**PREDICTIVE MODELLING**



**Gaurav Sood**

**PGP-DSBA Online July 2021**

**Date: 19th December, 2021**

**Table of Contents**

**Problem 1: Linear Regression…………………………………………………………………….7**

1.1.Read the data and do exploratory data analysis. Describe the data briefly. (Check the null values, Data types, shape, EDA, duplicate values). Perform Univariate and Bivariate Analysis. ………………...………………………………………8

**1.2** Impute null values if present, also check for the values which are equal to zero. Do they have any meaning or do we need to change them or drop them? Check for the possibility of combining the sub levels of a ordinal variables and take actions accordingly. Explain why you are combining these sub levels with appropriate reasoning……………………………………………………………………..…….………20

**1.3** Encode the data (having string values) for Modelling. Split the data into train and test (70:30). Apply Linear regression using scikit learn. Perform checks for significant variables using appropriate method from statsmodel. Create multiple models and check the performance of Predictions on Train and Test sets using Rsquare, RMSE & Adj Rsquare. Compare these models and select the best one with appropriate reasoning. ……………………………………………………………...……………..……22

**1.4**Inference: Basis on these predictions, what are the business insights and recommendations.…………………...……………………………………………………25

**Problem 2: Logistic Regression and LDA…….....……………………………………………26**

**2.1** Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, write an inference on it. Perform Univariate and Bivariate Analysis. Do exploratory data analysis. ………………………………………………….……………………...27

**2.2** Do not scale the data. Encode the data (having string values) for Modelling. Data Split: Split the data into train and test (70:30). Apply Logistic Regression and LDA (linear discriminant analysis). …………………………………………………………...……………...…37

**2.3**Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score for each model Final Model: Compare Both the models and write inference which model is best/optimized. …………………………………………………………………………………………………...……38

**2.4** Inference: Basis on these predictions, what are the insights and recommendations. ………………………….………………………………………………………………………….....46

