Case Study 1: Predictive Data Analytics
At Risk Prediction: "Is the householder a high or low risk for a loan?"

Due date: 11:59 PM, 8th September 2019
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 **8th Sept** via **Blackboard Assignment** submission. It is a firm deadline (already includes weekend).
2. The case study models and their outcomes **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 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 dataset required for this assignment can be found on BlackBoard with the file named as **HouseholderAtRisk.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 screenshots.
8. Name the case-study report as **casestudy1.docx**. The word file should include a cover page with the Student ID number and full name (as in

QUT-Virtual) for all students, along with the group name. Combine this file with 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 a single report per group after discussing the final project components.

In addition to your submitted report, 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 Dataset

The data set ***HouseholderAtRisk*** is based on the US Current Population Survey March Supplement. It contains 40,000 observations and 18 variables. Each observation indicates a householder and the 18 variables describe their characteristics. Detail of each variable is described in Table 1. You have been hired as a data analyst consultant by a bank. Your task is to inform decision makers the (characteristics of) householders using their attributes to predict which householders are at risk for providing a loan.

Table 1: List of Variables

Attribute	Description
ID	Householder's Identification Number
Age	Householder's age
WorkClass	Householder's class of work – Private, Federal-gov, State-gov, etc.
Weighting	This estimate represents the number of people that have the same socio-economic characteristics as the given householder. For more information https://www.census.gov/programs-surveys/cps/technical-documentation/methodology/weighting.html
Education	Householder's highest level of education
NumYearsEducation	Householder's years of education
Marital-Status	Householder's marital status - Never-married, Separated, Widowed, Divorced, etc.
Occupation	Householder's type of occupation - Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, etc.
Relationship	Householder's relationship in the family - Wife, Own-child, Husband, Not-in-family, etc.
Race	Householder's race - White, Asian-Pac-Islander, Amer-Indian-Eskimo, etc.
Gender	Householder's gender
CapitalLoss	Householder's investment loss
CapitalGain	Householder's investment gain
CapitalAvg	Average of Householder's investment gain and loss.
NumWorkingHoursPerWeek	Number of Hours of work per week
Sex	Householder's sex(0-Male or 1-Female)
CountryOfOrigin	Householder's country of origin - United-States(US, USA or United-States), Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, etc
AtRisk	Householder's risk for a loan High or Low

Case Study Tasks

Your task is to build various predictive models including 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) householder having high or low risk. This information can be utilized in multiple ways to assist various stakeholders.

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

Specific tasks for each data analytics process are listed below.

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

1. What is the proportion of householders who have high risk?
2. Did you have to fix any data quality problems? Detail them.

   Apply the imputation method(s) to the variable(s) that need it. List the variables that needed it. Justify your choice of imputation if needed.
3. The dataset may include irrelevant and redundant variables. 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 variables that are used for the second 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 variables that are used for the second 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 – default (Task 2.1) and using GridSearchCV (Task 2.2)? How do they compare performance-wise? Explain why those changes may have happened.
4. From the better model, can you identify which householders to target for providing loan? Can you provide some descriptive summary of those householders?

Task 3. Predictive Modeling Using Regression (5.5 marks)

1. Describe why you will have to do additional preparation for variables to be used in regression modelling. 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 have completed it, build another model 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 either by RFE or the selection by model method. 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 the 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 householders to target for providing loan? Can you provide some descriptive summary of those householders?

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 are the parameters used? Explain your decision. What is the 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. Answer the following:
    a. What are the parameters used? Explain your decision. What is the 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?
3. Would feature selection help here? Build another Neural Network model with inputs selected from RFE with regression (use the best model generated in Task 3) and from the decision tree (use the best model from Task 2). Answer the following for the best neural network model:

a. Did feature selection help here? Which method of feature selection produced the best result? 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.

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:
      (i) ROC Chart and Index;
      (ii) Accuracy Score;
   b. Which model would you use in deployment based on these findings? Discuss why?
   c. Do all the models agree on the householder's characteristics? How do they vary?
2. How the outcome of this study can be used by decision makers?
3. Can you summarise the positives and negative aspects of each predictive modelling method based on this data analysis exercise?

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 results with the substantial but incorrect implementation of at least one of the five 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 five tasks: Three of the five tasks are fundamentally correct, with substantially correct process 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. The written application is required to be of a 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. Good presentation of the report and team agreement is included.	24-25