Homework 3

Due Wednesday, October 20th at 11:59pm on Moodle

Deliverables: Please use this template to knit an HTML document. Convert this HTML document to a PDF by opening the HTML document in your web browser. *Print* the document (Ctrl/Cmd-P) and change the destination to "Save as PDF". Submit this one PDF to Moodle.

Alternatively, you may knit your Rmd directly to PDF if you have LaTeX installed.

Project Work

(Note: This includes HW2 investigations plus a few tasks for dealing with non-linearity.)

Goal: Begin an analysis of your dataset to answer your **regression** research question.

Collaboration: Form a team (2-3 members) for the project and this part can be done as a team. Only one team member should submit a Project Work section. Make sure you include the full names of all of the members in your write up.

Data cleaning: If your dataset requires any cleaning (e.g., merging datasets, creation of new variables), first consult the R Resources page to see if your questions are answered there. If not, post on the #rcode-questions channel in our Slack workspace to ask for help. *Please ask for help early and regularly* to avoid stressful workloads.

Required Analyses:

1. **Initial investigation: ignoring nonlinearity (for now)**
    a. Use ordinary least squares (OLS) by using the `lm` engine and LASSO (`glmnet` engine) to build a series of initial regression models for your quantitative outcome as a function of the predictors of interest. (As part of data cleaning, exclude any variables that you don't want to consider as predictors.)
        - You'll need two model specifications, `lm_spec` and `lm_lasso_spec` (you'll need to tune this one).
    b. For each set of variables, you'll need a `recipe` with the `formula`, `data`, and pre-processing steps
        - You may want to have steps in your recipe that remove variables with near zero variance (`step_nzv()`), remove variables that are highly correlated with other variables (`step_corr()`), normalize all quantitative predictors (`step_normalize(all_numeric_predictors())`) and add indicator variables for any categorical variables (`step_dummy(all_nominal_predictors())`).
        - These models should not include any transformations to deal with nonlinearity. You'll explore this in the next investigation.
    c. Estimate the test performance of the models using CV. Report and interpret (with units) the CV metric estimates along with a measure of uncertainty in the estimate (`std_error` is readily available when you used `collect_metrics(summarize=TRUE)`).
        - Compare estimated test performance across the models. Which models(s) might you prefer?
    d. Use residual plots to evaluate whether some quantitative predictors might be better modeled with nonlinear relationships.
    e. Which variables do you think are the most important predictors of your quantitative outcome? Justify your answer. Do the methods you've applied reach consensus on which variables are most important? What insights are expected? Surprising?
        - Note that if some (but not all) of the indicator terms for a categorical predictor are selected in the final models, the whole predictor should be treated as selected.

Note: after this process, you might have a set of models (one of which has predictors chosen using LASSO, one model with all the predictors of interest, and perhaps some models with subsets of predictors that were chosen based on the data context rather than an algorithmic process)

2. **Accounting for nonlinearity**
    - Update your models to use natural splines for some of the quantitative predictors to account for non-linearity (these are GAMs).
        - I recommend using OLS engine to fit these final models.
        - You'll need to update the recipe to include `step_ns()` for each quantitative predictor that you want to allow to be non-linear.
        - To determine number of knots (`deg_free`), I recommend fitting a smoothing spline and use `edf` to inform your choice.
    - Compare insights from variable importance analyses here and the corresponding results from the Investigation 1. Now after having accounted for nonlinearity, have the most relevant predictors changed?
        - Do you gain any insights from the GAM output plots (easily obtained from fitting smoothing splines) for each predictor?
    - Compare model performance between your GAM models that the models that assuming linearity.
        - How does test performance of the GAMs compare to other models you explored?
    - Don't worry about KNN for now.

3. **Summarize investigations**
    - Decide on an overall best model based on your investigations so far. To do this, make clear your analysis goals. Predictive accuracy? Interpretability? A combination of both?

4. **Societal impact**
    - Are there any harms that may come from your analyses and/or how the data were collected?
    - What cautions do you want to keep in mind when communicating your work?

Portfolio Work

Length requirements: Detailed for each section below.

Organization: To help the instructor and preceptors grade, please organize your document with clear section headers and start new pages for each method. Thank you!

Deliverables: Continue writing your responses in the same Google Doc that you set up for Homework 1. Include that URL for the Google Doc in your submission.

Note: Some prompts below may seem very open-ended. This is intentional. Crafting good responses requires looking back through our material to organize the concepts in a coherent, thematic way, which is extremely useful for your learning.

Revisions:

- Make any revisions desired to previous concepts. **Important note:** When making revisions, please change from "editing" to "suggesting" so that we can easily see what you've added to the document since we gave feedback (we will "accept" the changes when we give feedback). If you don't do this, we won't know to reread that section and give new feedback.
- General guidance for past homeworks will be available on Moodle (under the Solutions section). Look at these to guide your revisions. You can always ask for guidance in office hours as well.