**List of Tables**

Table no.1: Dataset Sample………………………………………………………………………………………………………………8

Table no.2: Dataset Sample…………………..…………………………………………………………………………………………..8

Table no.3: EDA…………………………….…………………………………………………………………………………………………..9

Table no.4: Null Value Check…………………………………………………………………………………………………………….9

Table no.5: Data Description………..………………………………………………………………………………………………...10

Table no.6: Dataset Sample………………………………………………………………………………………………..……………20

Table no.7: Dataset Sample..………………..………………………………………………………………………………………..20

Table no.8: EDA……………………………………………….………………………………………………………………………………27

Table no.9: Null Value Check…………………..……………………………………………………………………………………….27

Table no.10: Data Description……….………………………………………………………………………………………………...27

Table no. 11: Null Value Check …………………..……………………………………………………………………………………28

Table no. 12: Data Description…………………………………………………………………………………………………………28

Table no. 13: Logit Train Set…………………….………………………………………………………………………………………38

Table no. 14: Logit Test Set…….…………………..……………………………………………………………………………………39

Table no. 15: LDA Train Set………………………………………………………………………………………………………………41

Table no. 16: LDA Test Set…….…………………………………………………………………………………………………………42

**List of Figures**

Figure no. 1: ……………………………………………………………………………………………………………………………………11

Figure no. 2: ……………………………………………………………………………………………………………………………………11

Figure no. 3: ……………………………………………………………………………………………………………………………………12

Figure no. 4: ……………………………………………………………………………………………………………………………………12

Figure no. 5: ……………………………………………………………………………………………………………………………………13

Figure no. 6: ……………………………………………………………………………………………………………………………………13

Figure no. 7: ……………………………………………………………………………………………………………………………………14

Figure no. 8: ……………………………………………………………………………………………………………………………………14

Figure no. 9: ……………………………………………………………………………………………………………………………………15

Figure no. 10: ………………………………………………………………………………………………………………………………….15

Figure no. 11: ………………………………………………………………………………………………………………………………….16

Figure no. 12: ………………………………………………………………………………………………………………………………….16

Figure no. 13: ………………………………………………………………………………………………………………………………….17

Figure no. 14: ………………………………………………………………………………………………………………………………….18

Figure no. 15: ………………………………………………………………………………………………………………………………….19

Figure no. 16: ………………………………………………………………………………………………………………………………….23

Figure no. 17: ………………………………………………………………………………………………………………………………….23

Figure no. 18: ………………………………………………………………………………………………………………………………….24

Figure no. 19: ………………………………………………………………………………………………………………………………….25

Figure no. 20: ………………………………………………………………………………………………………………………………….29

Figure no. 21: ………………………………………………………………………………………………………………………………….30

Figure no. 22: ………………………………………………………………………………………………………………………………….30

Figure no. 23: ………………………………………………………………………………………………………………………………….31

Figure no. 24: ………………………………………………………………………………………………………………………………….31

Figure no. 25: ………………………………………………………………………………………………………………………………….32

Figure no. 26: ………………………………………………………………………………………………………………………………….32

Figure no. 27: ………………………………………………………………………………………………………………………………….33

Figure no. 28: ………………………………………………………………………………………………………………………………….34

Figure no. 29: ………………………………………………………………………………………………………………………………….34

Figure no. 30: ………………………………………………………………………………………………………………………………….35

Figure no. 31: ………………………………………………………………………………………………………………………………….35

Figure no. 32: ………………………………………………………………………………………………………………………………….38

Figure no. 33: ………………………………………………………………………………………………………………………………….39

Figure no. 34: ………………………………………………………………………………………………………………………………….39

Figure no. 35: ………………………………………………………………………………………………………………………………….40

Figure no. 36: ………………………………………………………………………………………………………………………………….41

Figure no. 37: ………………………………………………………………………………………………………………………………….42

Figure no. 38: ………………………………………………………………………………………………………………………………….42

Figure no. 39: ………………………………………………………………………………………………………………………………….43

Figure no. 40: ………………………………………………………………………………………………………………………………….44

Figure no. 41: ………………………………………………………………………………………………………………………………….44

Figure no. 42: ………………………………………………………………………………………………………………………………….45

**Problem 1: Linear Regression**

You are hired by a company Gem Stones co ltd, which is a cubic zirconia manufacturer. You are provided with the dataset containing the prices and other attributes of almost 27,000 cubic zirconia (which is an inexpensive diamond alternative with many of the same qualities as a diamond). The company is earning different profits on different prize slots. You have to help the company in predicting the price for the stone on the bases of the details given in the dataset so it can distinguish between higher profitable stones and lower profitable stones so as to have better profit share. Also, provide them with the best 5 attributes that are most important.

**Data Dictionary:**

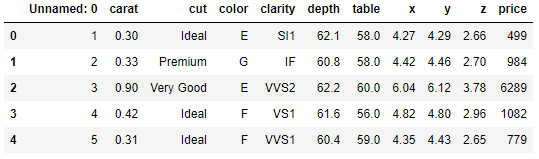
|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| Carat | Carat weight of the cubic zirconia. |
| Cut | Describe the cut quality of the cubic zirconia. Quality is increasing order Fair, Good, Very Good, Premium, Ideal. |
| Color | Colour of the cubic zirconia.With D being the worst and J the best. |
| Clarity | Clarity refers to the absence of the Inclusions and Blemishes. (In order from Worst to Best in terms of avg price) IF, VVS1, VVS2, VS1, VS2, Sl1, Sl2, l1 |
| Depth | The Height of cubic zirconia, measured from the Culet to the table, divided by its average Girdle Diameter. |
| Table | The Width of the cubic zirconia's Table expressed as a Percentage of its Average Diameter. |
| Price | the Price of the cubic zirconia. |
| X | Length of the cubic zirconia in mm. |
| Y | Width of the cubic zirconia in mm. |
| Z | Height of the cubic zirconia in mm. |

**1.1. Read the data and do exploratory data analysis. Describe the data briefly. (Check the null values, Data types, shape, EDA, duplicate values). Perform Univariate and Bivariate Analysis.**

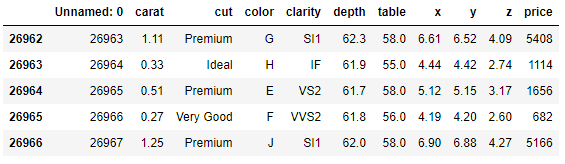
First, we load all the necessary libraries for model building.

Then, we read the head and tail of the dataset to check whether the data has been properly fed.

**Sample of the Dataset**

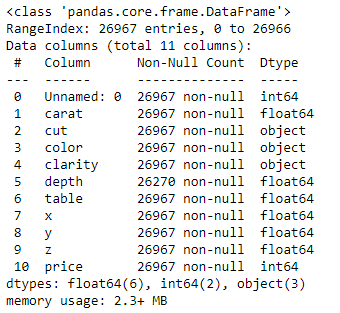


**Table no. 1: Dataset Sample**



**Table no. 2: Dataset Sample**

**Exploratory Data Analysis**

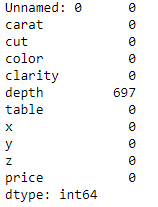




**Table no. 3: EDA**

The given dataset has 26967 entries across 11 columns, with six being of float type, two being of integer type and three being of object type.

**Checking for Null Values**



**Table no. 4: Null Value Check**

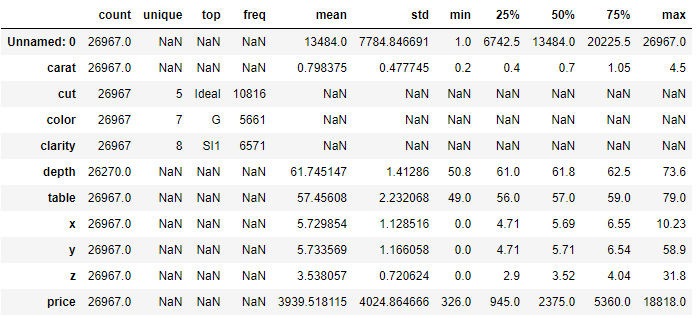
The given dataset is found to have 697 null values, all of them in the ‘depth’ variable.

