M.Tech. IN ADVANCED INFORMATION TECHNOLOGY – SOFTWARE TECHNOLOGY (MTECHST)

Term-End Examination

00020

December, 2014

MINE-022 : BUSINESS INTELLIGENCE

Time : 3 hours   *Maximum Marks : 100*

Note :

(i)   Section I is **compulsory** and carries **30** marks

(ii)   Section II carries **70** marks. Answer any **five** questions.

(iii) Assume suitable data wherever required.

(iv)   Draw suitable sketches wherever required.

(v)   Italicized figures to the right indicate maximum marks.

SECTION I

1. As part of a promotional campaign, the marketing manager of "Dilshan Retails" wishes to identify the potential buyers to send promotional catalog offers. The manager uses predictive modelling tools to find the potential customers for promotional offers. He takes help from one of the statisticians working in the same company. The objective of the selection is to "Maximize profit with cost". The variables for the selection are as follows :

Variables	Variables
Purchase(c,2)	Total Returns
Rupees Spent	Mens Apparel
Yearly Income	Home Furniture
Home Value	Lamps Purchase
Order Frequency	Linens Purchase
Recency	Blankets Purchase
Married(c,2)	Towels Purchase
Name Prefix(c,4)	Outdoor Product
Age	Coats Purchase
Sex(c,2)	Ladies Coats
Telemarket Ind.(c,2)	Ladies Apparel
Rents Apartment	His/Her Apparel
Occupied <1 Year	Jewelry Purchase
Domestic Product	Date 1st Order
Apparel Purchase	Telemarket Order(c,5)
Leisure Product	Account Number(c,128)
Luxury Items(c,2)	State Code(c,55)
Kitchen Product	Race(c,5)
Dishes Purchase	Heating Type(c,4)
Flatware Purchase	Number of Cars(c,4)
Total Dining (kitch+dish+flat)	Number of Kids
Promo : 1 – 7 Months	Travel Time
Promo : 8 – 13 Months	Education Level(c,4)
₹ Value per Mailing	Job Category
Country Code	
	Note: 'C' indicates class levels

Answer the following :

(a) What variables (input and target) are important for identifying the potential customers ? Which method do you prefer for variable selection ? *3+3*

(b) Identify the suitable predictive analytics technique (model). *2*

(c) How do you apply the model using SAS/SEMMA methodology ? Explain in detail. *8*

2. (a) The following table shows 24 records with their actual class and the probability of them being Class 1 members, as estimated by a classifier.

Actual class	Probability of Class 1	Actual class	Probability of Class 1
1	0·9959	1	0·5055
1	0·99875	0	0·4713
1	0·9844	0	0·3371
1	0·9804	1	0·2179
1	0·9481	0	0·1992
1	0·8892	0	0·1494
1	0·8476	0	0·0479
0	0·7628	0	0·0383
1	0·7069	0	0·0248
1	0·6807	0	0·0218
1	0·6563	0	0·0161
0	0·6224	0	0·0035

Based on the above data, find the misclassification rate when the cutoff is 0·5, 0·25 and 0·75. *10*

(b) Describe the roles assumed by the validation, partition and test partition in supervised learning algorithms. *4*

SECTION II

3. A large number of insurance records are to be examined to develop a model for predicting fraudulent claims. Of the claims in the historical database, 1% were judged to be fraud. A sample is taken to develop a model, and oversampling is used to provide a balanced sample in light of the very low response rate. When applied to this sample (N = 800) the model ends up correctly classifying 310 frauds, and 270 no frauds. It missed 90 frauds, and classified 130 records incorrectly as frauds when they were not.

   (a) Produce the confusion matrix for the sample as it stands. *5*

   (b) Find the adjusted misclassification rate. *5*

   (c) What percentage of new records would you expect to be classified as fraudulent ? *4*

4. (a) For the following classification matrix, find the measures : *2+2+2+2*

   (i) Overall error

   (ii) Overall accuracy

   (iii) Sensitivity

   (iv) Specificity

Predicted	1	0
1 (actual)	201	85
0 (actual)	25	2689

(b) For the following cases, identify whether the task required is supervised or unsupervised learning. Write two reasons for each case. 2+2+2

(i) Deciding whether to issue loan to an applicant based on demographics and financial data (with reference to data base of similar prior customers).

(ii) Identifying segments of similar customers.

(iii) Printing of customer discount coupons at the conclusion of a grocery store checkout based on what you just bought and what others have bought previously.

5. Write different distance measures for clustering. Apply k-means clustering algorithm for the following dataset and show clusters with proper distance measures. 5+9

Student	Age	Subject 1	Subject 2	Subject 3
S1	18	73	75	57
S2	18	79	85	75
S3	23	70	70	52
S4	20	55	55	55
S5	22	85	86	87
S6	19	91	90	85
S7	20	70	65	60
S8	21	53	56	59
S9	19	82	82	60
S10	47	75	76	77

6. (a) What is Artificial Neural Network (ANN) ? Explain the general learning process of ANN. Also, judge the following : *3+4+3*

"A neural network has been constructed to predict the creditworthiness of applicants. There are two output nodes, one for yes (1, yes : 0, no) and one for no (1, no : 0, yes). An applicant received a score of 0·83 for the "yes" output node and a 0·44 for the "no" output node. Discuss what may have happened and whether the applicant is a good credit risk.

(b) Explain the factors for choosing the Neural Networks architecture. *4*

7. (a) Apply association rules for the following table, containing telephone faceplate purchases. The numeric 1 (one) indicates presence of an item in the transaction and numeric 0 (zero) indicates absence of transaction.

Transaction	Red	White	Blue	Orange	Green	Yellow
1	1	1	0	0	1	0
2	0	1	0	1	0	0
3	0	1	1	0	0	0
4	1	1	0	1	0	0
5	1	0	1	0	0	0
6	0	1	1	0	0	0
7	1	0	1	0	0	0
8	1	1	1	0	1	0
9	1	1	1	0	0	0
10	0	0	0	0	0	1

From the above table, find the association rules with minimum support and minimum confidence of above 50% for red, white and green transactions. *10*

(b) Compare and contrast between the measures Confidence and Lift. *4*

8. (a) Why are variable measurement level and model role important for model creation ? *7*

(b) Business intelligence applications enable managers to take decisions in time. Explain it. *7*