DSA/ISE 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 for instructor approval before beginning your work.

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 you and your team. If you use someone else's code (e.g., from a notebook, etc.), you must cite the source. Failure to do so could result in one or more OU academic integrity violations.

In every aspect of this project you must be *clean, clear, concise* and *professional* with the quality of the deliverable (e.g., formatting, spelling, grammar, good quality images, etc.) For example: do not include poor quality screen shots, pixelated graphics, or tables/figures/text that do not fit well on the page.

Your final report should be written in the style of a scientific article and include appropriate citations and references if used.

Individual components of your course project include the following:

- Proposal and team formation (2%)
- Initial data analysis (10%)
- Draft (10%)
- Presentation (38%)
- Final report (40%)

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. *This should be 25% to 50% complete* and must meet the following expectations:
  - Title page with project name, names of all team members, team number, date, class information

  - (10%) One or more paragraphs explaining expected approach for dealing with missing values, outliers, skews, factors, and/or other data issues
  - (75%) Five or more *important* visualizations. Provide a brief explanation of the *insight* you found useful from each visualization, i.e., *why* is it important?
  - (15%) Appendix: Data quality report
  - 5 page limit (Appendices and title page do not count toward page limit)

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

- The first draft of the project report. Should be 50% to 75% complete and must contain certain content:
  - (20%) Introduction to the problem
  - (20%) Data description and exploratory analysis (e.g., updated from the initial data analysis task)
  - (20%) Description of modeling approach
  - (30%) Initial results
  - (10%) Team allocation/contribution page: A one oro two sentence description, per team member, of the contribution (past and future) of that team member to the draft, e.g.
    - Johnny led the exploratory data visualization component, he will lead the unsupervised modeling in the next phase.
    - Ursula led the code testing and review sessions and will write the conclusion.
    - Nila wrote the background and led the supervised modeling.

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

- Presentations should between 10-15 minutes long (i.e., about 10 slides – make the *important* points!)
- Include all team member names on the title slide; also include the group number.
- 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
- You must have a concluding slide that summarizes the work and key insights, points out critical assumptions or limitations, and discusses the potential impact or implementation issues with your work.

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

- One PDF report file, 10 (minimum) to 15 pages (max) + references (unlimited) + appendix (unlimited); the title page does not count toward the page limit.
  - Exceeding the page limit will result in significant grade penalties.
  - Going under the page limit may also result in penalties.
  - Font size 10, 11, or 12. Font such as Aptos, Calibri, Times New Roman, Palotino Linotype, etc.
  - Stylistically this should be written as a scientific paper. Number and label all figures and tables. You can use any scientific reference style that you want (just be consistent)
  - Every figure or table in the report body must be referred to and discussed appropriately in the manuscript text.
  - Be consistent with your fonts, line spacing, figure sizes. Make sure all figure information is legible.

  - In most cases, R code should not be included in the PDF file at all. If R code is included, it must be exceptionally important to the findings or it will result in a grade penalty.
  - R console output in terms of errors or warnings or other such messages should not be included in the PDF file. Inclusion will result in a grade penalty.
- Complete, commented, and “compilable” R script

The final project report shall include the following titled sections of approximate grade weight:

- Title page – include project name, team member names, group number, instructor name, course name, semester, and date (this does not count toward the page limit)
- (20%) **Executive Summary**: 1 full page summarizing your work. Do not go over one page.
  - Concise problem statement
  - List of major concerns/assumptions (if any)
  - Summary of findings (include quantifiable information, e.g., $CVR^2 = 0.984$)
  - Recommendations
- (15%) **Problem background**
  - Problem description, context, background
  - Data description
  - Exploratory data analysis – the highlights; not the kitchen sink
- (15%) **Methodology**
  - Feature selection, engineering, missing value imputation, outlier processing, etc.
  - Modeling choices
  - State model validation approach (e.g., 5-fold CV) — note: Do not simply do a single holdout (80/20, 70/30, etc.) for anything unless you have a very good, *defensible* reason for doing so!
- (25%) **Results**
  - Model results and performance summary
    - ∗ for supervised learning models, this is pretty straightforward
    - ∗ for unsupervised learning or other model types, you need to convince the reader of the validity of your work
  - You should provide a detailed analysis of your best model (e.g., residual diagnostics, coefficients, feature importance, hyperparameter tuning visualizations, confusion matrices, etc.)
  - Key findings of analysis! Tie this back to your models, assumptions, problem description, etc.
- (25%) **Conclusion**
  - Summary of problem, approach, findings – yes, this is a bit redundant, but this is your chance to help the reader really connect the problem statement with your work.
  - Key insights, points out critical assumptions or limitations, etc.
  - Final recommendation and potential impact of the work – take into account the performance of your models! For example, if your accuracy was 0.54, don’t say ”Company X can use this model to improve their profits next quarter”, such a model is not very good, so they probably can’t do that!
- 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/`