**Checking for Duplicates**



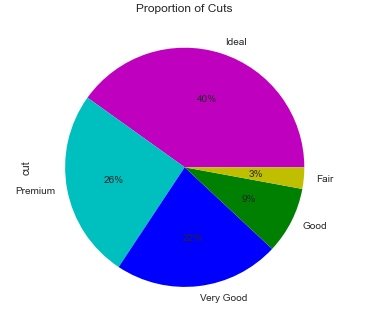
There are no duplicate rows in the given dataset.

**Data Description**

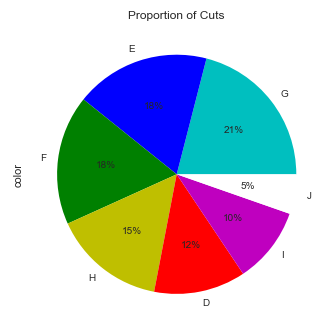


**Table no. 5: Data Description**

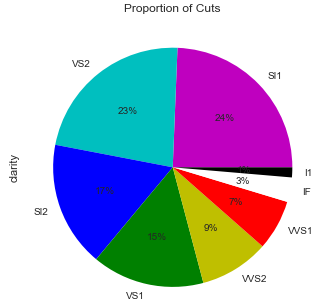
* The dataset contains 26967 observations over 11 variables. The dependent variable is ‘price’ of the zirconia stones.
* There are three categorical variables: cut, color and clarity, which represent the respective quality of the stone from lower to higher order.
* The other variables are of continuous numeric types, where x, y and z indicates dimension of the cubic stones (length, width and height).
* The average price of a zirconia stone is approximately 3940, and with a price range of 326 to 18,818 indicates the presence of outliers and a skewed distribution.

