MS&E 226 *Mini-project: Part 1*
Fundamentals of Data Science

For the first part of the course project there five steps:

a) Register a team (**due October 18, 2019 at 11:59pm**)

b) Choose a dataset, and set aside a holdout set

c) Investigate and explore the dataset

d) Build predictive models

e) Write a report and submit on Gradescope (**due November 13, 2019 at 11:59 PM**)

1 Introduction

This is the first part of the mini-project. You will be applying the concepts you learn in the class, while analyzing a dataset of your choosing. You can choose to do this alone, or in teams of two. (Working together with someone else is *strongly encouraged!*.)

In this first part, the focus is on obtaining and understanding your data, and on prediction. The second part is focused on inference.

2 Register a team

Once you have formed a team and have picked your data set, **register your team** by filling out this form: `https://tinyurl.com/2019-msande226-project`

We use the responses to assign TAs to projects. **In order to have everything run smoothly, the deadline to submit a team at the link above is October 18, 2019 at 11:59pm.**

3 Choosing a dataset

You can either choose a dataset we have selected, or find a dataset of your own. You are strongly encouraged to find your own dataset, ideally in an area you find more interesting and are personally motivated to explore. You have to like your data; you're going to spend a lot of hours staring at it, so you should find it fun and interesting to work with the dataset you've chosen!

If you choose your own dataset, make sure the dataset is rich enough to let you play with it, and see some common phenomena. In other words, it must have at least a few thousand rows ($> 3.5 - 4K$), and at least $15 - 20$ columns. Of course, larger is welcome. Some data sets might have fewer columns but very rich structure and also be viable, in that case, please come talk to one of the TAs.

We have provided two datasets that you could use: data on real estate from Ames, Iowa; and data from the U.S. College Scorecard. Both datasets are available online (in the "Datasets" section of the course site), together with some information about their origin, as well as a data dictionary (that explains what the columns mean).

The following link contains a list of many other datasets that are available through various R packages: `https://vincentarelbundock.github.io/Rdatasets/datasets.html`.

4 Setting aside a holdout set

Before doing anything with your data, *randomly* choose a test set (representing 20% of your rows), and keep it for later. (You don't need to report anything to us for this part.)

You will not touch this test set again until the end of the course! Fix this set from the beginning, and use the remaining 80% for exploration, model selection, and validation.

Note: It may be harder to do this properly with some type of datasets, like time series; if you have selected time series data, let us know and we can help you find a testing strategy. In general, for such data, you want to train on earlier data and test on later data.

5 Analyzing your data and building a predictive model

You will now (1) carry out some data exploration and then (2) build a predictive model. Write a report of **3-4 pages**; your report must be submitted by **November 13, 2019 at 11:59 PM on Gradescope**.

For each of the two components below, we provide some guidance on what we are expecting; however, data analysis is a dynamic process, rather than just ticking of checkboxes. You will be judged on the overall quality of your analysis and presentation, in addition to the whether you satisfy the minimum required steps we outline below.

Be succinct; the goal of the report is to highlight your findings, rather than describe everything you did. **You are welcome to attach appendices containing code, extra plots, etc., but you will be graded on at most the first 4 pages you submit.**

Grading: Out of **10 total points**, **6 points** will be based on whether you satisfy the minimum required steps. The remaining **4 points** will be based on subjective overall quality, calibrated across teaching assistants to ensure consistency.

5.1 Investigating and exploring your data

Once you've chosen your dataset, it is time to explore the data to get a better grasp of the structure, and to find interesting questions to explore.

In the report, you will describe your dataset and interesting findings; the write-up for this component should be **1-2** pages.

Required: Make sure your report addresses the following.

- Describe the dataset you have selected. Explain how the data was collected; Do you have any concerns about the data collection process, or about the completeness and accuracy of the data itself?

  Note: This is also a good time to go through some basic *data cleaning*: if there are columns that are obviously extraneous to the data analysis (e.g., IDs or metadata that have no bearing on your analysis), you can remove those now to make your life easier.

- Describe the continuous response variable you will use for your prediction task in the next component. Explain your choice.

