Problem Set 10

Clustering

[YOUR NAME]

Due Date: 2024-04-05

Open `RStudio` and create a new RMarkDown file (`.Rmd`) by going to `File -> New File -> R Markdown...` . Accept defaults and save this file as `[LAST NAME]_ps10.Rmd` to your `code` folder.

Copy and paste the contents of this `.Rmd` file into your `[LAST NAME]_ps10.Rmd` file. Then change the `author: [Your Name]` to your name.

We will be using the `pres_elec.rds` file from the course github page (https://github.com/jbisbee1/DS1000_S2024/raw/main/data/pres_elec.rds).

All of the following questions should be answered in this `.Rmd` file. There are code chunks with incomplete code that need to be filled in.

This problem set is worth 8 total points, plus two extra credit points. The point values for each question are indicated in brackets below. To receive full credit, you must have the correct code. In addition, some questions ask you to provide a written response in addition to the code.

You are free to rely on whatever resources you need to complete this problem set, including lecture notes, lecture presentations, Google, your classmates…you name it. However, the final submission must be complete by you. There are no group assignments. To submit, compiled the completed problem set and upload the PDF file to Brightspace on Friday by midnight. Also note that the TAs and professors will not respond to Campuswire posts after 5PM on Friday, so don't wait until the last minute to get started!

Good luck!

*Copy the link to ChatGPT you used here: _____________________

Question 0

Require `tidyverse` and `tidymodels` , and then load the `pres_elec.rds` (https://github.com/jbisbee1/DS1000_S2024/raw/main/data/pres_elec.rds) data to an object called `dat` .

Question 1 [2 points]

Describe the data. What is the unit of analysis? What information do the columns provide? What is the period described (i.e., how far back in time does the data go?). Is there any missing data? If so, "where" is it, in terms of both columns and in terms of the observations that have missing data?

```
# INSERT CODE HERE
```

> - Write response here

Question 2 [2 points]

Perform *k*-means analysis on the Republican and Democrat votes with *k* = 2, and then plot the results, coloring the points by cluster assignment. Then predict the `GOP_win` binary outcome as a function of the cluster assignment using a logit regression. Make sure to `factor(cluster)` in the regression. What is the AUC for this model? Finally, use cross validation with an 80-20% split to re-calculate the AUC. Overall, would you say that the *k*-means algorithm helps you predict which counties vote Republican?

```
set.seed(123)
# K-means with k = 2
# INSERT CODE HERE

# Plotting the result
dat %>%
  select(...) %>%
  drop_na() %>%
  mutate(cluster = ...) %>%
  ggplot(aes(x = ...,y = ...,color = factor(...),group = 1)) +
  geom_point() +
  labs(x = '',
       y = '',
       color = '')
```

```
## Error in dat %>% select(...) %>% drop_na() %>% mutate(cluster = ...) %>% : could not
find function "%>%"
```

```
# Create dataset for analysis
toanal <- dat %>%
  select(...,...,...) %>%
  drop_na() %>%
  mutate(cluster = ...)
```

```
## Error in dat %>% select(..., ..., ...) %>% drop_na() %>% mutate(cluster = ...): could
not find function "%>%"
```

```
# Estimate logit model
summary(m <- glm(...,toanal,family = binomial))
```

```
## Error in summary(m <- glm(..., toanal, family = binomial)): '...' used in an incorrec
t context
```

```
# Calculate AUC
roc_auc(toanal %>%
          mutate(prob_win = ...,
                 truth = ...),
        truth,prob_win)
```

```
## Error in roc_auc(toanal %>% mutate(prob_win = ..., truth = ...), truth, : could not f
ind function "roc_auc"
```

```
# Calculate cross-validated result
cvRes <- NULL
for(i in 1:100) {
  # INSERT CODE HERE
}

# Cross-validated AUC
# INSERT CODE HERE
```

> - Write response here.

Question 3 [2 points]

Now create an elbow plot by looping over potential values of *k* from 1 to 30 and plotting the *k* on the x-axis and the total Within Sum of Squares (total WSS) on the y-axis. What value of *k* would you use? Then re-run the preceding analysis with that value of *k* and interpret the results. Does the model improve?

```
# Looking at multiple values of k
kRes <- NULL
for(k in 1:30) {
  # Calculate k-means cluster solution for given value of k
  kResTmp <- # INSERT CODE HERE

  # Save result including value of k and the total WSS
  kRes <- data.frame(withinSS = ...,
             k = ...) %>%
    bind_rows(kRes)
}
```

```
## Error in data.frame(withinSS = ..., k = ...) %>% bind_rows(kRes): could not find func
tion "%>%"
```

```
# Plotting the elbow plot. Looks like k=4 is the elbow?
# INSERT CODE HERE

# Rerunning with optimal k
# INSERT CODE HERE

# Plotting again
# INSERT CODE HERE

# Create dataset for analysis
# INSERT CODE HERE

# Estimate logit model
# INSERT CODE HERE

# Calculate AUC
# INSERT CODE HERE

# Calculate cross-validated result
# INSERT CODE HERE

# Cross-validated AUC
# INSERT CODE HERE
```

> - Write response here

Question 4 [2 points]

Re-do the preceding analysis except instead of using total votes, calculate the percent vote share for Democrats and Republicans in each county. Then identify the optimal value of *k* using the elbow plot visualization, use this as your value of *k* for the clustering solution and again plot the results. Now use a logit regression to predict `GOP_win` as a function of the cluster membership for each county and calculate the AUC using the same cross validation method. Does your answer change?

```
# INSERT CODE HERE
```

> - Write response here.

Extra Credit [2 points]

Provide an explanation for why the *k*-means results are so much more helpful when run using vote shares instead of total votes.

> - Write response here