Advanced Data Mining – Team Project

Task: Propose and execute a research project using data mining techniques as a team of 2 participants.
Contribution to ADM: 60%

Deliverables

The project comprises of 4 components submitted in 2 parts as outlined below.

Component	*Weight*	*Description*	*Due*	*Submission #*
Proposal	33%	Detailed overview of the project	Wk 3	Submission 1
Paper	64%	12-page IEEE format report	Wk 13	Submission 2
Distribution of work	N/A	Overview of contribution of each team member to the project	Wk 13	Submission 2
Code	N/A	.zip of all code used in the project	Wk 13	Submission 2

First Submisson – Project Proposal 20%

The first submission is a business case outlining the proposed research project. This submission should focus on:

- The motivation for the research – why is this an important question to explore either from a business point of view or a research perspective?
- What business or research question will this analysis allow you to answer?
- Who are the potential beneficiaries of your research (academic, business or both)?
- What related work has been conducted in the field and how this has influenced your choice of topic?
- How you intend to carry out the work?
    - What process methodology do you intend to follow?
    - What models do you intend to build?
    - What metrics will you use to evaluate the models you build?

Note: You have full freedom to choose any appropriate method(s) and model (s) but must use appropriate literature to defend your choice(s).

Each team member is required to submit a project proposal.

Please follow the guidelines below when preparing the first submission.

Section	*Length Limit*	*Description*
Motivation	300 Words	Motivate your choice of topic
Research Question	15 Words	What question do you plan to answer
Potential Beneficaries	2 paragraphs	Who will benefit from your proposed research?
Related Work	2 pages	Briefly review 8-12 papers that inform your choice of topic.
Proposed Methodology	1 page	Briefly describe the process methodology do you intend to follow, the models you intend to build and how you will evaluate your proposal
Data Source(s)	5 lines	List the sources of data the project is considering
Bibliography	Exempt from length limits	

Grade (of 20)	*Criteria*
0	One or more of the required sections are missing. – thus cannot provide meaningful feedback
1-9	Project lacks: • direction, • perceivable novelty, and/or • complexity
10-20	Proposal suggests that the project will meet the module requirements. Grade differentiation will be based on general quality of the individual sections and coherence of the proposal.

Second Submission - Conference-style Report – 40%

The second submission is a conference-style report describing your project work. The overarching goal of the project is to leverage methods learned in the module and the course as a whole to execute a significant data mining study. You are expected to source your own data sets for this project. See https://www.dataquest.io/blog/free-datasets-for-projects/ for examples of data available for research projects.

Submitted papers will be assessed based on their novelty, technical quality, potential impact, insightfulness, depth, clarity, and reproducibility. Code and data sets are to also be submitted with the paper. Algorithms and resources used in a paper should be described as completely as possible to allow reproducibility. This includes experimental methodology, empirical evaluations, and results. The reproducibility factor will play an important role in the assessment of each submission.

Suggested Paper Structure

- Title
- Author
- Abstract: 150-250 words
- Introduction: remainder of 1st page (+ up to 1 column). Should motivate the work, present and discuss the research question(s) / objective(s) of the paper and (optionally) provide a concise overview of the following sections (max 1-2 lines per each).
- Related Work: ideally 2 pages (20 or more references in total) – this should not only summarise related work, but also critically evaluate (positive and negative aspects) related work with respect to the topic question, i.e. how well/badly does the related work artefact answer this question, what aspects are useful to consider, what are the limitations... Also discuss here are any foundational papers that substantiate your study design or upon which you build.
  Note: Introduction and Related Work aspects can be reused from the Project Proposal if they are of good quality. It is likely you will have found new literature in the course of your project and this may be added to the related work in the paper. Preparing a good initial project proposal will save you significant time in preparing this aspect of the paper.
- Data Mining Methodology (can be named differently): how have you approached answering your question. Additional (technical) details can also be discussed here. Essentially, you should recount how you applied either KDD or CRISP-DM to answer your research question(s). Specifically you should describe how you prepared your data (data cleaning and transformation or encoding, feature selection methods applied), give some details about the dataset (descriptive statistics about independent variables, number of missing values and how they were handled, comment on the class balance, visualizations should also be used to describe the dataset), describe how you prepared samples for building your models and outline how you developed your model(s) or ensemble model(s).
- Evaluation/results – how have you used your method(ology) to answer the research or business question (evaluation methodology), how do you know your approach is good, results of your evaluation, and a discussion on their implications / impact. If you have to parameterise part of your approach how have you done that, and why were these choices made, and what impacts can different parameterisations have on your results? You should also discuss the results in detail in this section: what are their implications? What do they show / not show? Etc.
- Conclusions and future work: summarise your findings, and discuss limitations / extensions that were you to have more time, you would do next to improve / extend your study. Summarise the (partial) answer to the research question(s) at a high level, and note the contribution to knowledge the paper has made.
- References
- Appendix A: Overview of contribution of each team member to the project

Each team member is required to submit a conference-style report.

Paper Formating and Length

Papers **must** follow the IEEE conference format and should be 12 **double column** pages in length (this includes all figures and references). For this exercise IEEE style referencing, **not Harvard referencing**, should be used.

Papers over 12 pages (even if it is only 1 word) will be subjected to a 5 percentile point penalty, i.e. the maximum mark for the paper will be 45%.