****

**Figure no. 1**

****

**Figure no. 2**

****

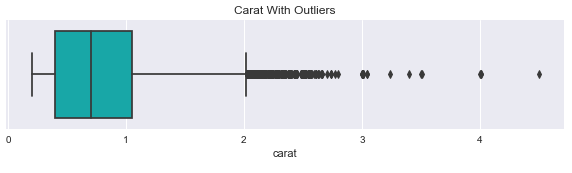
**Figure no. 3**

****

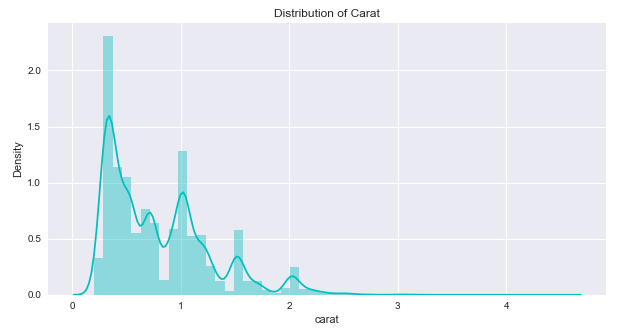
**Figure no. 4**

****

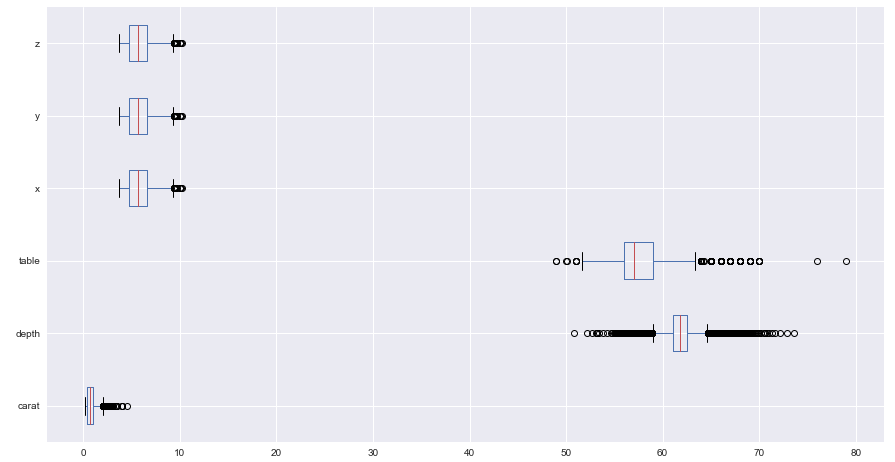
**Figure no. 5**

****

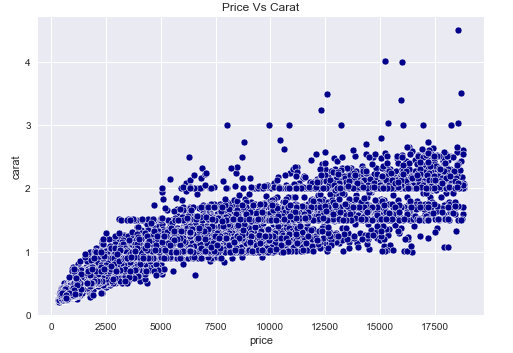
**Figure no. 6**

****

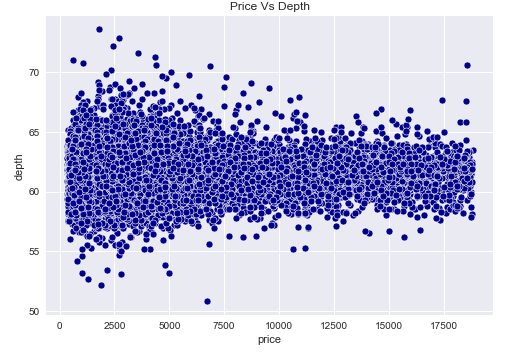
**Figure no. 7**

****

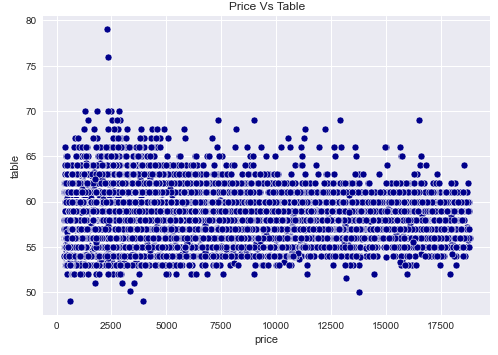
**Figure no. 8**

****

**Figure no. 9**

****

**Figure no. 10**

****

**Figure no. 11**

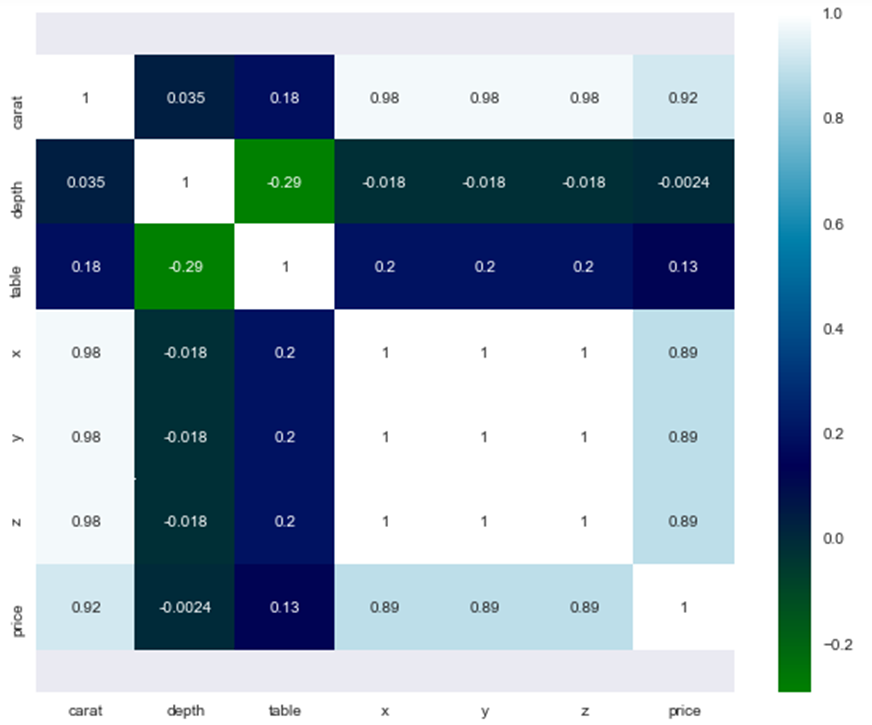
****

**Figure no. 12**

From the above figures, we can understand that:

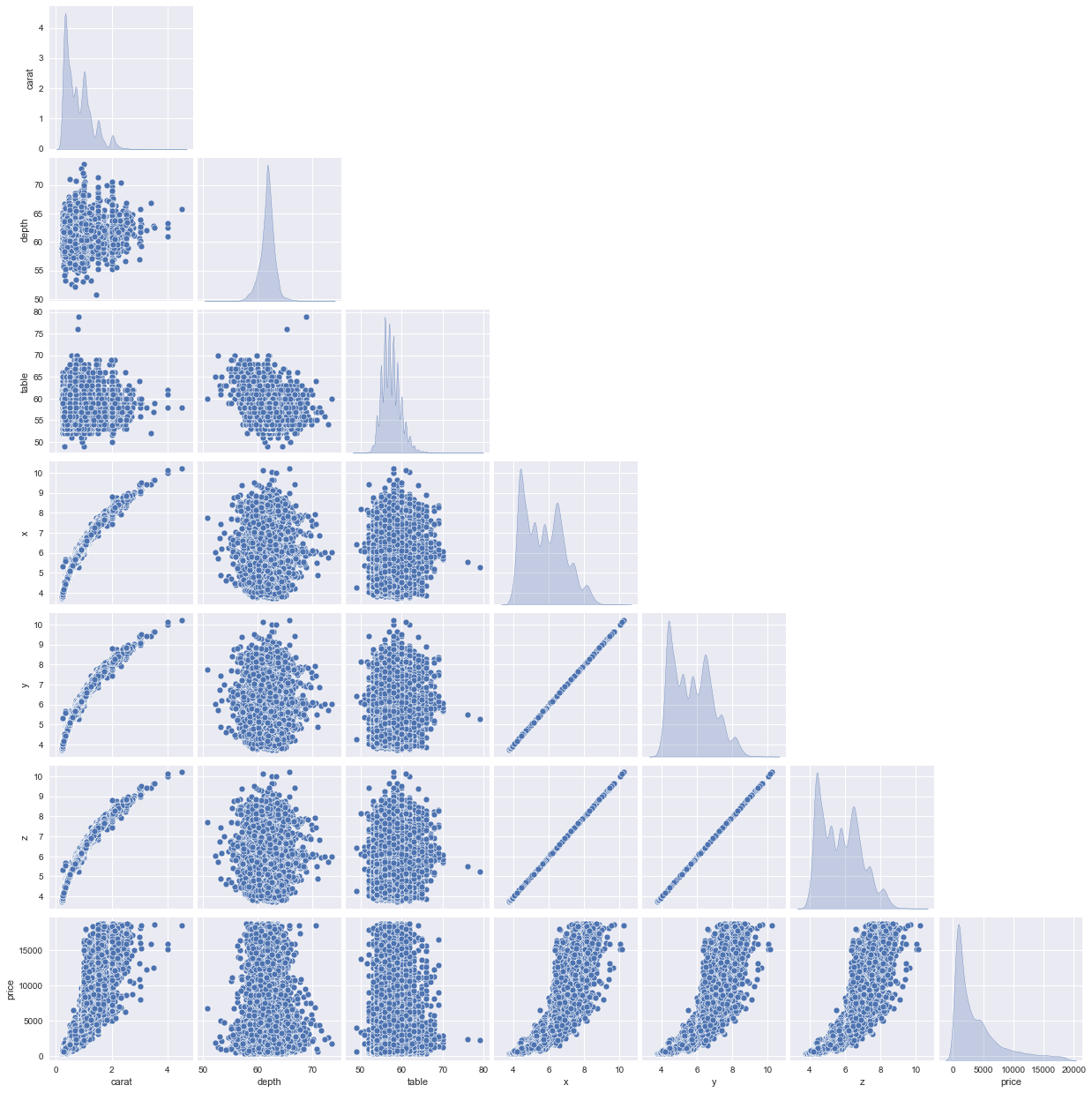
* The categorical quality variables have a proportion as shown in the pie charts. The highest grade for Cut is the ‘Ideal’ category with 10,816 stones. ‘G’ grade for Color is the most commonly occurring data point with 5661 stones, and similarly ‘Sl1’ for Clarity with 6571 stones.
* All the continuous variables in the dataset have a significant number of outliers and these variables are in different scales of values, which are removed so that the outliers do not influence the regression line towards them.
* The scatterplots showing the relationship between the response variable ‘price’ to other variables indicate that ‘price’ has a collinear relationship with the ‘carat’ variable and the dimensional variables ‘x’, ‘y’ and ‘z’. However, there is no significant relation observed with the variables ‘depth’ and ‘table’.

**Correlation Heatmap**



**Figure no. 13**

**Pairplot**

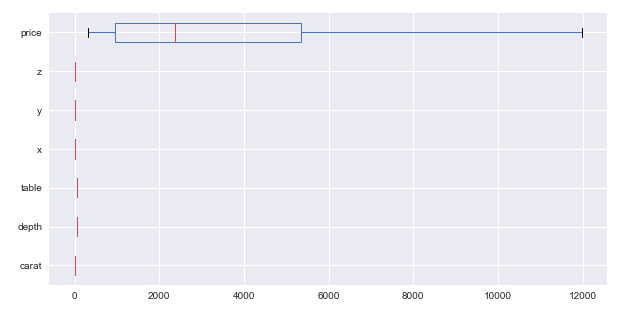
****

**Figure no. 14**

* From the above pairplot we can see the distribution of the continuous variables and the collinearity between them.
* From the heatmap of the Pearson correlation of continuous variables, we can see that the response variable ‘price’ is highly correlated with ‘carat’ as well as the dimension variables (x,y,z).
* Also, the collinearity of ‘carat’ is very high with the dimension variables as evident from the heatmap & pairplot, indicating that ‘carat’ is influenced by the size of the stone.
* The collinearity and correlation of x,y and z indicate that for a given observation the values of x, y and z will be the same. This is because the stones are cubic in shape.

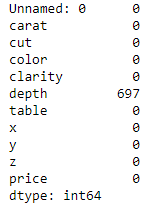
Outliers are removed from the dataset and the resulting boxplot is shown below.

