MSc in Data Analytics 2018-2019

CA660P Statistical Data Analysis

Assignment (worth 20% of final result)

Deadline 9th December, 2018

Instructions:

Students should work in groups (ideally in groups of 5) and produce a single report for the group, with the names and Student IDs of contributors clearly indicated on the top sheet. A statement should also be included on the top sheet to indicate that group-members have contributed equally (or reasonably so) to the work. The mark given for the assignment will apply equally to each member of the group, unless an individual student fails to make a contribution. Plagiarism between different groups will be strictly penalised.

Report:

Objective

The aim is to provide a report, describing a good exploratory analysis by addressing what the ‘group’ feels are the key questions that the data pose. The objectives should be to describe methods and solutions and to draw conclusions in a clear and concise way.

Contents

The report is expected to be a coherent description of analyses, results, and conclusions using text and annotated tables/figures/diagrams as appropriate.

Detailed code or listing of software commands should not be necessary, but students must indicate when they have used or adapted available code and where they have drawn on previously available analyses or sought to extend these. Since the dataset used is publicly available, it would be unrealistic to expect that no-one has looked at it before!

If a group feels that inclusion of additional/adapted/corrected code is required to make a specific point, it can include a disk. The report and inclusions (e.g. analyses) should be typed. Large amounts of printout should not be appended. If further details are necessary to make a point, this should be indicated clearly in the text and the relevant appendix cited, which should then appear in a readily accessible and annotated format on the accompanying disk. Any printout which is included without annotation, (i.e. clear description of what it purports to show), either in hard or electronic copy, will not be marked.

Structure

The report should have a clear structure containing sections providing an overall summary of your main findings as well as more detailed sections on the objective of the analysis, the analysis process and finally the conclusions (see below for more detail).

Length

The report is not expected to exceed 10 A4 pages in length, inclusive of any analyses reported in detail, background references or (short) code.

Background:

The finance sector has been aggressively pursuing the adoption of advanced data analytics methods to improve services and maximise gains. In this assignment we will use the bank marketing dataset provided by the University of California Irvine. The data contains information related to a direct marketing campaign of a Portuguese banking institution and its attempts to get its clients to subscribe for a term deposit. The data in its raw form is available at the following link: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing#. However, this data consists of several missing values and some attributes that are beyond the scope of this module. Therefore, a pre-processed version of this dataset is made available to you called ‘bankdata-cleaned.csv’ that you are expected to use for the purpose of this assignment. A ‘readme.txt’ file is also available to provide details about this dataset.

The purpose of this assignment is to show how the data can be used to extract meaningful information and to draw helpful conclusions about term deposit subscriptions. It can be used as a basis for comparison, modelling and testing, integrating and combining these approaches to build an overall picture of the most useful information provided by the data.

While you can be reasonably well assured of the quality of the data you should carry out your own data integrity checks (for example, you should produce an account of the data and you may want to produce statistical summaries for the variables). Your report should provide a clear statement of what you set out to do and why, the basis on which you selected variables, indicators, data items, the completeness of these data for your aims and a description of the results of the various stages of your investigation. All information sources (including previously produced reports, processed tables, etc.) must be acknowledged. State your conclusions clearly and indicate limitations and suggestions for future work. Key references should be included, particularly where these have been heavily drawn upon.