New concepts to address:

- **Splines:**
  - Algorithmic understanding: Explain the advantages of natural cubic splines over global transformations and piecewise polynomials. Also explain the connection between splines and the ordinary (least squares) regression framework. (5 sentences max.)
  - Bias-variance tradeoff: What tuning parameters control the performance of this method? How do low/high values of the tuning parameters relate to bias and variance of the learned model? (3 sentences max.)
  - Parametric / nonparametric: Where (roughly) does this method fall on the parametric-nonparametric spectrum, and why? (3 sentences max.)
  - Scaling of variables: Does the scale on which variables are measured matter for the performance of this algorithm? Why or why not? If scale does matter, how should this be addressed when using this method? (3 sentences max.)
  - Computational time: When using splines, how does computation time compare to fitting ordinary (least squares) regression models? (1 sentence)
  - Interpretation of output: SKIP - will be covered in the GAMs section
- **Local regression:**
  - Algorithmic understanding: Consider the R functions `lm()`, `predict()`, `dist()`, and `dplyr::filter()`. (Look up the documentation for unfamiliar functions in the Help pane of RStudio.) In what order would these functions need to be used in order to make a local regression prediction for a supplied test case? Explain. (5 sentences max.)
  - Bias-variance tradeoff: What tuning parameters control the performance of the method? How do low/high values of the tuning parameters relate to bias and variance of the learned model? (3 sentences max.)
  - Parametric / nonparametric: Where (roughly) does this method fall on the parametric-nonparametric spectrum, and why? (3 sentences max.)
  - Scaling of variables: Does the scale on which variables are measured matter for the performance of this algorithm? Why or why not? If scale does matter, how should this be addressed when using this method? (3 sentences max.)
  - Computational time: In general, local regression is very fast, but how would you expect computation time to vary with span? Explain. (3 sentences max.)
  - Interpretation of output: SKIP - will be covered in the GAMs section
- **GAMs:**
  - Algorithmic understanding: How do linear regression, splines, and local regression each relate to GAMs? Why would we want to model with GAMs? (5 sentences max.)
  - Bias-variance tradeoff: What tuning parameters control the performance of the method? How do low/high values of the tuning parameters relate to bias and variance of the learned model? (3 sentences max.)
  - Parametric / nonparametric: Where (roughly) does this method fall on the parametric-nonparametric spectrum, and why? (3 sentences max.)
  - Scaling of variables: Does the scale on which variables are measured matter for the performance of this algorithm? Why or why not? If scale does matter, how should this be addressed when using this method? (3 sentences max.)
  - Computational time: How a GAM is specified affects the time required to fit the model - why? (3 sentences max.)
  - Interpretation of output: How does the interpretation of ordinary regression coefficients compare to the interpretation of GAM output? (3 sentences max.)
- **Evaluating classification models:** Consider this xkcd comic. Write a paragraph (around 250 words) that addresses the following questions. Craft this paragraph so it flows nicely and does not read like a disconnected list of answers. (Include transitions between sentences.)

  - What is the classification model here?
  - How do the ideas in this comic emphasize comparisons between overall accuracy and class-specific accuracy measures?
  - What are the names of the relevant class-specific accuracy measures here, and what are there values?

- **Logistic regression:**
  - Algorithmic understanding: Write your own example of a logistic regression model formula. (Don't use the example from the video.) Using this example, show how to use the model to make both a soft and a hard prediction.
  - Bias-variance tradeoff: What tuning parameters control the performance of the method? How do low/high values of the tuning parameters relate to bias and variance of the learned model? (3 sentences max.)
  - Parametric / nonparametric: Where (roughly) does this method fall on the parametric-nonparametric spectrum, and why? (3 sentences max.)
  - Scaling of variables: Does the scale on which variables are measured matter for the performance of this algorithm? Why or why not? If scale does matter, how should this be addressed when using this method? (3 sentences max.)
  - Computational time: SKIP
  - Interpretation of output: In general, how can the coefficient for a quantitative predictor be interpreted? How can the coefficient for a categorical predictor (an indicator variable) be interpreted?

Reflection

Ethics: Read the article Getting Past Identity to What You Really Want. Write a short (roughly 250 words), thoughtful response about the ideas that the article brings forth. What skills do you think are essential for the leaders and data analysts of organizations to have to handle these issues with care?

Reflection: Write a short, thoughtful reflection about how things went this week. Feel free to use whichever prompts below resonate most with you, but don't feel limited to these prompts.

- How is collaborative learning? What are you working on to make it easier to work in groups? What is not working?
- What's going well? In school, work, other areas? What do you think would help sustain the things that are going well?
- What's not going well? In school, work, other areas? What do you think would help improve the things that aren't going as well? Anything that the instructor can do?

Self-Assessment: Before turning in this assignment on Moodle, go to the individual rubric shared with you and complete the self-assessment for the general skills (top section). After "HW3:", assess yourself on each of the general skills. Do feel like you've grown in a particular area since HW2?

Assessing yourself is hard. We must practice this skill. These "grades" you give yourself are intended to have you stop and think about your learning as you grow and develop the general skills and deepen your understanding of the course topics. These grades do not map directly to a final grade.