**Boxplot after removing outliers**

****

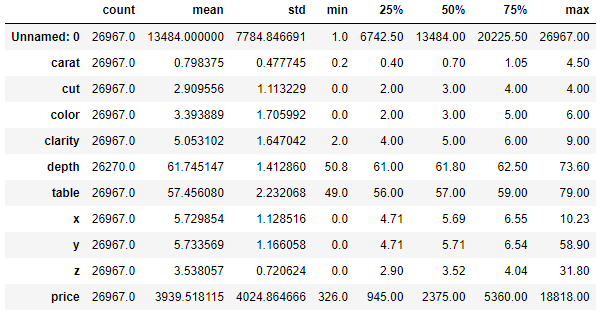
**Figure no. 15**

**1.2 Impute null values if present, also check for the values which are equal to zero. Do they have any meaning or do we need to change them or drop them? Check for the possibility of combining the sub levels of a ordinal variables and take actions accordingly. Explain why you are combining these sub levels with appropriate reasoning.**



**Table no. 6: Null Value Check**

There are 697 null value observations in the variable depth, which are imputed with the mean of the variable.



**Table no. 7: Data Description**

The dimension variables, x, y and z show a minimum value of zero, which are dropped as the shape of the stone cannot be zero and there are only 3 such observations.

**Ordinal Encoding**

* All the three categorical variables such as cut, color and clarity represent the respective quality of the stone from lower to higher order.
* Original Encoding is applied in this case as observation with higher quality takes priority over the lesser ones in training the model.
* The variables are encoded from zero to higher values in the increasing order of the respective quality attribute.

**Scaling the Data**

* The different factors in the given dataset are in different scales, which can be standardized.
* We shall use two methods called LinearRegression function from sklearn and OLS function from statsmodels for this business problem.
* Both of them use ordinary least square method which may not be influenced by scaling.
* Scaling would be effective if the algorithm used gradient descent method to converge faster.
* Scaling using z-score has been applied to evaluate the impact of scaling.
* No significant improvement is observed with scaling.

**1.3 Encode the data (having string values) for Modelling. Split the data into train and test (70:30). Apply Linear regression using scikit learn. Perform checks for significant variables using appropriate method from statsmodel. Create multiple models and check the performance of Predictions on Train and Test sets using Rsquare, RMSE & Adj Rsquare. Compare these models and select the best one with appropriate reasoning.**

**Ordinal Encoding**

* All the three categorical variables such as cut, color and clarity represent the respective quality of the stone from lower to higher order.
* Original Encoding is applied in this case as observation with higher quality takes priority over the lesser ones in training the model.
* The variables are encoded from zero to higher values in the increasing order of the respective quality attribute.

**Split Data**

The data was split into 70:30 ratio for train and test datasets. The predictor variables as X and response variable ‘price’ as Y and in train and test datasets were created.

**Modelling**

For this exercise, two linear regression methods from sklearn and statsmodels packages have been applied. OLS method from statsmodel has been applied to get an R like summary of the model.

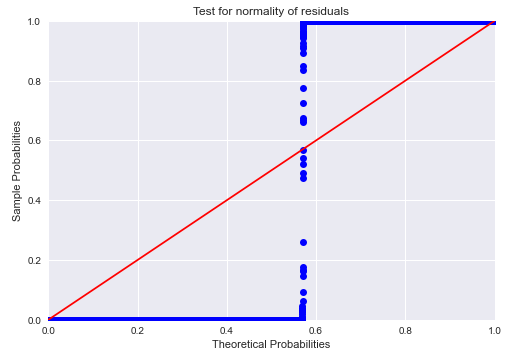
**Model Performance**

The predicted labels and test labels are plotted on a scatterplot to validate the predictions.

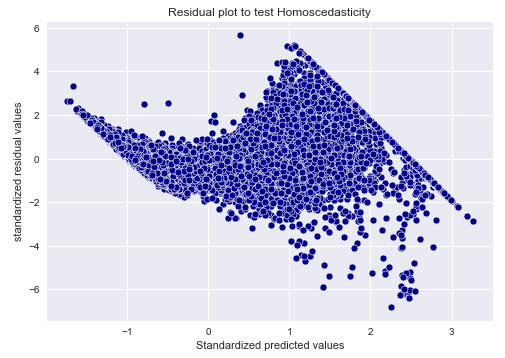
The assumptions of linear regression such as normality of residuals and homoscedasticity of residuals were also validated using a probability plot and residual plot, respectively.

****

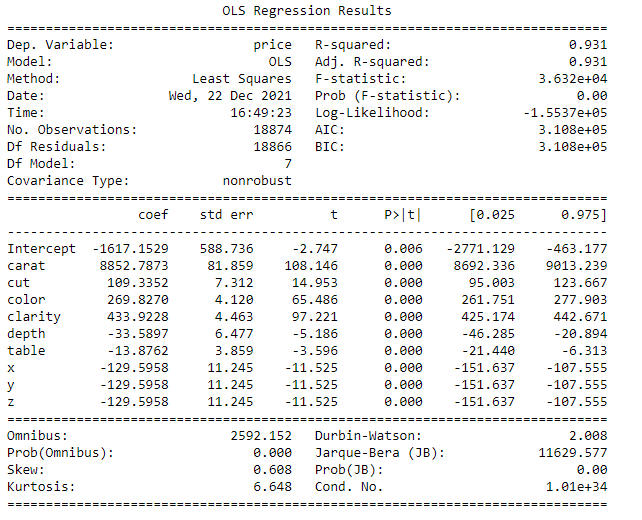
**Figure no. 16**

****

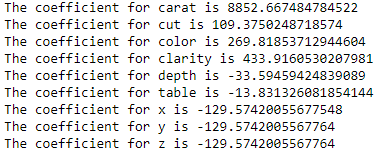
**Figure no. 17**

****

**Figure no. 18**

****

**1.4 Inference: Basis on these predictions, what are the business insights and recommendations.**

****

**Figure no. 19**

Following are some of the business insights and future recommendations for the given business problem:

