Case Study 1: Predictive Data Analytics
Identifying high school students performance
(with Python)

Due date: 9th September, 2018
Weighting: 25%

Introduction

This assignment is intended to allow you to display your knowledge and understanding of predictive data analytics. In this assignment, you will use classification algorithms implemented in Python to display your technical competence gained from the practical and lectures.

Instructions

1. The assignment report is due on **9th Sept** via **Blackboard Assignment** submission. It is a firm deadline (already includes weekend).
2. The assignment (data mining results) **will also be marked in the practical class**. Each group member will be asked specific questions about the case study in **week 8** practical labs. A 15% marks (out of 25 marks) will be assigned to you on the individual performance.
3. This is a group assignment. It is your responsibility to form a team of 3 members and you should do so preferably before the end of week 3. Groups are to be ARRANGED and MANAGED by you. As in real life, the performance of individuals in the team shall be judged by the performance of the team together, so choose your partners carefully.
4. Once the team is formed, you need to register the team on Blackboard. Choose "Tools" from the left side of the panel. Select the "Groups" tool and choose one of the CAB330 groups to register. This should be done by the end of week 3. To ensure that everyone agrees as to their responsibilities in the team and how you will work together, we have asked that you complete a Team Contract. This should be done before the team is registered. You can find the team agreement template and guidance under the Assessment Item 1 link.
5. Of course, the work you (group) hand in must be your own; no collaboration or borrowing from other groups is permitted. We will use the usual methods of detection of any plagiarism.
6. The datasets required for this assignment can be found on BlackBoard with the file named as **casestudy1-data.zip**.
7. The case study report should include response to the questions set in the case-study. There is no need of including an introduction, summary, conclusion or references in the report. Some answers may require screen shots.
8. Name the case-study report as **casestudy1.docx**. The word file should include a cover page with Student ID number and full name (as in QUT-Virtual) for all students, along with the group name. Combine this file with Jupyter notebook file, your **team contract**, and name the compressed file as **casestudy1.zip.** Submit this

file on **Blackboard (under the Assignment 1 link)**.

9. This assignment follows the standard QUT policy for late submission or plagiarized submission. Read the Assessment Policies on Blackboard or QUT Website.

Marks Distribution

In data analytics, there is hardly ever a single solution. The solution depends upon various setting such as input variables role and measurements, training size, underlying algorithm and the selected algorithm parameters. You may find that your project partner may have a different solution as yours. Your group should decide on a single project that you would like to be marked. Submit the report discussing the final project components.

We would mark the case study in the Week 8 practical class to explore your understanding of the data analytics concept. You should be prepared to show your final code and results to your marker. The marker will ask each student different questions and will assign individual mark (~15%).

Assignment Components	Marks
Data Pre-processing	4
Decision Tree Models	4
Regression Models	5.5
Neural Network Models	5.5
Comparison: Predictive Models	4
Report Presentation	1
Team Agreement	1

Case Study Scenario

In Michigan, the secondary education consists of 6 years of schooling, preceding 6 years of primary school education. Most of the students join the public and free education system. There are several courses (e.g. Sciences and Technologies, Visual Arts) that share core subjects such as the Spanish Language and Mathematics. A 20-point grading scale is used, where 0 is the lowest grade and 20 is the perfect score. Real values are permitted in the grade. During the school year, students are evaluated in three periods and the last evaluation (G3) corresponds to the final grade(PASS/FAIL). This study will consider data collected during the 2016- 2017 school year from two public schools, from Detroit Central High School(DCHS) and Troy High School(THS). Using surveys student's details about these attributes were collected such as mother's education, family income, social/emotional attributes (e.g. alcohol consumption) and school related (e.g. number of past class failures) variables that were expected to affect student performance. The questionnaire was reviewed by school professionals and tested on a small set of 15 students in order to get a feedback.

The department of education would like to identify which among these high school students will pass or fail. You have been hired as a data analyst consultant by this department. Your task is to inform decision makers the (characteristics of) secondary students using their past school grades (first and second periods), demographic, social and other school related data.

Case Study Dataset

The data set STUDENT contains 1044 observations and 35 variables. Variables are described in Table 1. You would note that some information is presented in ultiple ways. This is an example of the presence of redundant variables in a dataset.

