SINGAPORE UNIVERSITY OF SOCIAL SCIENCES

ANL307e
Predictive Modelling

Group-based Assignment

January 2018 Presentation

GROUP-BASED ASSIGNMENT

This assignment is worth 44% of the final mark for ANL307e Predictive Modelling.

The cut-off date for this assignment is **20 February 2018, 2355hrs.**

This is a group-based assignment. You should form a group of **3 members** from your seminar group. Each group is required to upload a single report to Canvas via your respective seminar group. Please elect a group leader. The responsibility of the group leader is to upload the report on behalf of the group. Those submitting individually will be given a 10 marks deduction.

It is important for each group member to contribute substantially to the final submitted work. All group members are equally responsible for the entire submitted assignment If you feel that your group members did not contribute sufficiently to the work submitted, please highlight this to your instructor as soon as possible. Your instructor will then investigate and decide on any action that needs to be taken. It is not necessary for all group members to be awarded the same mark.

You must attempt all the questions in this assignment. There are 80 marks allocated to the questions and 20 marks allocated for your report writing.

Please keep a copy of your report before submission.

Background

Autistic Spectrum Disorder (ASD) is a neurodevelopment condition associated with significant healthcare costs, and early diagnosis can significantly reduce the financial impacts. Unfortunately, waiting time for an ASD diagnosis is lengthy and procedures are not cost-effective. The economic impact of autism and the increase in the number of ASD cases across the world reveals an urgent need for the development of easily implementable and effective screening methods. Therefore, a time-efficient and accessible ASD screening protocol is imminent to help health professionals and inform individuals whether they should pursue formal clinical diagnosis. Behavioral test scores and demographics characteristics have been shown to be effective in detecting ASD cases, and recently, it has been suggested that analytics can be used to speed up the screening time and to improve the sensitivity, specificity and accuracy of the diagnosis process. You and your team have just joined the analytics department of the CDC (Centers for Disease Control and Prevention) as summer interns. Your manager is keen to explore the use of analytics in the screening for adult autism. As a result, you are given a subset of the historical screening data and this is available as a dataset known as ***Autism.csv***. The description of the attributes in the dataset are given as follows:

FIELD	DESCRIPTION
A1_Score to A10_Score	Assessment scores on 10 carefully curated behavorial tests
Age	Age of subject taking the screening test
Gender	This variable gives the gender of the subject Value: m - 'male' or f - 'female'
Juandice	This variable indicates if the subject is born with juandice Value: yes or no
PDD	This variable states whether a family member of the subject has pervasive developmental disorder Value: yes or no
Class/ASD	This is the target variable. It indicates whether the subject has traits of autism Value: YES or NO

Task

In this assignment, you are tasked to use analytics to perform the required analyses on the given dataset. You will be drawing from what you have learnt in ANL307e to create the analytics models. The task will involve the use of the Modeler to develop the models that are capable of producing reasonably accurate outcomes. For all models, use the default parameter settings unless otherwise stated in the questions. Strictly follow the given guidelines when writing your report, paying particular attention to the stipulated page limit. Failure to comply with the given page limit will result in marks deduction.

Your report should answer all the following questions:

(a) Use the **Var. File** source node to read the dataset into the Modeler. Set the appropriate data measurements (such as nominal, continuous, etc) and define the data roles (such as Input, Record ID, Target, etc). In your report, discuss how you distinguish the

various data measurements and data roles. State whether this is a classification or regression problem.

(8 marks)

(b) Analyse this dataset using the **Data Audit** node and report appropriate dataset characteristics. Comment on any data quality issues. Explain how the Modeler distinguishes extreme values and outliers from the normal data instances.

(6 marks)

(c) Based on your understanding of the background and context of the problem, explain whether this is a descriptive, predictive or prescriptive analytics task. Be sure to provide the definitions to the three different types of analytics stated above and use appropriate examples to distinguish between them. Keep your response to no more than two pages.

(14 marks)

(d) Partition the dataset for training and testing using a ratio of 80% and 20% (Do not change the seed setting in the Partition node). Clearly state the respective number of screening records in the resultant training and testing sets. Explain whether this ratio is appropriate for the given dataset. Using an appropriate diagram, explain the training-testing framework. Following that, describe the concept of k-fold cross-validation in predictive modeling, and explain when such a paradigm should be applied. How will you interpret the results from a k-fold cross-validation evaluation process?

(16 marks)

(e) Using the *partitioned* data prepared in Part (d), apply and construct a Logistic Regression model with the input selection method set as "Forwards". Appraise the suitability of the application of this model for the given problem. Evaluate its predictive performance and report the overall training and testing accuracies, as well as the hit rates and the sensitivity measures of the different diagnosis outcomes for the testing scenario. State also the equation of the trained model.

(12 marks)

(f) Construct a CHAID decision tree model using the same partitions in Part (d). Compare and contrast the characteristics of the CHAID decision tree model with the Logistic Regression model you have used so far. Comment on its predictive performance by analysing the training and testing results.

(8 marks)

(g) Repeat the modeling task using the C&RT node in the Modeler. Compare and contrast its performance to the Logistic Regression and CHAID models that you have built in Part (e) and (f) and comment on the effectiveness of the C&RT decision tree model for this problem. In addition, explain the difference in the total value of 'n' in Node 0 of the CHAID and C&RT models.

(8 marks)

(h) Describe how you can apply predictive modeling to solve a practical problem related to your work, or in a business problem that you are familiar with. Be sure to give enough details of the organisation, the business problem, and the data used in your proposed solution so that your recommendation can be appropriately assessed. Describe the format of the data set to be used and discuss the possible results or outputs of the

selected predictive modeling technique. Novel answers will receive higher marks.
(8 marks)

Another 20 marks are allocated for your writing.

Your writing should be succinct but not at the expense of excluding relevant details. Highlight only the points that are relevant to your discussion. Use plain and simple language. Some questions may not come with absolutely right or wrong answers. For such questions you have the liberty to express your views about the problem. However, your points have to be supported by evidence and good reasoning. It's the quality and not the length that counts. Make sure you follow the report guidelines and style specified in this assignment.

Make sure you indicate your group number, student names and student numbers of the team leader and members.

The topics in the main report should be presented in the order according to the sequence of the tasks/questions listed in the assignment; that is, in the order of (a), (b), ..., etc. You can have several sub-sections within a section if you deem appropriate.

The report must be self-contained. It is important to include all relevant tables and figures in the report as evidence to support the answers given.

The follow are some details of report format:

- Length: should not exceed 12 pages (including the relevant graphs and tables, but excluding the cover page).
- Font Style: Times New Roman
- Font size: 12
- Line spacing: 1.5
- Margins: 1" for top, bottom, right and left
- Include the page number on each page.

Some further suggestions:

- Ensure minimal grammatical and typographical errors.
- Write clearly in plain English.
- Write appropriately to the context.
- Cite appropriate sources.
- Provide a reference or bibliography at the end of the main report.
- Include less relevant details in the Appendix (maximum 4 pages).
- Good overall presentation of the report.

---- END OF ASSIGNMENT ----