

|  |  |  |
| --- | --- | --- |
| **Semester: Jan – Mar 25**  **Maximum Marks: 50 Examination: ETE Exam Date: 05.04.2025 Duration: 3 Hours** | | |
| **Programme code: 01**  **Programme: MBA** | **Class:** SY | **Semester/Trimester: VI** |
| **College:**  **K. J. Somaiya Institute of Management** | **Name of the department/Section/Center: Business Analytics** | |
| **Course Code: 217P01M627** | **Name of the Course: Machine Learning approach for Multivariate Data Analysis** | |
| **Instructions:**   1. **You must attempt 5 questions in all. All questions carry equal marks.** 2. **Question 1 is compulsory.** 3. **All subparts to a question must be answered.** 4. **Calculator is allowed.** | | |

|  |  |  |
| --- | --- | --- |
| **Question No.** |  | **Max.**  **Marks** |
| **1** | A researcher examined customers’ impulsive buying behaviour. The aim is to determine the influencing factors that lead a person for impulsive buying. The researcher used logistic regression analysis to find the influencing factors. The variables used for the analysis are mentioned below.  Dependent variable:  Impulsive buying behaviour (Nominal scale)  0-Not an impulsive buyer  1-An impulsive buyer  Independent variables:  Gender (Nominal scale)  Age (Ratio scale)  Marital status (Nominal scale)  Income (Ratio scale)  Frequency of buying (Ratio scale)  Preferred mode of payment (Nominal scale)  The output is given below:   |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | | **Variables in the Equation** | | | | | | | | |  | | B | S.E. | Wald | df | Sig. | Exp(B) | |  | **Gender** |  |  |  |  |  |  | | Male® |  |  |  |  |  |  | | Female | .515 | .262 | 3.884 | 1 | .049 | 1.674 | | **Age** | -.127 | .047 | 7.313 | 1 | .007 | .881 | | **Marital Status** |  |  |  |  |  |  | | Single® |  |  |  |  |  |  | | Married | .254 | .120 | 4.454 | 1 | .035 | 1.289 | | **Income** |  |  |  |  |  |  | | 50000-100000® |  |  |  |  |  |  | | 100000-150000 | .842 | .985 | .730 | 1 | .393 | 2.321 | | 150000 & above | 1.091 | .174 | 39.315 | 1 | .000 | 2.978 | | **Frequency of buying** | -.025 | .020 | 1.636 | 1 | .201 | .975 | | **Preferred mode of payment** |  |  |  |  |  |  | | UPI® |  |  |  |  |  |  | | Credit card | 1.788 | .901 | 3.942 | 1 | .047 | 5.980 | | Constant | -6.570 | .781 | 70.683 | 1 | .000 | .001 | | 1. Variable(s) entered on step 1: Gender, Age, Marital status, Income, Frequency of buying, Preferred mode of payment. | | | | | | | | | Note: ® refers to reference category | | | | | | | |  1. Interpret the odds ratios. 2. Construct the null and alternative hypothesis and conclude the result at 5% level of significance. | 10 |
| **2** | A new method to determine the amount of low-calorie sweetener in different food samples has been introduced by a company. The company wants to apply this method on four food samples. The company has four labs. So the tests that involve the application of this new method to each of the food samples will be carried out in each of the four labs. Each of the labs have reported the mean recovery percentages of the amount of low-calorie sweetener they could detect on each of the food samples. The data are given below.   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | Labs | Food samples | | | | | |  | Food sample 1 | Food sample 2 | Food sample 3 | Food sample 4 | | Lab 1 | 99.5 | 83 | 96.5 | 96.8 | | Lab 2 | 105 | 105.5 | 104 | 108 | | Lab 3 | 95.4 | 81.9 | 87.4 | 86.3 | | Lab 4 | 93.7 | 80.8 | 84.5 | 70.3 |   They run Two-Way ANOVA and get the output. The output is mentioned below.   |  |  |  |  |  | | --- | --- | --- | --- | --- | | Anova: Two-Factor Without Replication | | |  |  | |  |  |  |  |  | | *SUMMARY* | *Count* | *Sum* | *Average* | *Variance* | | Row 1 | 4 | 375.8 | 93.95 | 55.11 | | Row 2 | 4 | 422.5 | 105.625 | 2.895833333 | | Row 3 | 4 | 351 | 87.75 | 31.65666667 | | Row 4 | 4 | 329.3 | 82.325 | 93.6825 | |  |  |  |  |  | | Column 1 | 4 | 393.6 | 98.4 | 25.28666667 | | Column 2 | 4 | 351.2 | 87.8 | 140.0466667 | | Column 3 | 4 | 372.4 | 93.1 | 78.94 | | Column 4 | 4 | 361.4 | 90.35 | 257.1766667 |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | ANOVA |  |  |  |  |  |  | | *Source of Variation* | *SS* | *df* | *MS* | *F* | *P-value* | *F crit* | | Rows | 1201.723 | 3 | 400.574 | 11.913 | 0.002 | 3.863 | | Columns | 247.408 | 3 | 82.469 | 2.453 | 0.130 | 3.863 | | Error | 302.628 | 9 | 33.625 |  |  |  | | Total | 1751.758 | 15 |  |  |  |  |   **Answer the following questions:**   1. State the Null Hypothesis (Ho) and Alternate Hypothesis (H1). 2. How do you get degrees of Freedom for Rows? 3. How do you get degrees of Freedom for columns? 4. How do you get degrees of Freedom for Total? 5. Conclude the result at 5% level of significance. | 10 |