- Describe the binary response variable you will use for your prediction task in the next component. Explain your choice. *Note*: You can always create a binary response variable by starting with a continuous variable Y, and then defining $Z = 1$ if Y exceeds a fixed threshold, and $Z = 0$ otherwise.

- Loosely speaking, what questions might you be able to answer using this dataset? What makes this dataset exciting?

Recommended: The following list serves as a guide of the kinds of questions your report might address; good reports will make decisions about which of these are most meaningful to include, or even other features of your data analysis that are not described below.

- Are any values in your dataset `NULL` or `NA`? Think of what you will do with rows with such entries: do you plan to delete them, or still work with the remaining columns for such rows?

- Are there any columns that appear to be irrelevant to the questions you would like to answer?

- Find covariates that are most strongly positively correlated, as well as most strongly negatively correlated, with your choices of response variable.

  Are there variables you think should affect your response variable, that nevertheless have weak correlation with your response variable?

- Look for mutual correlations between these variables you identified in the last part. Create scatterplots for pairs of covariates you believe correlates well to the response variable.

  Are correlations *transitive* in your data? That is, if A is correlated strongly with B, and B with C, is A also correlated strongly with C in your data?

- Are there variables you would like to *add* to your dataset as you embark on your analysis? For example, are there interactions or higher order terms that might be relevant?

- Can you visualize interesting patterns in your data?

Note: These steps are just the tip of the iceberg! Ideally, you will look at your data many different ways; for example, it's useful to look at means and variances of columns, grouped based on the level of a categorical variable. (E.g., in the College Scorecard data, you might look at how future earnings differ for public vs. private colleges.)

Try to play with and understand your data as much as you can *before* you start building models in the next component!

5.2 Prediction

In this component, your goal is to find a model that predicts your outcome variable(s) well. This component of the report should highlight your work in building predictive models for both a classification and regression task, and should be **2-3** pages in length (though recall the overall project report length should not exceed 4 pages).

We want to see how you approach building a good predictive model. Be careful about your approach and methodology; it is far more important to be correct in your approach and precise in describing it, than it is to use fancy methods.

The following are **required** steps:

a) As part of your report, you should build a model for both a **regression** task (i.e. continuous outcome variable) and a **classification** task (i.e., binary outcome variable) on your data set. You should use the continuous and binary outcomes you identified in the first part of the report.

b) Describe the best model you built for each task, and how you evaluated them. Turn in code for the best model that you built for each task; code *does not* count towards the report page count. Note that best is subjective here; motivate why this model is the best.

c) **DO NOT** use the test set you previously created. For now we just want you to focus on *building* the best model you can. You will use whatever model you build on the test set in the final part of the project.

d) For the best of all the models you tried, give an estimate of what you think the test error will be when you run your model on the previously held out test set; explain how you arrived at your estimate.

The following are some **suggested** steps to get you thinking about how you might approach this part of the project:

- Before you launch into creating your predictive models, first think about what a good outcome would be: What would be acceptable performance? Should the model be interpretable, or can it be a black box? If you binarize a continuous variable as outcome, what is a threshold that makes sense? etc.

- The next step is to start with some baselines to give you some sense of what you are trying to achieve. For example, some baselines to compare to:

  - For regression, you could always predict the sample mean; for classification, you could predict whichever label occurs most frequently in the data.
  - For regression, you could build an OLS model using all the covariates; for classification, you could build a logistic regression model using all the covariates.

  Feel free to use your own baseline, if you can justify it! Based on your knowledge, how much improvement over a simple baseline do you think is feasible?

- What is your objective for the classification task? 0-1 loss or something else?
- Make sure you have an evaluation strategy in mind; recall that if you're not sure what to do, cross validation is always a reasonable way to compare model performance.
- Consider using transformations, interaction terms, regularization, etc.
- Feel free to try out any other methods you are curious about: not only methods you learned in class (e.g., k-nearest-neighbors or naive Bayes), but also other methods you may have heard about elsewhere.
- Try to use bias and variance to help structure your thinking about what different predictive models might be doing on the data.
- Be reflective: try to notice and comment on any practices in your model building procedure that you think might lead your estimate of test error to be overly optimistic.