* The regression coefficient of the predictor variable ‘carat’ is found to be highest (8852.67), meaning that for 1 unit increase in carat, the price of a Zirconia stone increases by 8852 units.
* ‘Clarity’ is found to be the next most important influencer in deciding the price with a coefficient of 433.92.
* ‘Color’ and ‘Cut’ are the next influencers, with coefficient values of 269.82 and 109.38 respectively.
* The factors ‘Depth’ and ‘Table’ are found to be quite insignificant in deciding the price of Zirconia stones.
* The test for homoscedasticity was done by plotting the residuals and it was found that it forms a funnel like shape. This is indicative that the residuals are heteroscedastic in nature, which violates one of the assumptions.
* This gives us scope for further improvement and would recommend further features to be added to improve the prediction of the response variable ‘price’.
* The recommendation for the company to identify higher profitable stones is to classify them based on the carat, dimension, clarity, color and cut in the respective order of significance.

**Problem 2: Logistic Regression and LDA**

You are hired by a tour and travel agency which deals in selling holiday packages. You are provided details of 872 employees of a company. Among these employees, some opted for the package and some didn't. You have to help the company in predicting whether an employee will opt for the package or not on the basis of the information given in the data set. Also, find out the important factors on the basis of which the company will focus on particular employees to sell their packages.

**Data Dictionary:**

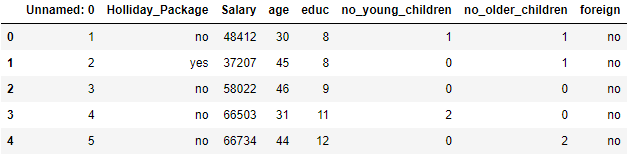
|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| Holiday\_Package | Opted for Holiday Package yes/no? |
| Salary | Employee salary |
| age | Age in years |
| edu | Years of formal education |
| no\_young\_children | The number of young children (younger than 7 years) |
| no\_older\_children | Number of older children |
| foreign | foreigner Yes/No |

**2.1 Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, write an inference on it. Perform Univariate and Bivariate Analysis. Do exploratory data analysis.**

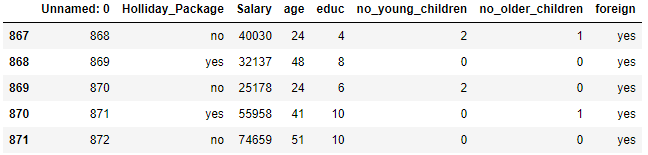
First, we load all the necessary libraries for model building.

Then, we read the head and tail of the dataset to check whether the data has been properly fed.

**Sample of the Dataset**

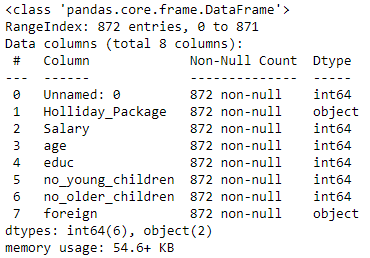
****

**Table no. 8: Dataset Sample**

****

**Table no. 9: Dataset Sample**

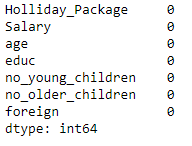
**Exploratory Data Analysis**

****

**Table no. 10: EDA**

The dataset contains 872 rows across 8 columns, of which six are of integer type and two are of object type.

**Checking for Null Values**

****

**Table no. 11: Null Value Check**

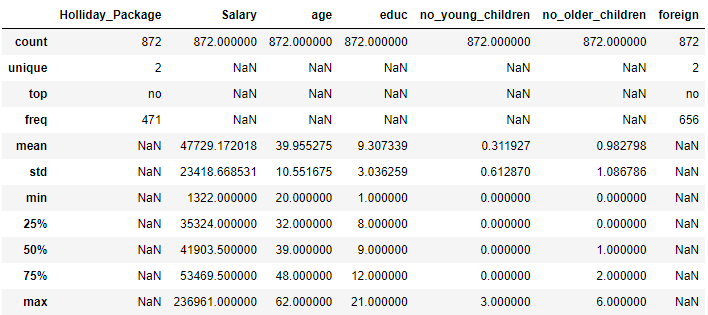
There are no null values in the given dataset.