| **3** | A researcher conducted a survey on middle-aged men in a fitness club to determine the relationship between physiological variables (weight in pounds, waist in inches and pulse rate) and exercise variables (sit\_ups and jumps).   |  |  | | --- | --- | | Physiological variables:  Weight in pounds  Waist in inches  Pulse rate | Exercise variables:  Sit-ups  Jumps |   The researcher then conducted canonical correlation analysis. The output is mentioned below:   |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | | **Canonical Correlations** | | | | | | | | |  | Correlation | Eigenvalue | Wilks Statistic | F | Num D.F | Denom D.F. | Sig. | | 1 | .771 | 1.461 | .391 | 2.996 | 6.000 | 30.000 | .021 | | 2 | .194 | .039 | .962 | . | . | . | . | | H0 for Wilks test is that the correlations in the current and following rows are zero | | | | | | | |  |  |  |  | | --- | --- | --- | | **Set 1 Canonical Loadings** | | | | Variable | 1 | 2 | | Sit\_up | -.847 | .532 | | Jumps | -.172 | .985 |  |  |  |  | | --- | --- | --- | | **Set 2 Canonical Loadings** | | | | Variable | 1 | 2 | | Weight | .638 | -.744 | | Waist | .933 | -.357 | | Pulse | -.355 | -.063 |  |  |  |  |  |  | | --- | --- | --- | --- | --- | | **Proportion of Variance Explained** | | | | | | Canonical Variable | Set 1 by Self | Set 1 by Set 2 | Set 2 by Self | Set 2 by Set 1 | | 1 | .373 | .222 | .468 | .278 | | 2 | .627 | .024 | .228 | .009 |  1. Identify the set of dependent and independent variables. 2. Interpret the output. | 10 |
| **4** | ACE Ltd. is an apparel brand that operates both offline (physical stores) and online (e-commerce platform) channels to sell its products. The company offers a wide range of clothing items, including but not limited to, shirts, pants, dresses, and accessories. ACE Ltd. prides itself on providing fashionable and high-quality apparel to its customers. The company is interested to identify the factors which are discriminating between online and offline customers. The company has considered five variables that can be recognized as probable discriminators; these include:   * Offers/discounts motivate me to buy (Offers) * Wide product variety is important to me (Product\_variety) * Exchange/return policies are essential for purchase (Exchange\_Return) * Good brand name motivates me to buy (Brand\_name) * Influenced by reviews and recommendations (Influencers)   Dependent variable:  Types of customers   1. Online 2. Offline   The researcher then ran a discriminant analysis to find the predictors which are differentiating between two groups. The output is given below:   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | **Group Statistics** | | | | | | | Type of customers | | Mean | Std. Deviation | Valid N (listwise) | | | Unweighted | Weighted | | Online customers | Offers | 4.25 | 0.637 | 167 | 167.000 | | Product\_variety | 3.52 | 1.212 | 167 | 167.000 | | Exchange\_Return | 2.88 | 1.074 | 167 | 167.000 | | Brand\_name | 2.51 | 1.069 | 167 | 167.000 | | Influencers | 2.96 | 1.064 | 167 | 167.000 | | Offline customers | Offers | 3.87 | 0.757 | 118 | 118.000 | | Product\_variety | 3.74 | 0.852 | 118 | 118.000 | | Exchange\_Return | 4.00 | 0.654 | 118 | 118.000 | | Brand\_name | 2.69 | 0.676 | 118 | 118.000 | | Influencers | 2.79 | 0.846 | 118 | 118.000 | | Total | Offers | 4.09 | 0.713 | 285 | 285.000 | | Product\_variety | 3.61 | 1.081 | 285 | 285.000 | | Exchange\_Return | 3.34 | 1.075 | 285 | 285.000 | | Brand\_name | 2.59 | 0.929 | 285 | 285.000 | | Influencers | 2.89 | 0.982 | 285 | 285.000 |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | **Tests of Equality of Group Means** | | | | | | |  | Wilks' Lambda | F | df1 | df2 | Sig. | | Offers | 0.931 | 20.851 | 1 | 283 | 0.000 | | Product\_variety | 0.990 | 2.787 | 1 | 283 | 0.096 | | Exchange\_Return | 0.736 | 101.551 | 1 | 283 | 0.000 | | Brand\_name | 0.992 | 2.367 | 1 | 283 | 0.125 | | Influencers | 0.992 | 2.231 | 1 | 283 | 0.136 |  |  |  |  |  |  | | --- | --- | --- | --- | --- | | **Eigenvalues** | | | | | | Function | Eigenvalue | % of Variance | Cumulative % | Canonical Correlation | | 1 | .640a | 100.0 | 100.0 | 0.625 | | a. First 1 canonical discriminant functions were used in the analysis. | | | | |  |  |  |  |  |  | | --- | --- | --- | --- | --- | | **Wilks' Lambda** | | | | | | Test of Function(s) | Wilks' Lambda | Chi-square | df | Sig. | | 1 | 0.610 | 138.834 | 5 | 0.000 |  |  |  | | --- | --- | | **Standardized Canonical Discriminant Function Coefficients** | | |  | Function | | 1 | | Offers | -0.349 | | Product\_variety | -0.611 | | Exchange\_Return | 1.348 | | Brand\_name | -0.385 | | Influencers | 0.072 |     Interpret the output. | 10 |
