STATS/CSE 780
Homework Assignment 2

Pratheepa Jeganathan

03 February, 2023

Instruction

- **Due before 10:00 PM on Friday, February 17, 2023.**
- **Submit a copy of the PDF with your report (2-3 pages) and technical supplemental material (less than 10 pages) to Avenue to Learn using the link that was emailed to you.**
  - Technical supplemental material can only include R codes for the results reported.
- **Late penalty for assignments: 15% will be deducted from assignments each day after the due date (rounding up).**
- **Assignments won't be accepted 48 hours after the due date.**

Assignment Standards

Your assignment must conform to the Assignment Standards listed below.

- RMarkdown or LaTeX is strongly recommended to write the report, and RMarkdown must be used to write the supplemental material.
- Report is about the results by applying data science methods and how you interpret or discuss the results. Don't show in the report how you do the analysis using R or Python.
- Technical supplemental material is how you produce the report results using R or Python. Don't print chunk messages, warnings, or extended data frames in the PDF.
- Write your name and student number on the title page. We will not grade assignments without the title page.

- Eleven-point font (times or similar) must be used with 1.5 line spacing and margins of at least 1~inch all around (The RMarkdown for this assignment used these formats).
- Your report may not exceed **three pages**, including tables and figures. It would help if you chose the tables and figures for the report. You can also keep other tables and figures in the supplementary material (**less than 10 pages**) and refer to the report. In addition, you may use one page for the bibliography and one page for the title page.
- You may discuss homework problems with other students, but you must prepare the written assignments yourself.
- No screenshots are accepted for any reason.
- The writing and referencing should be appropriate to the graduate level.
- The instructor may use various tools, including publicly available internet tools, to check the originality of the submitted work.

Question

Find a dataset that is suitable for classification. Some sites for dataset search are 1) Google Dataset Search, or 2) Kaggle Datasets, or 3) UCI Machine Learning Repository.

Do not use datasets that have been used in class or collected for your research (not publicly available) or in the textbooks used in this course or R or Python package data.

The dataset must have **at least five variables**.

1. Briefly describe your chosen dataset and clearly explain where it was sourced.
2. Produce some numerical and graphical summaries of the data. Do there appear to be any patterns? (what are variables, summaries, number of observations, data types, correlation, association analysis, outliers, and missing values analysis.)
3. Split the data into a training set and a hold-out set. Describe your choice of splitting.
4. Perform the following methods on the training set: logistic regression, K-nearest neighbor, and decision tree.
   (i) For each classifier, describe if you used any data or statistical transformation or select subset of predictors.
   (ii) For each classifier, describe the choice of tuning parameters (if any).
   (iii) For each classifier, describe the most important predictor variable(s) to classify the response or explain why can't you find this from the classifier.
   (iv) For the logistic regression, interpret the regression coefficient of the most important variable to classify the response.
5. Use the hold-out set to evaluate the performance of the classifiers. Compare and contrast the performance of the classifiers using the miss-classification error rate. If the classifier needs a cutoff to classify the labels, use sensitivity and specificity analysis to find the cutoff.
6. Perform logistic regression with shrinkage (lasso) on the training set.
   (i) Which shrinkage value seems to perform the best on this data set?
   (ii) Compare and contrast the interpretation with and without shrinkage.
   (iii) Compare and contrast the performance with and without shrinkage on the hold-out set.

7. Clearly state what conclusions (at least two) can be drawn from your analysis — these conclusions should be cast in the context of your chosen dataset.

Grading scheme

1.		source of dataset [1] describe your dataset [1] explain why the dataset is fit to the classifiers [1]
2.		variable description (selected or group of variables) [1] statistical summaries - no points if the graphs and tables are not readable [2] number of observations versus variables [1] data types [1] correlation, association analysis - no points if the graphs and tables are not readable [2] outliers detection and handling - no points if the graphs and tables are not readable [2] missing value detection and handling - no points if the graphs and tables are not readable [1]
3.		describe the (stratification) splitting [1]
4.	(i)	data and statistical transformation or subset of predictors for each classifier [3]
	(ii)	Choice of tuning parameters for each classifier [3]
	(iii)	the most important predictor for each classifier [3]
	(iv)	Interpret the logistic regression coefficient [1]
5.		sensitivity and specificity analysis to find the cut-off(s) [2] Compare and contrast the performance of the classifiers (at least three statements, use graphs) - no points if the graphs and tables are not readable [3]
6.	(i)	Perform logistic regression with shrinkage and find the value of shrinkage [2]
	(ii)	Compare and contrast the interpretation (at least one statement) [1]
	(iii)	Compare and contrast the performance (at least one statement) [1]
7.		at least thwo conclusions drawn from your analysis (should be cast in the context of your chosen dataset) [2]

References	Reference list starts on a new page, references are appropriate and list out in the **report** [2]
Supplementary material	Supplementary material starts on a new page, code readability, all codes are within the margins, the R codes and the outputs for the questions are presented [3]

The maximum point for this assignment is 40. We will convert this to 100%.