CIS/STA 3920: Term Project Guidelines

Project proposal	15%	Mon, 4/8
Final presentations	10%	Tues, 5/7 & Thurs 5/9 (in-class)
Final report	75%	Tues, 5/14
Peer evaluation	*participation grade*	Wed, 5/15

Overview

Groups are expected to integrate and apply core data mining concepts covered throughout the semester to address a business problem. This process includes:

(1) designing a methodology/approach to answer the business question,
(2) gathering and cleaning data,
(3) performing basic exploration of the data,
(4) building models using 2-3 different learning algorithms,
(5) optimizing performance of the models,
(6) comparing and evaluating the models, and
(7) presenting the findings in a business context.

Project Proposal *(1-2 pages)*

- Present the business problem your group will be addressing.
  - What is the issue? What is the goal (what are you predicting)? Who is the target audience? How will they benefit from this?
  - What would a new observation look like? Give an example.
- Identify the dataset(s).
  - Introduce the dataset. Where is the data from? Who originally collected the data and for what purpose? What is the size of the dataset? What does each observation represent?
- Describe the variables.
  - Provide brief descriptions for variables of interest: Basic definition (if not obvious), data type (int, num, char, factor), unit of measure (i.e., days, miles, $).
  - Identify which will be your target variable (y).
- Outline of methodology.
  - Numeric prediction or classification problem (or both)?
  - Validation method (how will you compare models?)
  - Data preparation (merging datasets, handling missing values, checking for data error, creating new predictors, converting continuous variables to categorical, binning values, etc.)

Good proposals will be **clear** and **concise**. They will be presented with clean and professional formatting (arial font; size 11; single-spaced; normal margins).

Final Report *(3-5 pages)*: **SUBMISSION LINK**

Content

- Introduction. *(1-2 paragraphs)*
  - Present the business problem and goals of the study.
  - Briefly state the methodological framework to be adopted (numeric prediction, classification, both?)
- Data description.
  - Describe the data (dimensions, observations, predictors, and target variable).
  - Specify key data cleaning tasks (i.e., how missing values were handled, binning values, creating new variables, merging datasets, etc.).
  - Summarize the data with EDA (descriptive statistics, visualizations).
- Model building. *(2-3 paragraphs)*
  - Briefly define the supervised learning algorithms used.
  - Describe the model development process for each (i.e., feature selection, parameter tuning, validation approach).
- Performance evaluation, model comparison. *(2-3 paragraphs)*
  - Compare and evaluate the performance of the models.
- Conclusions. *(2-3 paragraphs)*
  - How well does the model solve the business problem? Interpret metrics in context of business problem, compare against baseline, etc.
  - Summarize interesting findings regarding impactful predictors

Formatting

- Cover page. Include team name, members, title, etc.
- Text body: Arial font. Size 11. Single-spaced. Normal margins.
- Use space wisely. If large (or many) tables/figures are necessary for a section, attach them as an appendix and reference it in the text *(i.e., see Figure 3)*.
- Each table/figure must be properly captioned or titled. Plots should include titles, axes labels, etc. **Presentation counts.**
- Any references should be properly cited.
- No R code should be present in the report. R scripts will be submitted separately.

Final report deliverables (3 items)

1. Final report will be submitted as a PDF. Name the file after the team name.
2. R script will be submitted as an R file. Name the file after the team name. Code should be clean and coherent for my viewing. Include appropriate comments to identify sections and purpose. Do not include unnecessary code that was not used for data cleaning, model building, or analysis.

3. Dataset will be submitted as a CSV file. Name the file after the team name.

Peer Evaluation SUBMISSION LINK

- Each individual will briefly describe the contributions of his/her teammates and evaluate each with a score of 0-3 pts based on their effort.
- This score will be factored into the course participation grade (5% of total grade).
- This is to be submitted individually. Content of your evaluation WILL NOT be known to your peers.