Final project outline

Joshua Loftus

Note: This is not final, since there may need to be clarifications as questions arise in the process. That much flexibility is necessary for open-ended projects where teams can pick their own datasets.

Grading expectations

For grading purposes you should keep in mind three high level goals.

1. **I must be absolutely certain that you (meaning your team) did the work**. If you use any resources online to find examples of code then you should cite the resource. It's also helpful to put comments in your code explaining it, for example:

```
dataset %>%
  some_function() %>% # to do something
  another_function()  # for something else
```

2. **I should be able to understand what you did**. State your motivating question(s), describe your dataset, and explain your models and methods **clearly** and **concisely**. Make sure that your markdown document is formatted properly and the plots and other outputs display correctly. Double check this for the final PDF file before you submit it.
3. **I should be convinced that you understand what you did**. Some software libraries are easy enough to use that you can get impressive results without knowing how or why. But this is not a course on how to use `tidymodels`, it's a course on machine learning. Without sacrificing clarity or brevity, your explanations should show deep understanding.

Notably absent from this list: getting significant results, using the most cutting edge methods, etc. Those are not required, and in fact putting too much energy toward them puts us at higher risk of *overfitting* our whole process to one dataset.

Basic requirements

Proposal stage

1. Each member of your team should read the questions in this survey: https://forms.gle/EmBSz5KgaiS9RbBR9 (https://forms.gle/EmBSz5KgaiS9RbBR9)
2. Working together, your team should decide on a dataset and write an initial description that can be included in the final project write-up. Then have one (**only one**) member of your team complete the survey by copy/pasting sentences from your initial description

- Optionally, you can also send me an email to let me know when your team has filled out the survey so I can reply sooner.

The sooner you finish this proposal stage and receive confirmation from me that the dataset is approved the sooner you can get started on the rest of the project. Of course, you can also work on the rest of the project before hearing back from me if you are OK with taking the risk that you might need to redo some work if I ask you to change something about the dataset choice.

Analysis stage

Your team should work together to create one R Markdown (`Rmd`) document with all of your analysis and interpretation.

In order to streamline the process of fitting and comparing multiple models you should **learn about and use the `tidymodels` package**. This blog post (https://rviews.rstudio.com/2019/06/19/a-gentle-intro-to-tidymodels/) shows some basic usage, and this reference page (https://recipes.tidymodels.org/reference/index.html) shows the names of a lot of useful helper functions. The standardized approach of this package guarantees each model/method uses the same training/testing data with the same pre-processing steps. (Although it's relatively new I expect proficiency with `tidymodels` to become a more common expectation for data science roles).

Your project should use multiple machine learning models or methods, meeting these minimum requirements:

- At least one model must be simple enough to consider as a baseline for comparison to the more sophisticated models. Regression models or nearest neighbors methods, based on only a few predictors, are good candidates for baseline methods.
- At least one *non-baseline* model must be (relatively) interpretable. For this model you should write a brief sub-section including your interpretation of the results. You could compare to a baseline model on both predictive accuracy and (in)consistency of interpretations.
- At least one model must be (relatively) high-dimensional. If your dataset has many predictors, and the number of observations is not much larger, then for example you could fit a penalized regression model using all the predictors. If your dataset does not have many predictors you could consider models that include non-linear transformations, interaction terms, and/or local smoothing to increase the effective degrees of freedom.
- At least one model must be (relatively) more focused on predictive accuracy without interpretability. Imagine that you would submit this model to a prediction competition where the winner is chosen using a separate set of test data from the same data generating process (in-distribution generalization).

Note that it may be possible to satisfy multiple requirements with one model, e.g. the last two requirements in the above list.

Write-up stage

Structure your R Markdown report to make it easy to read and navigate. Use sections/subsection structure. Break up code into multiple chunks, include comments, break lines of code into multiple lines so they don't extend off the side of the page.

Every team member should check every part of the document, ask each other clarifying questions, and write whatever comments and explanations are necessary to make sure that **everyone is responsible for all parts of the project**. It's very important that you have some time for this stage before the submission deadline so you can make edits and changes if necessary.