NANYANG TECHNOLOGICAL UNIVERSITY

SEMESTER 1 EXAMINATION 2015-2016

BC2406 Analytics I: Visual and Predictive Techniques

Nov/Dec 2015 Time Allowed: 2½ hours

INSTRUCTIONS

1. This paper contains **FOUR(4)** questions and comprises **EIGHT(8)** pages
2. Answer all **FOUR(4)** questions.
3. The number of marks allocated is shown at the end of each question.
4. Begin your answer to each question on a separate page of the answer book.
5. Answers will be graded for content and appropriate presentation.

Question 1

For each of the following activities listed below, indicate if it is a data mining task. If the activity is a data mining task please state the kind of data mining (e.g. classification, prediction or clustering?)

(a) Dividing the customers of a bank according to their loyalty scores.

No

(2 marks)

(b) Identify the sales person with the highest sales.

No

(2 marks)

(c) Estimating the future box office of movies using online reviews of these movies.

yes. prediction

(2 marks)

(d) Guessing the outcomes of tossing a (fair) pair of dice.

No

(2 marks)

(e) Forecasting if a visitor to the website is a hacker based on the visitor's past online activities.

yes. classification

(2 marks)

(TOTAL: 10 marks)

Question 2

ABC.com is an apparel web retailer launched in 2000. ABC.com defined "heavy spenders" as visitors who spend more than $100 on an average order at its site. The retailer wants to build a model to predict the types of visitors who potentially can be heavy spenders. From the model, the retailer hopes to learn the profiles and the online activities of these heavy spenders.

(a) What is the dependent variable for this predictive model?

Type of Spender ---- 1 = Heavy Spender / 2= Non-Heavy Spender

(1 marks)

Demo info (Profiles)

Online act.

(b) Can you suggest 5 predictor variables to be used in this predictive model?

Gender, Age, Income, Number of visit per month, Ave time on website

(5 marks)

(c) What are some possible insights the retailer can get from predicting the heavy spenders? Discuss two possible insights based on the predictor variables you have chosen.

Which Gender is higher likely to be "heavy spenders"
Which Age Group is higher likely to be "heavy spenders"

(4 marks)

(TOTAL: 10 marks)

We will see the effect of the factors/ attributes that will relate to heavy and non heavy spenders once we identify, we can target the certain cluster.
How this factors will affect the dependent variable. To suggest a certain group of product to certain group of customer.

Question 3

An insurance company wants to predict the amount of fraudulent automobile accidents claims. You will help the firm perform the prediction using regressional analysis and data cleaning is required before the analysis. Table 3.1 provides a sample of the data and the variables in the dataset are explained in Table 3.2. Summary statistics are shown in Table 3.3 and pairwise correlations among some variables are listed in Table 3.4. Before starting the formal analysis, you tried a preliminary regression model with three randomly selected variables, and the preliminary regression results are shown in Table 3.5.

Table 3.1 Sample of Fraudulent Claims Data

Claim Id	Action	City code	Gender	Num child	Income	House Value	Age	Credit Score	Fraud Or not	Fraud amount
1	phonecall	small	male	1	5	9	31	2	1	1073
2	mail	large	female	1	1	7	35	3	1	2146
3	phonecall	small	female	2	5	8	38	8	1	1073
4	phonecall	medium	male	1	4	8	40	4	1	1716.8
5	phonecall	small	female	1	4	8	57	5	1	2146
6	mail	large	female	1	1	7	28	10	1	4292
6	mail	small	female	1	4	5	61	6	1	1073
7	phonecall	medium	female	1	4	8	39	9	1	1502.2
9	phonecall	large	female	1	1	5	50	1	1	643.8
10	mail	large	female	1	5	8	58	4	1	2146

Table 3.2 Data Table

Claim id	id number of each claim
Action	action taken by the insurance firm to contact the claimant after receiving the claim (e.g. by phone or by mail)
CityCode	=large if the claimant lived in a large city, =medium if the claimant lived in a medium city, =small if lived in a small city
Gender	male or female;
Numchild	number of children
Income	household income (*1000)
HouseValue	market value of claimant's house (*Million $)
CreditScore	Credit score provided by a third party
Age	age
Fraudornot	indicate if claim is fraudulent. 1: fraud, 0: no fraud
Fraud amount	Fraud amount

Note: Question 3 continues on page 4

Question 3 (continued)

