Case Study 1: Predictive Data Analytics
Identifying Loyal Customers: Organic Products Offerings
(with Python)

Due date: 10th September, 2017
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 practicals and lectures.

Instructions

1. The assignment report is due on **10th 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 **organics (for python).csv**.
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. The source code should include plenty of comments for clarity.
8. Name the case-study report as **casestudy1.doc**. 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 your **Python source code**

and **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

A supermarket has just begun offering a line of organic products. The supermarket has a customer loyalty program. As an initial buyer incentive plan, the supermarket provided coupons for the organic products to all of their loyalty program participants and has now collected data from 22,000 participants that includes whether or not these customers have purchased any of the organic products. The supermarket's management would like to determine which customers are likely to purchase recently offered organic products.

The supermarket's management would like to determine which customers are likely to purchase these products. You have been hired as a data analyst consultant by this management. Your task is to inform decision makers the (characteristics of) potential buyers from the entire user base by building predictive models on this data set of selective customers.

Case Study Dataset

The data set ORGANICS contains over 22,000 observations and 18 variables. Variables are described in Table 1. You would note that some information is presented in multiple 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.

- Variables DOB, AGE, AGEGRP1, and AGEGRP2 are all different measurements for the same information. (Hint: It is not good to bias the data mining model by giving any pre-defined grouping of age. The best practice would be to let the model choose its own group grouping.)
- Variable NGROUP contains collapsed levels of the variable NEIGHBORHOOD.
- Variables LCDATE and LTIME represent the same information in two different formats.
- There are two target variables namely, ORGANICS and ORGYN, with different types. Choose the target that suits best according to the given task.

Name	**Description**
CUSTID	Customer Loyalty Identification Number
GENDER	M = male, F = female, U = unknown
DOB	Date of birth
EDATE	Date extracted from the daily sales database
AGE	Age, in years
AGEGRP1	Age group 1
AGEGRP2	Age group 2
TV_REG	Television Region
NGROUP	Neighborhood group
NEIGHBORHOOD	Type of residential neighborhood
LCDATE	Loyalty card application due
LTIME	Time as loyalty card member
ORGANICS	Number of organic products purchased
BILL	Total amount spent
REGION	Geographic region
CLASS	Customer loyalty status: tin, silver, gold or platinum
ORGYN	Organics purchased? 1 = Yes, 0 = No
AFFL	Affluence grade on a scale from 1 to 30

Table 1: List of Variables

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) potential buyers

Set up a new project for this task with **DMProj1.py** as the Python file and **ORGANICS** as the dataset. Include various models in this source file. Name all the models meaningfully and include plenty of comments for clarity.

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

1. What is the proportion of individuals who purchased organic products?
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. Which variable is used for the first split? What are the competing splits for this first split?
   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 customers to target for further marketing? Can you provide some descriptive summary of those customers?

Task 3. Predictive Modeling Using Regression (5.5 marks)

1. In preparation for regression, apply transformation/scaling 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 customers to target? Can you provide some descriptive summary of those customers?

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

1. Build a Neural Network model using the default setting. Answer the following:
   a. What is the network architecture of the model?
   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. What are the parameters used? Explain your decision. 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 customers to target? Can you provide some descriptive summary of those customers?

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 customers' characteristics? How do they vary?
2. 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?
3. 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 source code 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 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, misclassification/accuracy, average squared error, precision, recall, F1, ROC 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