CIS 670, Spring 2021
Course Project

1 Data

You must first choose a dataset to analyze. There are many datasets out there, many of them are interesting, and many are easy to import into R. Kaggle (www.kaggle.com) has a wide variety. However, many of these are designed for a specific task. It is not sufficient to build a model with good accuracy at a classification or regression task; you must analyze the data more broadly.

Proposal: You must state which dataset or datasets you plan to work with and briefly describe them (1-3 sentences).

Project: You must provide background on the dataset or datasets you worked with, such as who created it, why, how it was collected, for what purpose it was collected, what attributes are included, what representation decisions were made and how that affects the data, etc. A good reference for questions to ask about your data is the paper "Datasheets for Datasets," by Gebru et al. (2018) (`https://arxiv.org/abs/1803.09010`).

2 Questions

You must then choose some questions to answer, or attempt to answer. What relationships do you hypothesize exist? Feel free to look at the data to gain some intuitions before you pick a question to dive into more deeply. This project should represent several weeks of work, so choose questions that are broad enough to spend a whole project exploring. You should list your initial ideas in your project proposal, but it's fine if your ideas evolve. It's also fine to answer many small questions if you don't see a big question to answer.

Proposal: You must list some questions you plan to investigate and a little bit of information about why you're interested in those questions how you plan to investigate them.

Project: You must clearly state the questions you're attempting to answer. Some questions could be a bit vague or broad, such as determining the basic properties of the different attributes, but others should be more specific, such as determining if a particular relationship exists.

3 Visualization

(Project only.) An essential part of data science is visualization, to understand the data. R makes some of this easy, through powerful plotting tools. You can easily view scatterplots of all pairs of variables with the pairs function, you can generate plots or histograms for individual variables, and many more complex visualizations are available in other R packages. ggplot2 is particularly powerful.

You can learn more R techniques for manipulating, visualizing, and presenting data in the free book “R for Data Science” (`https://r4ds.had.co.nz/`).

4 Models

Your analysis must include fitting and interpreting models. This could include fitting multiple models and analyzing which one fits best and why. It could include analyzing regression coefficients to determine if there’s evidence of a relationship between certain variables and the outcome. It could include checking for interactions and confounders. It could also include preprocessing the data in different ways to obtain better fits — transforming variables, dimensionality reduction, etc.

Proposal: Your proposal should briefly mention your ideas about models to use and things to try, but I expect this will evolve a lot during your project.

Project: Your final project should include models and analysis, such as comparisons and contrasts, checking for statistical significance (if/when appropriate), discussions of modeling assumptions and limitations, etc.

5 Results, Analysis, and Discussion

(Project only.) Your project should take all the visualization and modeling and attempt to draw conclusions from it. What is the evidence for your conclusions? What is the evidence against your conclusions? What are the limitations of your analysis? What other data would you want to collect in order to make a better decision?

6 Impact

(Project only.) Finally, your project must discuss the impact that your methods or conclusions could have if adopted in the real world. What does your analysis suggest should be done? Do you agree? Who would be impacted, and how? Reducing costs? Increasing or decreasing bias? Shifting power? What would be the impact if your models were accurate? What would be the impact if they weren’t?

7 Grading

The paper will be graded on a 50-point scale.

- Proposal - 5 points. The proposal should describe the dataset, some of the questions to address, and a little bit about the proposed approach. I expect that your ideas will evolve, so your final project could look somewhat different. I expect most people will get all of these points.

- Visualizations and descriptive analysis - 10 points. Include appropriate visualizations to illustrate key properties of the dataset and the questions you're trying to answer.

- Modeling - 10 points. Use appropriate methods for answering the questions you seek to answer. Apply some creativity in your approach. Understand the assumptions and limitations of your methods. Use proper tuning and validation methods.

- Analysis and discussion - 10 points. Draw what conclusions you can and observe what remains inconclusive. Explain the evidence.

- Impact - 5 points. Explain the likely impact your methods could have if adopted widely. If your results have little impact beyond scientific curiosity, then this might just be a single paragraph.

- Written Presentation - 5 points. The project should be done as a notebook or markdown file in R, including all code, visualizations, and text within the notebook. Text should be clear and detailed enough to understand the background, approach, and results. (This will be wordier than a typical Jupyter notebook, which often has sparse text and a lot of code.)

- In-class Presentation - 5 points. Present your project to the class. Slides are recommended as visual aids.

Important Note: If you want to do something awesome that doesn't quite fit into this template, ask me! I'd rather have you do something great that you're excited about than check off a few boxes for a grade.