for statistical modeling. The last thing we want is for you to spend most/all of your time carrying out this data cleaning, leaving you almost no time to actually build models and play with them. (Note: this painstaking cleaning process has already been carried out for the four data sets on CatCourses.)

Project Proposal. First, you are required to write a proposal of what you plan to do for the project. The proposal should be between one and two pages in length. This is due by 11:59pm on Friday, October 18, 2019—we will set up a CatCourses assignment and allow uploads of type PDF.

The proposal should clearly state which data set you would like to work with—you may choose either (i) exactly one of the four data sets uploaded to CatCourses or (ii) a data set of your own choice. *If you want to use a data set other than one of the four on CatCourses, please read the guidelines above—in particular, your data must be approved in advance by me, before you can proceed.*

The proposal should clearly state what you plan to do, statistically speaking, with the data set:

- Which data set will you analyze?
- Who will you work with? (You are allowed to work solo if you prefer. You are also allowed to form groups; the maximum group size is 3 students.)
- With the data set that you have chosen, what is the research question you seek to answer?

- How will you use statistics and probability to answer the research question? What kinds of models will you build?
- What are some obstacles you envision? How will you overcome these obstacles?
- How will you develop your model(s)?
- How will you test your model(s)?

I want to see evidence that you have thought through the answers to these questions. Be creative and don't hold back!

Project First Draft. You are required to submit your first draft of the project (narrative plus R code—see requirements below under Expectations) by Friday, Nov. 8, 2019, 11:59pm. The TA's and I will then read the project drafts. We will get back to you rapidly (within 1 week) with our detailed comments/suggestions for what needs to be changed/improved for the final draft. All files related to your project must be uploaded to CatCourses—we will create an Assignment for you to upload your materials.

Project Final Drafts. The final draft (narrative plus R code—see requirements below under Expectations) will be due on Friday, Dec. 13, 2019, 11:59 PM PST. **This is the last day of the semester and there will be absolutely no extensions granted!** All files related to your project must be uploaded to CatCourses—we will create an Assignment for you to upload your materials.

Submit Your R Code! For full credit, any R code used to generate results in your project must be submitted. *Note: we will not accept code that must be "copy and pasted" into R in order to run. We will only accept ".R" files that can be "sourced" into R as standalone programs, just like the codes that have been posted on CatCourses all semester long.*

One Submission Per Group. For all of the above submissions, only one submission per group is required.

Office Hours / Email Help Policy. You may bring your laptop to office hours and ask us to take a look at your code. You are also allowed to email us your code as attachments. *But keep in mind one thing: getting your code to work is your job. We will try our best to point you in the right direction, and we will give you working snippets of code, but ultimately we are not going to do all your work for you. Please do not expect otherwise.*

Expectations. Here are some notes about what we expect from the project writeup, both for the first and final drafts.

- All R code that is submitted must be commented well enough that we can understand it. If you write some R code that you consider to be difficult, sneaky, or clever in any way whatsoever, you must comment it.

- R code by itself, even if well-commented, is not sufficient for a first draft. *Any group that submits only R code without an accompanying writeup will not receive any feedback whatsoever.*

- By *project writeup*, we mean a *scientific report*. Basic elements of a scientific report include:

    1. description of the data set
    2. description and motivation of the research question—what question are you trying to answer, and why is that question worth asking?
    3. discussion of statistical methods used to model the data set, together with an explanation of why the particular statistical models were chosen, i.e., why are the methods you chose suitable for the particular data set at hand?
    4. presentation of results obtained by applying the statistical model to the data—how good is the model?
    5. discussion of the results—what insights can you give regarding what worked and what didn't work? Are there particular aspects of the data set that are easy/difficult to model? **What did you learn by doing this project?**

- To clarify, in addition to whatever has been written above, the scientific report must contain descriptions of:

    1. what you did
    2. why you did what you did
    3. how well your model performs, both in absolute and relative terms

- The writeup must include citations of sources used.

- The writeup must be written in English using complete sentences.

- You are encouraged to include tables and figures in your project writeup. All tables and figures must have clearly written captions such that if the reader reads the caption, he/she can understand the gist of the table or figure *without* reading some other part of the project writeup. All figures must have legible labels on the axes.

- **In general, the more work you put into the first draft, the better quality feedback you will receive, enabling you to be better positioned to submit a top-notch final draft.**

Requirements, Ideas, and References.

- Two of the data sets will feature *classification* problems. These are a lot like regression problems; the only difference is that the response variable is discrete rather than continuous. We will discuss methods for dealing with such problems in the coming weeks. In general, if you would like to solve a classification problem, the way you model the data and the way you test your model will be different than if you try to solve a regression problem. This should be noted in your proposal and project drafts.

- All of the data sets contain many predictors. Not all of these variables will help you. Blindly fitting a model that includes all the predictors is totally not allowed. Instead, one of the main goals of this project is to describe to us how you went about intelligently choosing which variables to include in your statistical model. In short, complicated models are not necessarily better!

- Introducing transformations of columns is totally allowed, so long as you can demonstrate that this enables you to develop better (i.e., more predictive) models.

- Reading up on the scientific background that goes into these data sets is allowed—sometimes, there are things that are "well-known" in the scientific literature that are not included in the data set you are examining. For example, it might be the case that your reading will inspire you to include a ratio of variables in your model, and this ratio might be more predictive than any of the original variables.

- Do not forget about the power of exploratory data analysis (EDA). There is a reason we have spent and will spend time on histograms, empirical CDF's, and kernel density estimates—they are really useful tools. We would like to see you use these tools creatively, perhaps to model the predictors before doing regression, or perhaps to figure out if the residuals of the model that you fit are doing what they are supposed to do. **EDA by itself will not make a complete project, but preliminary EDA + predictive modeling + model testing + analysis of residuals will almost surely be enough for a complete project.**

- In terms of models that you are allowed to use: we are aware that certain people in the class have taken courses in artificial intelligence, machine learning, econometrics, and/or other classes that have featured statistical models. In general, you are allowed to use any modeling strategy that you understand and that you can explain to me on a whiteboard. **If you do not understand something well enough to explain the mathematics inside it, do not use it.**

- When you propose a statistical model, if you can write it down in mathematical language using random variables, then that should suggest to you *probabilistic simulations* that you can run in R. Creative use of such simulations is encouraged.

- Finally, if you are the sort of person who wants to get involved in research, this project could be a *beginning*, not just something to be completed. Groups that submit outstanding final projects may be contacted to expand upon their work for presentation/publication.