**ST. XAVIER’S COLLEGE (AUTONOMOUS)**

**SEM-III: BDA**

**End Sem Examination (Nov-2021)**

**Assignment Test**

|  |  |  |  |
| --- | --- | --- | --- |
| **Subject: DATA VISUALIZATION AND MODELING IN OPERATIONS MANAGEMENT(PBD- 3803)** | | | **Max Marks: 60** |
| **Date of assignment test: 24/11/2021** | | **Roll No.: \_\_\_\_\_\_\_\_\_\_\_\_\_\_** | **Mode of submission: In Moodle as an assignment submission** |
| **Instructions :**  1. Submit your answer script in Moodle **on or before** **5 PM today**  2. **Write your roll number on the first page of the answer script.**  4. **Rename your data and all variables in the data** with your roll number (i.e., data\_roll number and variable\_roll number) before start doing analysis in Q2.  5. Code should be written just above the output that you are reproducing in the file for evaluation in Q2.  6. Plots (if any) should be reproduced for evaluation in the answer script.  7. Attempt both Q1 and Q2. | | | |
| Q1 | Design a Dashboard and Story point for the following sheets. By taking suitable data store. Also, produce the results with the necessary interpretation.  1. Interactive Filter action on summary levels of sum and profit on dual axis.  2. Action filters on country’s sales with the hover option  3. Bar and stacked representation with parameter usage for Top 3 Sales & Bottom 4 profit along with the index.  4. Make an interactive treemap with the help of filters reflected on region-wise profit.  Do the necessary formatting for worksheets, dashboards, and story point to make it more attractive. | | |
| Q2 | The dataset *ESE 2* is a healthcare dataset which has 14635 observations with several attributes. The objective is to build a model that predicts the presence of complications of surgery of the patient. The data description is given in sheet 2 of excel sheet.  Answer the following questions.   1. Obtain the proportion of patients who had complications based on asa\_status. Which among these seems to have an indication of post-surgery complication according to the given data? Justify your answer.   (b)Logistic regression model   1. Build a binary logistic regression model to predict the probability of having complications of surgery of the patient based on the predictors. Comment on the overall model significance. 2. Find out and report which variables are statistically significant in the logistic regression model built in (i). 3. Build a new logistic regression model using only significant features. Report the model diagnostics followed to build this model. 4. Write an estimated logistic regression model obtained in (iii) 5. Use Youden’s index to find the most optimal cut-off probability value for the best model chosen in (iii).   (c) Decision tree classifier model   1. Build Random Forest decision tree and clearly identify and report predictor which is classifying the patient having complications of post-surgery. 2. Also plot that decision tree and report the most important splitting criteria (rule). 3. Use variance importance plot, report the variables which are classifying the surgery complication.   (d) Compare the logistic regression model and decision tree classifier performance using confusion matrix with specific accuracy measures or ROC and AUC?  OR | | |
| Q2 | (a)Criminologists are interested in the effect of punishment regimes on crime rates. This has been studied using aggregate data on 47 states of the USA for 1960. The data set given in the file “Crime” contains the following variables. Assume that none of the predictor variables are highly correlated.   |  |  | | --- | --- | | Variable | Description | | M | Percentage of males aged 14–24 in total state population | | So | Indicator variable for a southern state | | Ed | Mean years of schooling of the population aged 25 years or over | | Po1 | Per capita expenditure on police protection in 1960 | | Po2 | Per capita expenditure on police protection in 1959 | | LF | Labor force participation rate of civilian males in the age group 14-24 | | MF | Number of males per 100 females | | Pop | State population in 1960 in hundred thousands | | NW | Percentage of nonwhites in the population | | U1 | Unemployment rate of urban males 14-24 | | U2 | Unemployment rate of urban males 35-39 | | Wealth | Median value of transferable assets or family income | | Ineq | Percentage of families earning below half the median income (income inequality) | | Prob | Ratio of number of commitments to number of offenses (probability of imprisonment) | | Time | Average time in months served by offenders in state prisons before their first release | | Crime | Number of offenses per 100,000 population in 1960 (crime rate) |   Build a multiple linear regression model to predict the Crime based on the predictors given in the data   1. Which features have the symptom of multicollinearity? Report what action you have taken to treat multicollinearity. 2. Find outliers in the dataset using Cook’s distance and standardized residuals (greater than 2.5R). If required remove and report the observations from the dataset (maximum 2 for each for the time being). 3. Validate those assumptions of regression model you are trying to build. If all assumptions are meeting, go to (v) 4. If any of the assumptions in (iii) are not satisfied, try some transformation to validate the assumptions. 5. Which features are statistically significant in predicting the Crime? 6. Build a linear regression model with significant features and report model performance. 7. Write down the final model obtained by you. Can the model built by you used for point prediction? Why/Why not? Justify.   (b) The data “Smoking” contains measurements of weight and tar, nicotine, and carbon monoxide content for 25 brands of domestic cigarettes. The variables in the data are Brand name, Tar content (mg), Nicotine content (mg), Weight (g), Carbon monoxide content (mg).   1. Using all the variables except brand, run the k-means algorithm with k=2 to identify clusters within the data. Report the cluster membership of each observation. 2. Generate a silhouette plot of your cluster model and report the mean silhouette values for each cluster. Measure the goodness of the clusters using the mean silhouette values? Interpret the same. 3. Is there an observation in the particular brand which is clearly misclassified by looking at the silhouette value plot. If any, report the observation. | | |

**\* \* \* ALL THE BEST \* \* \***