STATS/CSE 780
Midterm Exam, Winter 2023
Project proposal

Pratheepa Jeganathan

16 February, 2023

Instruction

- **Due before 10:00 PM on Tuesday, March 7, 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 or Python codes for the results reported.
- **Late penalty for the project proposal: 15% will be deducted from the project proposal each day after the due date (rounding up).**
- **Project proposal won't be accepted 48 hours after the due date.**

Project Proposal Standards

Your project proposal must conform to the standards listed below.

1. RMarkdown or LaTeX is strongly recommended for writing the project proposal report, and RMarkdown or Jupyter Notebook must be used to write the supplemental material.
2. Write your name and student number on the title page. We will only grade the project proposal with the title page.
3. Report is about the results by applying data science methods, interpreting, and discussing the results. Don't show R or Python codes in the report.

4. Technical supplemental material is how you produce the results using R or Python codes. Don't print chunk messages, warnings, or extended data frames in the PDF.

- 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 project proposal report may not exceed **three pages**, inclusive of tables and figures. It would help if you chose the tables and figures accordingly 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 project proposals 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.
- Various tools, including publicly available internet tools, may be used by the instructor to check the originality of submitted work.

Question

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.
- Some sites for dataset search are
  1) Google Dataset Search, or
  2) Kaggle Datasets, or
  3) UCI Machine Learning Repository.

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.

Required workflow:

Step 1 (Introduction): Define the problem based on the dataset to be addressed (briefly describe your chosen dataset, clearly explain where it was sourced, and write a background/literature review on the dataset).

Step 2 (Introduction): Present your solution (identify the data science methods such as prediction or classification, unsupervised learning or machine learning or inference, etc., and how these methods solve the problem in Step 1.)

Step 3 (Methods): Define your deliverables and success criterion (what are the results of applying two algorithms, comparison criterion).

Step 4 (Methods): State your plan or approach (description of the algorithms, tuning parameters, feature engineering, computational cost, interpretability of the results, reproducibility).

Step 5 (Results): Report and interpret your preliminary results (conduct an exploratory data analysis, including variable names, summaries, number of observations, data types, correlation analysis, outliers analysis, handling missing values, and data transformation).

Step 6 (Project timeline): Outline your project completion timeline (completion date of Step 4, presentation slides (5%), oral presentation (10%), final written project report (30%)).

Grading scheme

The project proposal is 10% of your class project.

Step 1 (Introduction - problem description)	describe the problem, source of the dataset, literature review on the dataset, [3]
Step 2 (Introduction - present your solution)	prediction or classification, unsupervised learning or machine learning or inference, how these methods solve the problem in Step 1 [2]
Step 3 (Methods - define your deliverable and success criteria)	what are the results of applying two algorithms, what is/are the comparison criterion [2]
Step 4 (Methods - state your plan or approach)	description of the two algorithms [2] describe the variable selection, feature engineering, tuning parameters selection, computational cost, interpretability of the results, and reproducibility of the results [4]
Step 5 (Results - report your preliminary results)	exploratory data analysis, including variable names, data types, summaries, number of observations, correlation analysis, outliers analysis, handling missing values, data transformation [5]
Step 6 (Outline your project timeline)	completion date of Step 4, presentation slides (5%), oral presentation (10%), final written project report (30%) [1]
References	Reference list starts on a new page, references are appropriate and listed out in the **report** [2]
Supplementary material	Supplementary material starts on a new page; code readability; all codes are within the margins; the R or Python codes and the outputs for the questions are presented [3]

The maximum point for this assignment is 24. We will convert this to 100%.