

Task B “Moneyball” Instructions

Preparation. You read and understood the assignment requirements and are aware that this is part of the assessment. You understand that machine learning is grounded in rigorous logic and theory that should inform your practical analysis. You understand that there is no single right solution and that trying different approaches and discovering empirically what works best for a particular problem is natural and desirable in this type of analysis.

Business context and 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. You make sure that the dataset is free of errors and correctly processed for your analysis. You handle missing values and other issues appropriately. You describe the data processing steps in a clear and concise way.

Exploratory data analysis (EDA). Your report describes your EDA process, presenting only selected results. You studied key variables individually and pairwise using appropriate figures and descriptive statistics. You note any features 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 modelling. 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 process for feature engineering. Your choices are justified by data analysis, domain knowledge, logic, and/or trial and error. Data-driven choices are better than opinion-based choices.

Methodology and modelling. You clearly describe and justify the models, methods, and algorithms 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. Your methodology pays attention to statistical variability. You report all crucial assumptions and check them as relevant via formal and informal diagnostics. 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.

Analysis and conclusions. Your analysis is rich. 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. You are not misled by overfitting. Your analysis pays attention to statistical variability. You make no claims for which you have no evidence. You do not make statements that imply causation when discussing associations. You explicitly acknowledge when limitations of the data or methods lead to uncertainty about your answer to the substantive question.

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

Report. Your report is well organised and professionally presented and formatted, as if it had been prepared for a client later in your career. There are clear divisions between sections and paragraphs.

Tables. Your tables are appropriately formatted and have a clear layout. The tables have informative row and column labels. The tables are as much as possible easy to be understood 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 easy to understand and have informative titles, captions, labels, and legends. The figures are well formatted and laid out. The figures are placed near the relevant discussion in your report. Your figures have appropriate definition and were directly saved from Python into an image file format. There is no text around your figures.

Numbers. All numerical results are reported to four-decimal point.

Referencing. You follow the Harvard Referencing System.

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. Someone should be able to run your code and reproduce all the results that appear in your report. Your code has 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 and tutorial materials.