IST 5535 – Machine Learning Algorithms and Applications

Spring 2021

Project Description

1. Project Purpose

The purpose of the group project is for students to try all different techniques and methods covered in the class to analyze a real dataset. Small groups (usually 3-4 students in each group, individual project is NOT allowed) shall deliver a machine learning project by using R.

2. Deliverables

- Deliverable 1: Project Proposal + Project Plan (**Due Friday Mar 19**)
  Discuss with your classmates and form a project team with 3-4 members in total. Submit your project proposal. If you need to discuss your project proposal with the instructor, please book a meeting with the instructor.

  (1) Project Proposal (1 page, MS Word format)

  The one-page proposal should be written in MS Word format containing the following contents:
  - Group members
  - Problem description and research question
  - Dataset (source, number of rows, number of columns, response variable etc.)
  - Regression or classification
  - Potential problems or challenges
  - Attach an appendix describing the meaning of all variables in your dataset

  (2) Project Plan (MS Excel format)

  The project plan describes of how your project is expected to be conducted. It contains the following important elements:
  - Tasks
  - Expected start time and finish time of each task
  - Person(s) assigned to each task

  Note:

  Before submitting proposal, each group must send their dataset to the instructor for approval. It is recommended that each group discuss their dataset with the instructor in the office hours.

  Once your dataset has been approved by the instructor, you need to start your work as soon as possible. Do NOT wait till the last week to start your group project. Last week projects are usually not of high quality and thus will lead to a low score.

- Deliverable 2: Project Report + Project Execution Report + Presentation Slides + Technical Appendix with R code (**Due Monday May 3**)

  (1) Project Report (4-5 pages, MS Word format)
  - Describe your research question and dataset
  - Explain all methods you applied or tried (how the method works, why you choose it)
  - A summary of methods comparison
  - Results
  - Conclusion

  (2) Project Execution Report (MS Excel format)

  The project execution report records how your project is actually conducted. It contains the following important elements:
  - Tasks (refined from project plan, you can add more tasks if necessary)
  - Expected start time and finish time of each task (from project plan)
  - Person(s) assigned to each task (from project plan)
  - Actual start time and finish time of each task
  - Hours spent by each person for each task

  (3) Presentation slides (MS PowerPoint format)

  (4) Technical Appendix
  - Describe technical details of your project (maximum 10 pages, MS Word format)
  - R Markdown + HTML report directly generated from R Markdown

3. Datasets

Students should come up with their own datasets. If you have your own first-hand datasets, that is great. If you don't have datasets at hand, you can find some existing datasets from online data repositories such as:

- Open Government Data
  - https://www.data.gov/
- Kaggle Datasets
  - https://www.kaggle.com/datasets
- Harvard Dataverse
  - https://dataverse.harvard.edu/dataverse/harvard
- KDD Nugets
  - https://www.kdnuggets.com/datasets/index.html
- Google Dataset Search
  - https://toolbox.google.com/datasetsearch

The dataset that is appropriate for the course project must satisfy the following criteria:

- The dataset must have potential to tell an interesting story;
- There must be relatively large number of variables (>=20) that could be extracted from the dataset;
- The dataset should have a large number of observations (>=1000);
- The dataset and your research questions have not been fully analyzed on the Internet, especially using R.

4. Project Presentation Guideline

Project presentations will take place at the end of the semester. Each presentation will last about 20 minutes (depending on the size of the class).

Note:

- This is a short time, so be prepared to be concise.
- This is the time to show the fruit of your semester's labor.

Your creative efforts should be realized within the following format:

- ***Teamwork*** (1 minute)
  Introduce your team members. Briefly describe their contributions.
- ***Background*** (2 minutes)
  Briefly explain the data, research question, context and setting.
- ***Data Analyses and Results*** (12 minutes)
  Briefly explain your data analysis methods and results. If you have many things, try to present the important methods and interesting findings.
- ***Conclusion*** (3 minute)
  Summarize your project: What insights you get from the data analysis? What limitations does the project have? Can the project be extended? What are the next phases of development for this project?
- ***Open Floor*** (2 minute)
  Finally, the presenting team will respond to questions from the class. Your instructor will moderate the time remaining.

5. Project Grading

- At the end of the semester, you will evaluate your group members on contribution to the group project.
- Each student's final grade for the group project will be assigned according to both group performance and the level of his/her contribution to the group.