**INMR95 Assessment Details**

Module Convenor: Dr Markos Kyritsis

Email: [m.kyritsis@henley.ac.uk](mailto:m.kyritsis@henley.ac.uk)

**Assessment Details**

**IMPORTANT NOTE:**

The final report must be submitted to the Turnitin submission **as a word document** **by itself** (i.e., not as a zip file with other stuff such as source code).

The source code, new generated data, and anything else such as e.g. R shiny app (optional) will be **submitted separately in a different link on Blackboard.** This link is contained in the folder names “Submission for Supporting Files” under the folder “Assessment”.

**Section 1 – Descriptive Analytics**

**Descriptive statistics are brief informational coefficients that summarize a given data set, which can be either a representation of the entire population or a sample of a population. Descriptive statistics are broken down into measures of central tendency and measures of variability (spread). Measures of central tendency include the mean, median, and mode, while measures of variability include standard deviation, variance, minimum and maximum variables, kurtosis, and skewness.**

Introduction

Case study: **Student satisfaction** is a KPI for most, if not all higher education institutes. There are a range of reasons why students may or may not be satisfied with their courses. The Turkiye Student Evaluation Datasets gives us a small insight into the complexities that drive student experience;

you have been hired by a company called HigherEdCo ltd. as a higher education consultant to perform several multivariate analyses that will indicate the factors that impact student experience (according to the data collected). **Your analysis should be approached critically, and variable as well as method selections should be justified. You MUST reduce the dimensionality of this dataset.**

The dataset you will be working with is the Turkiye Student Evaluation Data Set (Gunduz & Fokue, 2013). The dataset is made up of the following variables:

学生体验满意度

|  |
| --- |
| instr: Instructor's identifier; values taken from {1,2,3} 三个导师  class: Course code (descriptor); values taken from {1-13} 13个课程  repeat: Number of times the student is taking this course; values taken from {0,1,2,3,...} 重修次数  attendance: Code of the level of attendance; values from {0, 1, 2, 3, 4} 出勤率  difficulty: Level of difficulty of the course as perceived by the student; values taken from {1,2,3,4,5} 课程难度  Q1: The semester course content, teaching method and evaluation system were provided at the start. 课程内容与教学方法  Q2: The course aims and objectives were clearly stated at the beginning of the period. 教学目标  Q3: The course was worth the amount of credit assigned to it. 课程性价比  Q4: The course was taught according to the syllabus announced on the first day of class. 教学大纲  Q5:The class discussions, homework assignments, applications and studies were satisfactory. 作业  Q6: The textbook and other courses resources were sufficient and up to date. 教材  Q7: The course allowed field work, applications, laboratory, discussion .实地操练  Q8: The quizzes, assignments, projects and exams contributed to helping the learning.对学习有帮  Q9: I greatly enjoyed the class and was eager to actively participate during the lectures. 学生主动性参加课程  Q10: My initial expectations about the course were met at the end of the period or year. 课程满足学生期望  Q11: The course was relevant and beneficial to my professional development. 课程相关并有益处  Q12: The course helped me look at life and the world with a new perspective. 课程帮我看到生活  Q13: The Instructor's knowledge was relevant and up to date. 导师的知识新鲜  Q14: The Instructor came prepared for classes. 导师有准备  Q15: The Instructor taught in accordance with the announced lesson plan. 导师根据计划教学  Q16: The Instructor was committed to the course and was understandable. 导师的贡献可以理解  Q17: The Instructor arrived on time for classes. 导师准时到教室  Q18: The Instructor has a smooth and easy to follow delivery/speech. 导师的演讲能力抓人  Q19: The Instructor made effective use of class hours. 导师课堂时间利用率很高  Q20: The Instructor explained the course and was eager to be helpful to students. 导师渴望帮助学生  Q21: The Instructor demonstrated a positive approach to students. 导师解释学习方法  Q22: The Instructor was open and respectful of the views of students about the course. 导师对评价开放  Q23: The Instructor encouraged participation in the course. 导师鼓励课堂参与  Q24: The Instructor gave relevant homework assignments/projects, and helped/guided students. 导师给作业并且帮助学生  Q25: The Instructor responded to questions about the course inside and outside of the course.  导师无处不回应  Q26: The Instructor's evaluation system (midterm and final questions, projects, assignments, etc.) effectively measured the course objectives.  导师的评价系统是有效衡量了课程目标  Q27: The Instructor provided solutions to exams and discussed them with students.  导师提供了对考试的解决方式并与学生讨论  Q28: The Instructor treated all students in a right and objective manner.  导师对待学生的方式很不错 |

It’s up to you to choose your independent and dependent variable(s), as well as the tests you will run. However, everything you do needs to be justified, i.e., you need to explain why you chose to use that particular test, why you treated a certain variable as e.g., categorical, and why you transformed variables (if applicable). In short, a good project will critically analyse the results obtained and identify its limitations. The more detailed and exhaustive your analysis, the more likely you are to score a high grade (see marking scheme). Make sure to include figures and tables to support your findings.