The following information would assist you in assigning the variables roles.

- There are three target variables namely, G1, G2 and G3, with different types. Choose the target that suits best according to the given task.
- Identify if the variable is an input variable or a supplementary variable.
- Data transformation is required for a few input variables to get improved accuracy.

Table 1: List of Variables

Attribute	**Description**
Id	student's id
InitialName	student's initial
School	student's school name
Sex	student’s sex
Age	student’s age
Address	student's home address type
Famsize	family size(≤ 3 or > 3)
Pstatus	parent’s cohabitation status (living together or apart)
Medu	mother’s education(0 – none, 1 – primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
Fedu	father’s education(0 – none, 1 – primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
Mjob	mother’s job
Fjob	father’s job
Reason	reason to choose this school
guardian	student’s guardian
traveltime	home to school travel time (1 – < 15 min., 2 – 15 to 30 min., 3 – 30 min. to 1 hour or 4 – > 1 hour)
studytime	weekly study time (1 – < 2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours or 4 – > 10 hours)
Failures	number of past class failures(n if 1 ≤ n < 3, else 4)
schoolsup	extra educational school support (yes or no)

Famsup	family educational support (yes or no)
Paid	extra paid classes (yes or no)
activities	extra-curricular activities (yes or no)
Nursery	attended nursery school (yes or no
Higher	wants to take higher education (yes or no)
Internet	Internet access at home (yes or no)
romantic	with a romantic relationship (yes or no)
Famrel	quality of family relationships (1 – very bad to 5 – excellent)
freetime	free time after school (1 – very low to 5 – very high)
Gout	going out with friends (1 – very low to 5 – very high)
Dalc	workday alcohol consumption (1 – very low to 5 – very high)
Walc	weekend alcohol consumption (1 – very low to 5 – very high)
Health	current health status (1 – very bad to 5 – very good)
absences	number of school absences (0 to 75)
G1	first period grade (0 to 20)
G2	second period grade (0 to 20)
G3	Final result(PASS/FAIL)

Case Study Tasks

Your task is to build various predictive models such as decision tree, regression function, and neural network on this data set and compare them. Results inferred by these models should inform decision makers the (characteristics of) students performance.

Set up a new project for this task with **DMProj1** as the Python file and **STUDENT** as the dataset. Include various models in this source file. Name all the models meaningfully.

Task 1. Data Selection and Distribution. (4 marks)

1. What is the proportion of students who will pass?
2. Did you have to fix any data quality problems? Detail them.

   Apply imputation method(s) to the variable(s) that need it. List the variables that needed it. Justify your choice of imputation if needed.
3. What variables did you include in the analysis and what were their roles and measurement level set? Justify your choice.
4. What distribution scheme did you use? What "data partitioning allocation" did you set? Explain your selection. (Hint: Take the lead from Week 2 lecture on data distribution)

Task 2. Predictive Modeling Using Decision Trees (4 marks)

1. Build a decision tree using the default setting. Examine the tree results and answer the followings:
    a. What is classification accuracy on training and test datasets?
    b. List the decision rules.
    c. What are the 5 important variables in building the tree?
    d. Report if you see any evidence of model overfitting.

2. Build another decision tree tuned with GridSearchCV. Examine the tree results.
    a. What is classification accuracy on training and test datasets?
    b. What are the parameters used? Explain your decision.
    c. What are the optimal parameters for this decision tree?
    d. Which variable is used for the first split? What are the competing splits for this first split?
    e. What are the 5 important variables in building the tree?
    f. Report if you see any evidence of model overfitting.

3. What is the significant difference do you see between these two decision tree models? How do they compare performance-wise? Explain why those changes may have happened.
4. From the better model, can you identify which students to target for further consultation? Can you provide some descriptive summary of those students?

