QBUS3820
Machine Learning and Data Mining in Business
Semester 1, 2020

Marking Scheme and Rubric for the Group Project

1. Marking Scheme

Problem formulation	5 marks
Exploratory data analysis	10 marks
Feature Engineering	20 marks
Model building	25 marks
Model validation	5 Marks
Model evaluation	5 marks
Data mining	20 marks
Writing and presentation.	10 marks
Total	**100 marks**

Marks can be deducted in some cases: please refer to Section 3.

2. Rubric

Preparation. You read and understood the assignment description and requirements and are aware that this is part of the assessment. You understand that there is no single right solution to complex problems, and that experimenting with different approaches and using the data to discover what works best is natural and desirable in this type of analysis.

Problem formulation. The report includes a discussion of the context for the analysis, the problem and questions/hypotheses to be addressed, and how you plan to measure the success of your proposed solutions.

Data processing. Your report includes an informative general description of the dataset. You make sure that the dataset is free of errors and correctly processed for your analysis. You describe the data processing steps concisely.

Exploratory data analysis (EDA). You study the variables individually as well as the bivariate relationships between the predictors and the response using appropriate figures and descriptive statistics. You note any characteristics of the data that are relevant for model building. You note the presence of outliers and any other anomalies that can affect the analysis. You explain the relevance of the EDA results to your subsequent feature engineering and model building. Your EDA section in the report is concise, leaving additional figures and tables to the appendix if needed.

Feature engineering. You describe and explain your feature engineering process. You understand and consider the range of feature engineering options listed in the lecture slides and potentially other resources. Your choices are driven and justified by data analysis, understanding of the material, domain knowledge, logical reasoning, the requirements of different learning algorithms, and trial and error (if necessary). Data-driven choices are better than opinion-based choices.

Model building. You clearly describe and justify the models, methods, and algorithms used in your analysis. The choice of methods is logically related to the assignment requirements, the substantive problem, underlying theoretical knowledge, and data analysis. This may involve systematic trial and error, but the report should focus on your final solutions.

Model validation. You obtain a high standard of predictive accuracy in line with what is achievable with the methods discussed in the course and the level of experience expected from students taking this unit. You correctly present and interpret the public leaderboard results according to the instructions of the assignment.

Model evaluation. This refers to the Kaggle Private Leaderboard (test) results, not the report. You will be awarded 5 marks for participation as long as your submission is correct and you have made a genuine effort to generate good predictions to the best of your abilities.

Data mining. You correctly interpret the results and discuss how they address the substantive question. The reasoning from methodology and results to your conclusions is logical and convincing. Your analysis takes statistical variability into account, where appropriate. You make no claims for which you have no evidence. You do not make statements that imply causation when discussing association. You explicitly acknowledge when limitations of the data or methods lead to uncertainty about your answer to a substantive question.

Assumptions. You report and check any crucial assumptions behind your analysis. You clearly recognise when an assumption is not satisfied or questionable. Some problems may be unfixable given the available data and methods. In this case you can identify what additional information or methodology could allow you to fix these problems.

Writing. Your writing is concise, clear and free of grammatical and spelling errors. You use appropriate technical terminology. Your paragraphs and sentences are well connected and follow a clear logic. There is a distinction between the essential parts of the report and less important material (use the appendix). Your text refers to meaningful variable names. If you use an abbreviation or label, you first have to define it.

Layout. Your report is professionally presented and formatted, as if it had been prepared for a client in an actual job. The report is well laid-out, with clear divisions between sections and paragraphs. Your report looks uncluttered. You display at least an intuitive understanding of basic design principles.

Tables. Your tables are well formatted a clear layout. The tables have informative row and column labels. The tables are as much as possible easy to understand on their own (in the real world, a significant part of your audience will skim-read by going straight to the tables). The tables do not contain information which is irrelevant to the discussion in your report. Your table is not an image. The tables are placed near the relevant discussion in your report. There is no text around your tables.

Figures. Your figures are professionally presented. Your figures have informative titles, captions, labels, and legends. The figures are placed near the relevant discussion in your report. Your figures have good definition and were directly saved from Python into an image file format (no screenshots). There is no text around your figures.

Numbers. All numerical results are reported to suitable precision (typically no more than three decimal places, in some cases fewer).

Referencing. You follow referencing guidelines and rules of the university.

Python code. The main text of your report is entirely free of Python code. The code is presented in a neat and compact way. The code uses meaningful variable names and can be easily followed by someone with training in Python and statistics. Your code have comments that clearly indicate which parts correspond to which sections of your report. You explicitly acknowledge when you borrow pieces of code from sources other than the lecture materials.

Reproducibility. Your results are reproducible. Your report provides enough information for the reader to reproduce the steps that lead to your final results. Someone should be able to run your code and reproduce all the results that appear in your report.

3. Deductions

Up to 10 marks will be automatically deducted from the writing and presentation item of the rubric in the following cases.

The report is poorly written.	-10 marks
The report has an excessive number of grammatical or spelling mistakes.	-5 marks
There is an excessive number of abbreviations or labels that the reader may be unfamiliar with.	-5 marks
The report is disorganised/has a poor layout.	-5 marks
The tables are difficult to read, for example due to poor layout or labelling.	-10 marks
A table is an image.	-5 marks
Numbers are not appropriately rounded.	-5 marks
The figures are difficult to understand due to poor layout or lack of labelling.	-10 marks

The following penalties also apply without limit.

No participation in the Kaggle competition.	-20 marks

The group or a member of the group cannot be immediately identified on Canvas, the first page of the report, and/or Kaggle.	Treated as a late submission.