Word and LaTeX templates are available here:
http://www.ieee.org/conferences_events/conferences/publishing/templates.html

Space saving tips:

- Never have a line less than half full at the end of a paragraph. Almost any paragraph can be rewritten so that this is not the case!
- Graphs / Flow diagrams / Tables are easy to do sub optimally – draw them properly and decide if they really need to be as big as they are. Or if they really should span both columns.
- Sub figures (e.g. 3 graphs as one figure prefixed a, b c that span both columns) usually are fairly space efficient.
- The LaTeX template is significantly cleverer than the Word one, and will do more work to save space.
- In LaTeX paragraph spacing is heavily optimised. This also means that cutting out a line or two before a new section can cause paragraph spacing to be recalculated thus saving significant space.

Submission Process

Papers should be submitted via Moodle and Turnitin.

Suggested time plan

Week 1	Teams formed
Week 3	Project proposal submitted **Submission Date and Time:** **07 October 2018, 23:55**
Week 4	Data curated and cleaned (at least started). Methodological approach slowly coming together.
Week 6	Any implementation and/or feature engineering needed for analysis complete. Exploratory data analysis done.
Week 8	Initial analysis complete
Week 10	Initial results / answer(s) to question(s)
Week 11	Most results and evaluation aspects complete
Week 13	Draft of paper ready for submission **Submission Date and Time** **16 December 2108, 23:55**

Criteria % Weight	> 80%	H1	H2-1	H2-2	Pass	Fail
Objectives and Motivation 10%	Challenging project objectives are well presented, met and thoroughly motivated as well as discussed.		Appropriate project objectives are well presented, met and thoroughly motivated as well as discussed.	Appropriate project objectives are presented, mostly met and motivated as well as discussed.	There are clear objectives, which are at least partially met.	Cannot discern project objectives, and/or if project objectives were met.
Discussion of related work 10%	Discussion of related work is excellent, and the choice of papers to discuss excellently situates the project within the literature.	Discussion of related work is v. good, and the choice of papers to discuss excellently situates the project within the literature	Discussion of related work is good and the choice of papers to discuss well situates the project within the literature	Discussion of related work is appropriate and the choice of papers to discuss well situates the project within the literature	Discussion of related work is appropriate, and the choice of papers appropriately situates the project within the literature	Discussion of related work lacks depth, and/or the choice of papers seems somewhat arbitrary.
Data Mining Methodology 40%	It is hard to find fault in the approach. All key decisions are justified with appropriate literature. The project extends beyond simply applying models to complex datasets, and also competes with or outperforms other relevant works.	All stages of KDD/CRISP-DM are rigorously applied. All key decisions are justified with appropriate literature. The project extends beyond simply applying models to complex datasets, and also attempts to compare with other relevant works.	All stages of KDD/CRISP-DM are rigorously applied. Some minor shortcuts or errors may be present. Most key decisions are justified with appropriate literature. The project extends beyond simply applying models to datasets, and also attempts to compare with other relevant works.	All stages of KDD/CRISP-DM are appropriately applied, but the general approach lacks some depth. There may be some mistakes in the approach taken. Key decisions are justified with appropriate literature, but more depth is needed. The project extends beyond simply applying models to a dataset.	All stages of KDD/CRISP-DM are appropriately applied, but the general approach lacks depth. There may be significant mistakes in the approach taken. Some key decisions are justified with appropriate literature, but more depth is needed. The project may lack depth, complexity, or uses only toy datasets.	KDD or CRISP-DM not appropriately followed and/or applied. The approach taken may also be hard to discern. Key decisions are not justified or substantiated with appropriate literature The project may also lack depth and/or complexity.
Contribution and Findings 20%	The project has publication potential	The project evidences a novel contribution but may require some additional work to round off the contribute and/or generalisability of the findings.	The project evidences a novel contribution but requires additional work to round off the contribute and/or generalisability of the findings.	The project has a clear non-arbitrary contribution but may lack some depth and/or rigor in either the findings, or interpretation of the results.	The project findings have value but may be somewhat arbitrary and/or lack generalisability. There may be some small mistakes in the interpretation of the results.	It is not clear what the relevance of the findings are. There may be some fundamental mistakes in the interpretation of the results.
Conclusions and FW 10%	Insightful conclusions, which appreciate limitations and implications of the project. Well-conceived and thought out future work is discussed and presented.		Implications and limitations well understood. Discussion also correctly highlights key takeaways. Appropriate future work is discussed and presented.	Implications and limitations well understood. Discussion also correctly highlights key takeaways. Future work lacks depth and creativity, but is appropriate.	Implications and limitations not well understood. Future work lacks depth and creativity, but is appropriate.	Implications and limitations not understood. Future work seems arbitrary or inconsistent with project findings
Presentation 10%	Well written, with no (large) language errors. All figures are well conceived and readable. The IEEE template is adhered to. Report does not exceed the length limits. References are appropriately and correctly used.		Main document has a few language and/or style errors. Figures are well presented. IEEE template and length limit are adhered to. References are complete, and correctly used.	Main document is readable with some language and/or style errors. Some figures may be hard to read or presented in a suboptimal manner. IEEE template is largely adhered to. References are mostly complete and correctly used.		Littered with typos, and/or poor use of English. IEEE template may have been broken. Figures may be hard to read. References (if any) are probably incomplete.

Note: failure to contextualise, research and utilise current data mining approaches, applications and technologies in order to provide strategies to address processing of datasets with a variety of characteristics caps the paper at 39%