| **5** | A car company planned to design a product considering four attributes namely Fuel type, Transmission, Category and Price. Mentioned below the attributes and their levels.   |  |  |  |  | | --- | --- | --- | --- | | **Fuel type**   1. Petrol 2. Diesel 3. CNG 4. EV | **Transmission**   1. Manual 2. Automatic | **Category**   1. Sedan 2. Hatchback 3. SUV | **Price**   1. 15 Lakhs 2. 25 Lakhs 3. 35 Lakhs |   Using the four attributes and their levels, the company generated some combinations using orthogonal design. The combinations were considered as questionnaire and circulated among the customers. After getting the response, the company ran a conjoint analysis to finalize the product. The questionnaire and the output tables are given below.   |  |  | | --- | --- | | **Importance value** | | | **Fuel type** | 34 | | **Transmission** | 16 | | **Category** | 24 | | **Price** | 26 |  |  |  |  | | --- | --- | --- | | **Utilities** | | | |  |  | Utility Estimate | | **Fuel type** | Petrol | -0.035 | | Diesel | 0.240 | | CNG | -0.320 | | EV | 0.115 | | **Transmission** | Manual | 0.153 | | Automatic | -0.153 | | **Category** | Sedan | -0.057 | | Hatchback | 0.231 | | SUV | -0.174 | | **Price** | 15 Lakhs | 0.070 | | 25 Lakhs | 0.303 | | 35 Lakhs | -0.373 | | **(Constant)** |  | 8.497 |  |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | **Fuel type** | **Transmission** | **Category** | **Price** | | Product 1 | EV | Manual | Sedan | 3500000 | | Product 2 | Petrol | Automatic | SUV | 1500000 | | Product 3 | Diesel | Automatic | Sedan | 3500000 | | Product 4 | EV | Automatic | Hatchback | 1500000 | | Product 5 | EV | Automatic | SUV | 2500000 | | Product 6 | CNG | Automatic | Sedan | 1500000 | | Product 7 | CNG | Manual | Hatchback | 1500000 | | Product 8 | Diesel | Automatic | Sedan | 1500000 | | Product 9 | Diesel | Manual | Hatchback | 2500000 | | Product 10 | EV | Manual | Sedan | 1500000 | | Product 11 | Petrol | Manual | Sedan | 1500000 | | Product 12 | Diesel | Manual | SUV | 1500000 | | Product 13 | Petrol | Manual | Sedan | 2500000 | | Product 14 | CNG | Manual | SUV | 3500000 | | Product 15 | Petrol | Automatic | Hatchback | 3500000 | | Product 16 | CNG | Automatic | Sedan | 2500000 | | Product 17 | EV | Automatic | SUV | 3500000 | | Product 18 | Diesel | Automatic | SUV | 3500000 | | Product 19 | EV | Automatic | Sedan | 3500000 |   The questionnaire does not contain any holdout cases.   1. Interpret the utilities and importance values tables. 2. Finalise the product. | 10 |
| **6** | A company is analyzing data to classify individuals into two categories based on specific attributes like glucose levels, BMI, and age. A decision tree model is used to make predictions by splitting data at various thresholds. For instance, if an individual's glucose level is above 127.5, the model checks further conditions like BMI and glucose levels to determine the final classification. The goal is to accurately group individuals based on patterns found in the data.  A diagram of a number of numbers  AI-generated content may be incorrect.   1. Explain the concept of entropy and information gain in decision trees and its role in determining the best split at each node. 2. Interpret the meaning of the class labels (0 and 1) in this decision tree and describe how they influence the classification at different levels. 3. Analyse the path from the root to a leaf node where class = 1. What does this path tell you about the conditions required for classification into class 1? 4. Compare the decision tree's performance for classifying instances when glucose ≤ 127.5 (left subtree) and glucose > 127.5 (right subtree) based on the given entropy values. 5. The given decision tree is built using attributes like glucose, BMI, and age. If the tree were allowed to grow deeper with more splits, how could this lead to overfitting? Explain in terms of the bias-variance tradeoff and suggest a possible way to prevent overfitting in decision trees. | 10 |