ISE/DSA 5103 Intelligent Data Analytics

Course Project

Requirement details

In teams of 3 to 4, define a data-intensive problem to explore and solve using a variety analytics techniques. The problem should be of sufficient complexity to challenge you. You are encouraged to use any techniques from this class *and/or* any additional techniques within the field of data science and analytics. You must submit a brief project proposal – multiple teams will be allowed to work on the same project.

Problems may be based on current research you are pursuing, problems from industry (e.g., from your place of employment), from analytics competition websites, or other public data sources. Do not use "tutorial" data sets or competitions with significant R code already available. Additionally, while using/reading/learning from other's work is valuable and permitted, the majority of the work should be conducted by your team. If you use someone else's code (e.g., from a notebook, etc.), cite the source.

Individual components of your course project include the following:

- Proposal and team formation
- Initial data analysis (12%)
- Initial draft (12%)
- Peer-review critique (6%)
- Presentation (35%)
- Final report (35%)

See course syllabus and website for individual component due dates.

Component details

Proposals: As a team, submit to the course website the following:

- Name of all team members
- Brief description of the data, problem, and proposed solution approach (e.g., regression, classification, clustering, association mining, etc.) – (half page is fine)
- URL of the problem/data website (if using public data)

Initial data analysis: As a team, submit to the course website the following:

- Initial exploratory analysis. Should be at least 50% complete and must contain certain content:
  - Data quality report
  - Bullet point explanation of expected approach for dealing with missing values, outliers, skews, factors, and/or other data issues
  - At least two *important* visualizations

Initial draft: As a team, submit to the course website the following:

- The first draft of the project report. Should be at least 75% complete and must contain certain content:
  - Introduction to the problem

  - Data description and exploratory analysis (e.g., updated from the initial data analysis task)
  - Description of modeling approach
  - Initial results
- Include the name and *emails* all team members on the report.
- This draft will be used in a peer-review process – evaluated by your classmates to help you improve your work. If the initial draft is not at least 75% complete, the grade will be penalized.

Peer-review: Each team member will submit to the course website one or more project report critique(s). A rubric to help you evaluate project reports will be provided.

Presentations: Each team will summarize their work in presentation form.

- Presentations should between 5 and 10 minutes long (i.e., about 10 slides – make the *important* points!)
- Presentation should be divided equally among team members
- The slide presentations and *recorded video* will be made available to the entire class
- Grading will be based on quality/clarity of content and the ability to present complex material under a given time/page limit
- Every team member should be ready to answer questions about any part of the project

---

Final project report: As a team, you will submit to the course website, the following:

- One PDF file 12-15 pages max + appendix (unlimited) – Note: exceeding the page limit will result in significant grade penalties.
- Complete, commented, and "compilable" R script

The final project report shall include the following titled sections:

- **Executive Summary**: 1 page
  - Concise problem statement
  - List of major concerns/assumptions (if any)
  - Summary of findings
  - Recommendations
- **Problem background**
  - Problem description, context, background
  - Data description
  - Exploratory data analysis – the highlights; not the kitchen sink
- **Methodology**
  - Feature selection, engineering, missing value imputation, outlier processing, etc.
  - Modeling choices
  - State model validation plan (e.g., 5-fold CV)
- **Results**
  - Model performance summary
  - Key findings of analysis!

- **Conclusion**
  - Summary of problem, approach, findings
  - Key issues, limitations, etc.
  - Final recommendation
- References (optional, does not count toward page limit)
- Appendix (optional, does not count toward page limit):
  - Data visualizations, tables, transformations, etc. which support the work, but are not of primary importance
  - Important code excerpts or algorithms used / developed (if any).

Do not write the final report as a narrative, e.g., "we tried X, but it didn't work, so then we tried Y, and it worked a little better, finally, after some pizza, taking a nap, and getting our heads together we all decided on Z" No! This is a scientific/industry paper, not a diary entry! Give me facts, don't use "we" in the manuscript unless you absolutely have to.

Possible sites for data/problems

Competition websites with data and problems

- crowdanalytix.com
- KDD Cup (sigkdd.org/kddcup/index.php)
- Kaggle.com – *only if there are little to no R notebooks available!*
- tunedit.org/challenges
- www.kdnuggets.com/competitions/past-competitions.html

Websites with many datasets

- Open Data New York: `https://opendata.cityofnewyork.us/`
- Analyze Boston `https://data.boston.gov/`
- Seattle Open Data `https://data.seattle.gov/`
- Oklahoma Open Data `https://data.ok.gov/`
- Data.gov `https://www.data.gov/open-gov/`
- Open Data Columbia `https://www.datos.gov.co/`
- World Bank Open Data `https://data.worldbank.org/`
- UCI Machine Learning Repository `https://archive.ics.uci.edu/ml/index.php`
- Open Data on AWS `https://registry.opendata.aws/`
- National Center for Education Statistics `https://nces.ed.gov/`
- American Economic Association `https://www.aeaweb.org/resources/data/us-macro-regional`
- Registry of Research Data Repositories `https://www.re3data.org/`

Websites for specific data

- Yelp Open Dataset `https://www.yelp.com/dataset`
- Million Song Dataset `http://millionsongdataset.com/`
- Global Burden of Disease Study `http://ghdx.healthdata.org/gbd-2016`
- MIT Lab for Computational Physiology `https://mimic.physionet.org/`