**Expected Project Output**

In the end you need to submit a section with the following headings:

1. Introduction (briefly what your aim was for the analysis and your research question)
2. Process (what types of statistical testing did you use to answer the research question and the rationale for using said methods).
3. Results (the results of all the analyses, including figures, tables, and test outputs). The output needs to be written for an **academic audience**.

You must also submit:

* Your new csv file with any new variables you extracted/modified from the data. You must name this **exploratory.csv**
* Your R script that shows, with comments, step by step the process you took to analyse the data. You must name this **exploratory.R**
* **Anything else that you feel is relevant is welcome (but not required)**

**Section 2 – Predictive Analytics**

**Predictive analytics is the process of using data to forecast future outcomes. The process uses data analysis, machine learning, artificial intelligence, and statistical models to find patterns that might predict future behaviour.**

Case study: You have been hired as a consultant to provide data-driven recommendations to the marketing department of the German-Hellenic bank. The bank has supplied you with anonymised data (the data we will be using has been supplied by Moro et al. (2014), and can be found in the UCI website).

Here is a list of the variables:

|  |
| --- |
| Input variables:  # bank client data:  1 - age (numeric)  2 - job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')  3 - marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)  4 - education (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')  5 - default: has credit in default? (categorical: 'no','yes','unknown')  6 – balance: Account balance  7 - housing: has housing loan? (categorical: 'no','yes','unknown')  8 - loan: has personal loan? (categorical: 'no','yes','unknown')  # related with the last contact of the current campaign:  9 - contact: contact communication type (categorical: 'cellular','telephone')  10 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')  11 - day\_of\_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')  ~~12 - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.~~  # other attributes:  13 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)  14 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)  15 - previous: number of contacts performed before this campaign and for this client (numeric)  16 - poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')  # social and economic context attributes  Output variable (desired target):  17 - y - has the client subscribed a term deposit? (binary: 'yes','no') |

**Note that variable 12 should be discarded**. The original data set has outcome variable 17 as the desired output. However, you do not necessarily need to focuson this variable. You are expected to explore other relationships in the data and present interesting findings to your client (hint: look at balance for example). You should build **multiple models** for comparisons, but **present two final models** on two different outcome variables. All actions taken need to be critically analysed and justified. The more detailed and exhaustive your analysis, the more likely you are to score a high grade (see marking scheme). Your outcome variables can be **categorical, continuous, or a mix of both (i.e., one model as a classification model, one as a regression model一个二元 一个回归).**

**Expected Project Output**

In the end you need to submit a section with the following headings:

1. Introduction (briefly what your aim was for the analysis, along with your research question)
2. Process (what types of models did you use to answer the research question and the rationale for using said modelling techniques).
3. Results (the results of the analyses and the models, including model performance and model comparisons). All the output needs to be written for an **academic audience**.

You must also submit:

* Your new csv file with any new variables you extracted/modified from the data. You must name this **analysis.csv**
* Your R script that shows, with comments, step by step the process you took to analyse the data. You must name this **analysis.R**
* **A R shiny app or anything else you feel is relevant (optional)**

**Section 3 – Prescriptive Analytics**

In this section you will f**orm data-driven recommendations** using your findings from Section 1 and Section 2 for your **two clients.** You can expect your audience to be a layman audience with little to no understanding of statistics and modelling. Therefore, unlike the results is section 1 and 2, your report needs to be written in such a way that a layman audience can understand it. Ultimately you need to make a **convincing argument** that states how your client should proceed based on the results of your findings. You are expected to use a critical approach by using the results obtained to both generate recommendations and identify limitations. You can include an interactive Shiny R app, which is optional but will increase your likelihood of delivering a more robust solution.

**Expected Project Output**

In the end you need to submit a document with the following headings:

1. **Client: HigherEdCo ltd**. - Executive summary
2. Aims and Objectives
3. Analysis
4. Recommendations
5. Limitations
6. **Client: German-Hellenic Bank**. - Executive summary
7. Aims and Objectives
8. Analysis
9. Recommendations
10. Limitations

**References**

Gunduz, G. & Fokoue, E. (2013). UCI Machine Learning Repository [[[https://archive.ics.uci.edu](https://archive.ics.uci.edu/)]]. Irvine, CA: University of California, School of Information and Computer Science.

S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014