STATS/CSE 780
Project Report

Instructor: Pratheepa Jeganathan

20 March, 2023

Instructions

- **Due before 10:00 PM on Monday, April 17, 2023.**
- **Submit a copy of a PDF with your project report (not exceeding 15 pages, inclusive of tables and figures, but not including one title page and 1-2 pages of bibliography) and technical supplemental material (not exceeding 20 pages) to Avenue to Learn using the link that was emailed to you.**
- **Late penalty for project report: 15% will be deducted from the project report each day after the due date (rounding up).**
- **Project report submitted more than 48 hours late will receive a zero grade.**

Report Standards

Your project report must conform to the standards listed below.

- RMarkdown or LaTeX is strongly recommended to write the report, and RMarkdown or Jupyter Notebook must be used to write the supplemental material.
- Report is about the results by applying data science methods and interpreting or discussing the results. Don't show how you do the analysis using R codes in the report.
- Technical supplemental material is how you produce the results using R codes. Don't print chunk messages, warnings, or extended data frames or outputs in the PDF.
- Write your name and student number on the title page. We will only grade the report with 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 document used these formats).
- Your report may not exceed **fifteen (15) pages**, inclusive of tables and figures. You can also keep other tables and figures in the supplementary material (**less than 20 pages**) and refer to the report. In addition, you may use one page for the title page and 1-2 pages for the bibliography.
- You may discuss project ideas with other students, but you must prepare the report 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 submitted work.

For the project, you will identify a problem, identify a dataset related to the problem, choose two algorithms to address the problem, do exploratory data analysis (graphical and numerical summaries, outlier and missing value analysis), apply the algorithms, compare and contrast the results, state conclusions in the context of your chosen dataset.

Requirements:

Dataset:

- The dataset must have at least ten variables, including at least one categorical variable.
  - The dataset must include a categorical variable as a predictor if you use a supervised learning method.
  - If there are no categorical variables, you can use the clustering method to identify clusters and use the clusters as one of the predictors.
- **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.

Algorithms:

- One algorithm can be from the topics covered in the lectures before decision trees (inclusive) with modifications.

- Another algorithm can be from the topics covered in the lectures after decision trees or a new one.

Highly encourage that the project report is an extension of the project proposal. In addition to the project proposal,

1) you will technically describe two different algorithms to address a problem on a real dataset and technically describe comparison method,

2) carry out a thorough analysis of the dataset using the chosen methods (exploratory analysis, training, choosing tuning parameters, model assessment, model comparison),

3) interpret the results, both in technical terms and in the context of the data,

4) clearly state what conclusions can be drawn in the context of your chosen dataset.

Prepare and submit a report, subject to the **Report Standards** specified above, and the sections are given in the grading scheme.

Grading scheme

Introduction or Background	Source of the dataset [1] Description of the problem being addressed [1]
	Describe the statistical problem. E.g., prediction or classification, unsupervised learning or machine learning or inference, sparse learning, etc. [1]
Methods	Description of the two methods used [6] (3 points for each method), including the rationale for the choice of tuning parameters, etc. [2] (1 point for each method),
	Description of the comparison criterion [2], including the rationale for the choice of criterion to compare the methods [1]
Results	Results of exploratory data analysis (you can revise the results in the project proposal), including details on variables, summaries, number of observations, data types, relationships between variables, outliers, missing values, etc. [4] Points will be deducted for not readable figures and tables
	Interpretation of results applying two methods, both in technical terms and in the context of the data [6] (3 points for each method - atleast three interpreation for each method) Points will be deducted for not readable figures and tables
Conclusion	At least two findings, in the context of the data and the problem being addressed [2] At least two conclusions on the comparison of two techniques [2]
	Discuss at least two analytical challenges, computational cost, interpretability of the results, reproducibility of the results, etc. [2]

References	Reference list starts on a new page, references are appropriate and listed out in the **report** [2] Points will be deducted if there are fewer than three references Points will be deducted if the reference is not following the report
Supplementary material	Supplementary material starts on a new page [1] Computational workflow is understandable [2] Presented only codes for the results discussed in the report [2] Points will be deducted for presenting codes for the results not in the report Points will be deducted for not defined workflow

The maximum number of points for this assignment is 37. We will convert this to 100%.

The project report is 30% of your final grade.

Note: If the dataset used is not publicly and freely available, the grade will be zero.