Homework 2

Due Friday, February 18 at 9:00am CST 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

Instructions

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.

Your Work a & b.

```
## See answer to Q2 to explain irregularities

# library statements
library(ISLR)
library(dplyr)
library(readr)
library(broom)
library(ggplot2)
library(tidymodels)
tidymodels_prefer() # Resolves conflicts, prefers tidymodel functions
# read in data

stroke <- read_csv("https://raw.githubusercontent.com/MarshallRoll/STAT_253_Project/main/healthcare-data
```

```
# data cleaning
stroke_clean <- stroke %>%
  filter(bmi != "N/A") %>%
  filter(smoking_status != "Unknown") %>%
  mutate(stroke_factor = factor(stroke))
```

```
# creation of cv folds

stroke_cv <- vfold_cv(stroke_clean, v = 6)
```

```
# model spec
lm_spec <-
    logistic_reg(mixture = 1, penalty = 0) %>%
    set_engine(engine = 'glmnet')

lm_lasso_spec <-
  logistic_reg() %>%
  set_args(mixture = 1, penalty = tune()) %>%
  set_engine(engine = 'glmnet') %>%
  set_mode('classification')

modAll <- fit(lm_spec,
            stroke_factor ~ .,
            data = stroke_clean)
```

```
# recipes & workflows
all_rec <- recipe( stroke ~ . , data = stroke_clean) %>%
    step_nzv(all_predictors()) %>% # removes variables with the same value
    step_novel(all_nominal_predictors()) %>% # important if you have rare categorical variables
    step_normalize(all_numeric_predictors()) %>%  # important standardization step for LASSO
    step_dummy(all_nominal_predictors())

model_wf <- workflow() %>%
  add_recipe(all_rec) %>%
  add_model(lm_spec)
```

```
# fit & tune models
# this is where our code broke
modAll_cv <- fit_resamples(model_wf,
  resamples = stroke_cv,
  metrics = metric_set(mae)
)
```

c.

```
#  calculate/collect CV metrics
mod1_cv %>% collect_metrics()
```

d.

```
# visual residuals
```

e.

2. **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?

Because we only have three quantitative variables in our model, we tried to predict a classification outcome. However, we've come to a stopping point due to our limitation on the knowledge of logistic regressions. We are unsure what would be a good evaluation metric as well as how to proceed with a logistic regression. We talked with Bryan and he told us that we should find a new dataset that has suitable outcome variables and resubmit the linear regression portion of the assignment next week.

3. **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?

We were unable to conclude where the data was sourced from, but we believe that there were no harms in how it was collected as the information seems to be from a reliable source. Additionally, there could be harms that stem from our analyses due to type 1 and 2 errors when predicting if an individual is at risk of having a stroke. we want to be absolutely clear that our analyses and prediction will not be 100% correct when predicting if an individual is at risk of having a stroke.

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 guideance 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:

- **Subset selection:**
  - Algorithmic understanding: Look at Conceptual exercise 1, parts (a) and (b) in ISLR Section 6.8. **What are the aspects of the subset selection algorithm(s) that are essential to answering these questions, and why?** (Note: you'll have to try to answer the ISLR questions to respond to this prompt, but the focus of your writing should be on the question in bold here.)
  - 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: What computational time considerations are relevant for this method (how long the algorithms take to run)?
  - Interpretation of output: What parts of the algorithm output have useful interpretations, and what are those interpretations? **Focus on output that allows us to measure variable importance. How do the algorithms/output allow us to learn about variable importance?**
- **LASSO:**
  - Algorithmic understanding: Come up with your own analogy for explaining how the penalized least squares criterion works.
  - 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: What computational time considerations are relevant for this method (how long the algorithms take to run)?
  - Interpretation of output: What parts of the algorithm output have useful interpretations, and what are those interpretations? **Focus on output that allows us to measure variable importance. How do the algorithms/output allow us to learn about variable importance?**
- **KNN:**
  - Algorithmic understanding: Draw and annotate pictures that show how the KNN (K = 2) regression algorithm would work for a test case in a 2 quantitative predictor setting. Also explain how the curse of dimensionality affects KNN performance. (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: The KNN algorithm is often called a "lazy" learner. Discuss how this relates to the model training process and the computations that must be performed when predicting on a new test case. (3 sentences max.)
- Interpretation of output: The "lazy" learner feature of KNN in relation to model training affects the interpretability of output. How? (3 sentences max.)

Reflection

Ethics: Read the article Automated background checks are deciding who's fit for a home. Write a short (roughly 250 words), thoughtful response about the ideas that the article brings forth. What themes recur from last week's article (on an old Amazon recruiting tool) or movie (Coded Bias)? What aspects are more particular to the context of equity in housing access?

In Portfolio

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 are class-related things going? Is there anything that you need from the instructor? What new strategies for watching videos, reading, reviewing, gaining insights from class work have you tried or would like to try?

So far, things in the class seem to be going well. It can definitely be challenging to understand the specific mathematical components of each of the models, but overall I feel like I generally understand what each of the models does and when we might want to use it.

- How is group work going? Did you try out any new collaboration strategies with your new group? How did they go?

Group work is going well, although we hit a snag with our dataset not being suitable for regression. We were able to meet outside of class and setup a Github repository as well as discuss our approach to the content conversation. The strategy of meeting up before or after class is proving to be effective.

- How is your work/life balance going? Did you try out any new activities or strategies for staying well? How did they go?

There is a lot of work at Macalester, but overall things are going well. I haven't tried anything new in this area.

Link to Portfolio: https://docs.google.com/document/d/1CcLU0YgUvIK9FL9SOQt_jWs6pnpCVbFnPaGuqiotCYQ/edit