Table 3.3 Summary Statistics

	Numchild	Income	HouseValue	CreditScore	Age	Fraud amount
Mean	1.05	4.00	6.44	5.55	55.95	2708.95
Median	1.00	4.00	8.00	5.00	61.00	2146.00
Standard Deviation	0.25	1.62	2.44	3.35	18.60	1772.16
Range	2.00	6.00	8.00	11.00	60.00	10515.40
Minimum	1.00	1.00	1.00	1.00	20.00	214.60
Maximum	3.00	7.00	9.00	12.00	80.00	10730.00
Count	15	216	218	218	218	218

Table 3.4 Correlation Matrix

	Numchild	Income	HouseValue	CreditScore	Age
Numchild	1				
Income	0.138	1			
HouseValue	0.103	0.939	1		
CreditScore	0.064	0.197	0.147	1	
Age	0.108	0.098	0.056	0.049	1

Table 3.5 Preliminary Regression Results

Dependent variable: Fraud amount

R Square	0.814454965
Adjusted R Square	0.813237509

	Coefficients	Standard Error	*p*-value
Intercept	16511.31	350.57	0
Age	-168.24	2.28	0
Income	-127.47	12.43	0.0000007
Numchild	68.92	88.75	0.438

Note: Question 3 continues on page 5

No outlier
Missing Value:
Num Child - exclude/ delete variable
Income - Replace the 2 missing value with mean

outliers : Meant +- 3SD
missing:
Normalisation: Mean <SD

Question 3 (continued)

preprossing/ manage data

(a) Based on the summary statistics in Table 3.3, what potential issue(s) does this data set have? State the reason(s) for your conclusion and suggest solution(s) to mitigate the issue(s)?

Size of number of child is smaller than the rest - change those with no kids to 0

(5 marks)

Fraud Amt have outlier/ huge variation in data. - normalisation

(b) Read all the tables and attributes. Some attributes cannot be used in this regression analysis, and you should omit these variables in the final regression model. Please list all the variables you are will exclude from the model, and provide reasons for each deletion.

(10 marks)

Clam ID - there are no information or meaningful data to be retrieved from it
Action - Cos is action after Fraud, Num of child - too many missing value, Fraud or not- will always be 1, Delete income/House Vaue -- High correlation

(c) Among all the variables you are considering to use for the regression, some of them cannot be analyzed by the software directly. Please list these variables. How will you recode these variables? Please explain with details.

(5 marks)

City Code, Gender - are now in string Convert to numerical value

(d) Refer to the preliminary results in Table 3.5. How does this model perform? Can you use this model to make the prediction?

(5 marks)

R^2 measure modelfit

Adjust R-square near 1.0, therefore can use to make prediction

Lower MSE means the model perform very well with prediction. Wit
R^2 cannot not good enough to make interprediction

(e) Based on the preliminary estimation in Table 3.5, how will you interpret the relationships between the predictors and "Fraud amount"?

(5 marks)

(TOTAL: 30 marks)

Interpret relation of age and the dependent variable- (magnitude and significant)
One unit increase in age will decrease the Fraud amount by $168.24. This relationship is at 0.1% significant level.

Income-
When income increase by $1,000 will decrease the fraud amount by $127.47. This relationship is at 0.1% signifcant level

Question 4

Last month, your supervisor, Sue, mailed 20,000 of your existing customers with a special promotional offer. The overall take-up rate for this offer is 5%. Out of the customers who responded, the average purchase quantum is $12. The postage and printing costs for each mail is $0.55. Sue now delegated this task to you and you do not wish target the customers randomly as before. To help with the customer selection, you built two decision tree models. First, you sample 2,000 customers from the 20,000 customers with 1,000 responding customers and 1,000 non-responding customers. To build the model, you have 1,000 out of the 2,000 customers randomly assigned as training data, and the remaining customers assigned as the testing data. "Class 1" corresponds to customers who responded to the offer, and "Class 0" corresponds to those who did not. Age, sex, income, education, location, and job industry were used as input variables for these customers. You build two trees (Decision Tree 1 and Decision Tree 2) from the training data and test them with the testing dataset. The performances of these two trees are listed in Table 4.1. Figure 4.1 shows the final tree of Decision Tree 1. In each node, the number of observations with Class 0 is given in the left part, and the number of observations with Class 1 in the right part. The leaf nodes have been drawn as rectangles. You also derived the lift charts for these two decision trees. Some values on Decision Tree 1's lift chart have been summarized in Table 4.2.

