STA 325: Project Description

Note: This document is not set in stone, and will be updated throughout the semester. Any updates will be communicated well before due dates.

This is a comprehensive machine learning / data mining project which will be worth 25% of your final grade. The project will consist of three parts: a proposal presentation (worth 5%), a final presentation (worth 10%), and a final report (worth 10%).

- *Groups*: You will work in groups of at most 4 people. You are free to form your own groups, but keep in mind that an ideal group should be interested in a common project topic, and should cover a diverse range of skills in modeling, coding, writing and presentation.
- *Topics*: You are free to choose your own project topics. The only requirements are that the topic must tackle an important real-world problem, and interpretable conclusions and meaningful decisions can be made from data analysis. Of course, all topics must be appropriate for the classroom setting; if you have any questions on this, please consult with me prior to the project proposal.
- *Class time*: While most work will be done outside of class, you will have some class time (around three classes total; see syllabus) to work on the project. *Use this time wisely*! Come prepared with ideas, questions, and analysis, and the TAs and I can then give advice on what may (or may not) be promising directions to work towards.
- *Data sources*: While I would imagine the project topics to be quite diverse, there are a few good data repositories which may be useful to all groups:
  - `https://toolbox.google.com/datasetsearch`: Google search for datasets.
  - `https://www.kdnuggets.com/datasets/index.html`: A list of data repositories for financial, economic and machine learning applications.
  - `https://www.data.gov/open-gov/`: Open data site for the federal government.
  - `https://www.kaggle.com/datasets`: Datasets used in Kaggle competitions.

  Some topic suggestions are also provided on the Google spreadsheet available via Piazza; you are welcome to use these or choose your own topic.
- *Tips*:

- *Start early*! Meaningful and interpretable predictive modeling requires careful thought and analysis, and is difficult to do well when crammed in a couple of days. Starting early will give you ample opportunities to seek feedback and improve your project.
- *Get feedback*! You are strongly encouraged to proactively seek feedback from me and the TAs during *all* stages of your project. Evaluations aside, we are here to *help* you develop this into a project you can be proud to present to potential employers. Of course, seeking and implementing feedback will improve the quality of your project, which may result in better grades.

Deliverables:

- *Proposal presentation*: For the proposal, you will make a short 5-10 minute informal presentation outlining a tentative framework for your project. You should make this presentation with your group in either mine or the TA office hours. We will be looking for:
  - *Executive summary*: A few slides highlighting (i) key project objectives, (ii) a high-level summary of project plans, (iii) the importance of this project and who it can benefit, and (iv) other relevant information. An executive summary should be concise and to-the-point. Think of it as a "sales pitch" to your boss, who will then decide the funding level for your project.
  - *Data description*: A few slides describing the data to be used. Topics to discuss may include (but is not exclusive to): (i) data sources – where are you getting the data? (ii) predictors and response – what variables will be used for modeling? (iii) data type – ordinal, nominal, continuous, etc., and (iv) data scraping / wrangling – how to extract and clean data for modeling?
  - *Project roadmap*: A few slides on how you intend to use the data to achieve project objectives. Some discussion points may include (but is not exclusive to): (i) inference and prediction goals, (ii) possible modeling strategies for the data, (iii) selecting a good model (or good models) which achieve inference and prediction goals. You may wish to include some exploratory data analysis (e.g., scatterplots, descriptive statistics, boxplots) to support your answers. Also consider including a flow chart to outline key steps in the roadmap.
- *Final report*: You will submit a final report (no more than 10 pages) with the following broad sections:
  - *Introduction*: A few paragraphs which (i) motivate problem importance & relevance (supported by relevant literature, if any), (ii) describe project goals and how such goals address the problem, as well as (iii) a high-level roadmap of the proposed methodology, and (iv) other relevant information for the reader. See project rubric for details.
  - *Data*: This should be an extension of the "Data description" section from your proposal. See project rubric for details.

- *Methodology*: Discussion & justification of model choice and features, and how the proposed model(s) fully addresses project goals. Any "downstream" uses of the model (e.g., for prediction, optimization, ranking) should be discussed in detail here. See project rubric for details.
- *Results*: Statistical analyses of the fitted model(s), and a translation of these findings into meaningful & understandable conclusions for the target audience (e.g., engineers, business managers, policy-makers, etc). See project rubric for details.
- *Conclusion*: A summary of key findings and potential impacts of your project.

We will be following the project rubric quite closely for grading. If any parts of the rubric are unclear, you should clarify with the instructor or TAs – we're here to help!

- *Oral presentation*: Your group will give a 20-minute presentation on the above topics. This presentation should highlight and summarize key points in your report. As such, your problem formulation should be convincing and well-motivated, your data / modeling approaches should be discussed and well-justified, and your findings should be clearly communicated and interpreted for the problem at hand. Your group will be graded on presentation poise and clarity, as well as how well your group members answer questions.