Task 3. Predictive Modeling Using Regression (5.5 marks)

1. In preparation for regression, apply transformation method(s) to the variable(s) that need it. List the variables that needed it.
2. Build a regression model using the default regression method with all inputs. Once you done it, build another one and tune it using GridSearchCV. Answer the followings:
    a. Report which variables are included in the regression model.
    b. Report the top-5 important variables (in the order) in the model.
    c. Report any sign of overfitting.
    d. What are the parameters used? Explain your decision. What are the optimal parameters? Which regression function is being used?
    e. What is classification accuracy on training and test datasets?
3. Build another regression model using the subset of inputs selected by RFE and selection by model methods. Answer the followings:
    a. Report which variables are included in the regression model.
    b. Report the top-5 important variables (in the order) in the model.
    c. Report any sign of overfitting.
    d. What is classification accuracy on training and test datasets?
4. Using the comparison statistics, which of the regression models appears to be better? Is there any difference between two models (i.e one with selected variables and

another with all variables)? Explain why those changes may have happened.

5. From the better model, can you identify which students to target? Can you provide some descriptive summary of those students?

Task 4. Predictive Modeling Using Neural Networks (5.5 marks)

1. Build a Neural Network model using the default setting. After that, tune it with GridSearchCV. Answer the following:
    a. What are the parameters used? Explain your decision. What is the optimal network architecture?
    b. How many iterations are needed to train this network?
    c. Do you see any sign of over-fitting?
    d. Did the training process converge and resulted in the best model?
    e. What is classification accuracy on training and test datasets?
2. Refine this network by tuning it with GridSearchCV. Report the trained model, same as Task 4.1
3. Build another Neural Network model with inputs selected from RFE with regression (use the best model generated in Task 3) and selection with decision tree (use the best model from Task 2). Answer the following:
    a. Did feature selection help here? Any change in the network architecture? What inputs are being used as the network input?
    b. What is classification accuracy on training and test datasets? Is there any improvement in the outcome?
    c. How many iterations are now needed to train this network?
    d. Do you see any sign of over-fitting?
    e. Did the training process converge and resulted in the best model?
    f. Finally, see whether the change in network architecture can further improve the performance, use GridSearchCV to tune the network. Report if there was any improvement.
3. Using the comparison methods, which of the models (i.e one with selected variables and another with all variables) appears to be better?
   From the better model, can you identify which students to target? Can you provide some descriptive summary of those students?

Task 5. Comparing Predictive Models (4 marks)

1. Use the comparison methods to compare the best decision tree model, the best regression model and the best neural network model.
    a. Discuss the findings led by (a) ROC Chart and Index; (b) Accuracy Score; (c) Classification Report.
    b. Do all the models agree on the students' characteristics? How do they vary?
2. Summarise your findings and present the results in a table.
3. Finally, based on all models and analysis, is there a particular model you will use in decision making? Justify your choice.
   How the outcome of this study can be used by decision makers?
4. Can you summarise positives and negatives of each predictive modelling method based on this analysis?

Assignment 1 Criteria Sheet:

Criteria	Comments and scoring
Non Submission of all components/ evidence of plagiarism	0
Has demonstrated a task with a working model with /without submission and demonstrates the ability to run the program and add some components. Questions were poorly answered.	1-5
Has demonstrated a task with a working model having a data source, and diagram with the substantial but incorrect implementation of at least one of the three components (predictive models). Questions were poorly answered.	6-9
Has implemented models for all three tasks (three data mining algorithms) with at least one being substantially correct. Shows some understanding of concepts with some success applying knowledge in basic questions	10-13
Has implemented models for all three tasks: Two of the three tasks are fundamentally correct, with substantially correct work flow diagrams which may contain minor errors. Response to questions shows a fundamental understanding of terms and concepts.	14-17
Has fundamentally correct implementation of all five tasks i.e. correct allocations of a target, rejections of variables according to instructions, running three models and comparing them. Includes a demonstration of the competent application of tools. Almost all questions have been reasonably answered. Demonstrate a strong understanding of the methods and terms including predictive mining, partitioning, imputation, comparison node, ensemble, misclassification, average squared error, sensitivity, specificity, lift, ROC chart, lift chart, support and confidence during written analyses. Some minor errors are allowed. Written application is required to be of reasonable standard.	18-20
Has implemented all of the requirements above with very few errors. A strong focus on the application on creative application of tools, and evaluation and interpretation of results is evident.	21-23
All of the criteria above are met; extensive model generation and analysis have been conducted to produce exceptional outcomes and have applied principles learnt in lectures to enhance the results.	24-25