Table 4.1 Performances of Decision Tree 1 and Decision Tree 2

	Decision Tree 1	Decision Tree 2
Accuracy	71.6%	77%
Sensitivity	0.804	0.796
ROC	0.762	0.848

ROC = sensitivity + sparsity

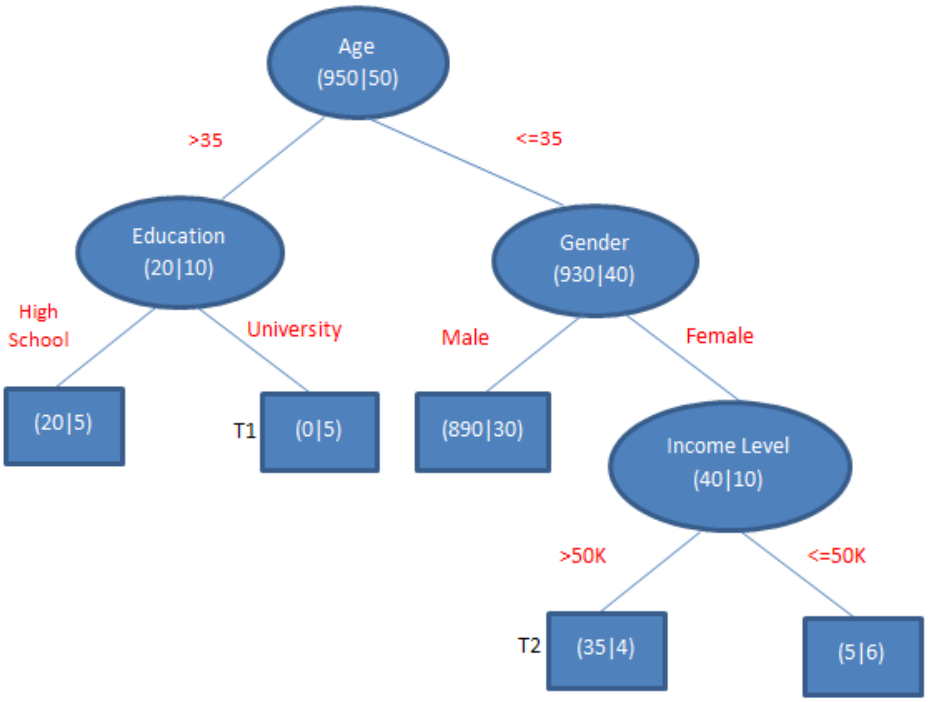
Table 4.2 Values from Lift Chart for Decision Tree 1

X Axis-Number of Customers	Y Axis-Cumulative Number of Respondents
0	0
50	50
100	92
200	147
250	188
300	209

Note: Question 4 continues on page 7

Question 4 (continued)

Figure 4.1 Decision Tree 1



Please answer the following questions.

(a) Refer to Table 4.1. Compare the performances between these two decision trees. Which model performs better?

Decision Tree 2 is better :ROC = the 2 S and therefore ROC and accuracy both is better (5 marks)

(b) Refer to Figure 4.1. The company would like to have a few, simple, comprehensible language rules that embody the decisions represented by the tree. Please derive all the rules from the right branch. (3 marks)

If age <= 35 & Male, Then Class 0 (No Response)
If Age < = 35 & Female & Income > 50k, Then class 0 (No Response)
If Age <= 35 & Female & Income <= 50k. Then Class 1 (Response)

(c) Refer to Figure 4.1. Explain why the attributes "location" and "job industry" are not used in Decision Tree 1?

This 2 valuable do not improve the purity of GINI index (5 marks)

(d) Refer to Figure 4.1. Without relying on additional calculations between node T1 and T2, which node will have higher Gini index. Explain how you arrive at this conclusion without additional calculations? (5 marks)

T1, all the instance are class 1. Very Pure
T2. Very inpure

(e) Refer to Figure 4.1. The Full tree has 4 levels and 6 leaves, but the final tree has 3 levels and 5 leaves. What technique do you think is used to reduce the full tree to the current tree? What is the main reason for using this technique? (5 marks)

Pruning. To prevent Overfitting.

Note: Question 4 continues on page 8

Question 4 (continued)

(f) Which sampling technique is used here? Why do we need to use this technique in this problem?

(5 marks)

(g) Describe the cost/benefit matrix. Please provide details of calculations.

(10 marks)

(h) Refer to Table 4.2. If you have a budget which allows to target and mail 100 customers, using Decision Tree 1 how much net revenue can you get? How much net revenue can you get if you use Sue's method of random mailing? What are the benefits for using the decision tree model? Please show the calculations to support your answer.

(12 marks)

(TOTAL: 50 marks)

- END OF PAPER -