**Checking for Duplicates**



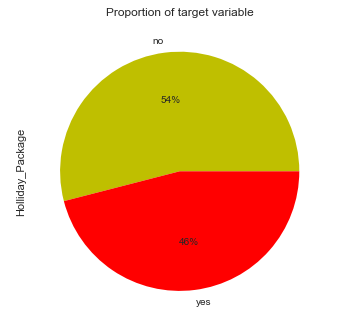
There are no duplicate values in the given dataset.

**Data Description**

****

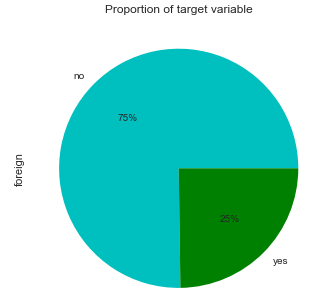
**Table no. 12: Data Description**

* The dataset contains 872 observations over 7 variables. The dependent variable is Holliday\_Package, indicating whether an employee has opted for a holiday package or not.
* Salary, Age, Education, no. of young children and no. of older children are continuous independent variables.
* The categorical variable foreign indicates whether the employee is a foreign national or not.
* The average salary of an employee is approximately 47729.2, with the maximum being 236961 and minimum 1322. The range of salary indicates presence of outliers.
* The average age of an employee is about 40 years, with the oldest being of 62 years and youngest being just 20 years of age.
* On average, the employees had over 9 years of formal education, with the highest being 21 years and lowest being only 1 year.

****

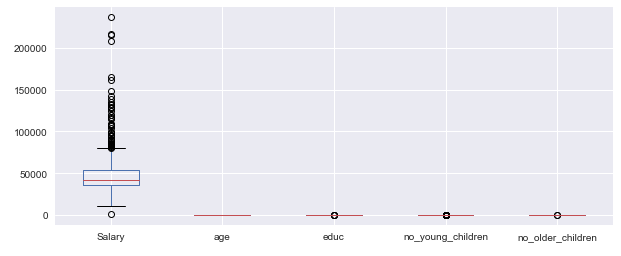
**Figure no. 20**

The proportion of the target variable ‘Holliday\_Package’ indicates that 46% of the employees opted for the holiday package and the other 54% did not.

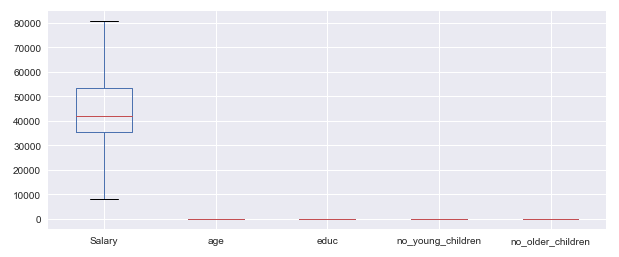
****

**Figure no. 21**

The above figure shows us that the proportion of foreign nationals in the company is 25%.

****

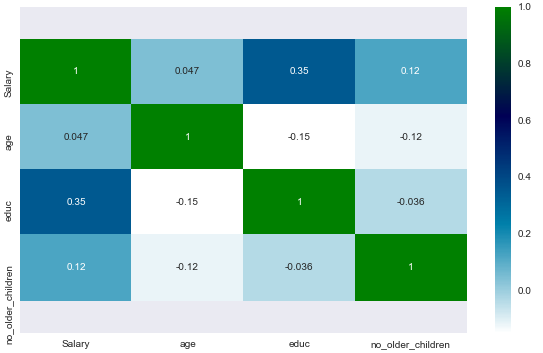
**Figure no. 22**

****

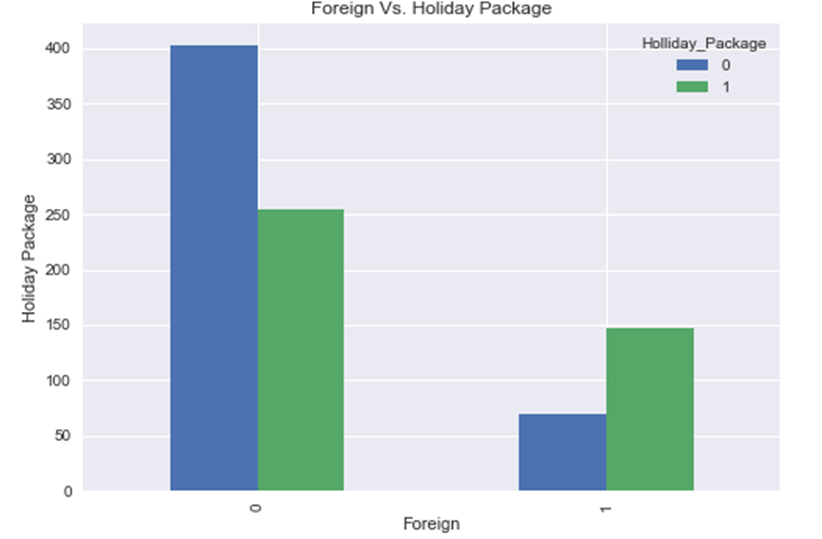
**Figure no. 23**

****

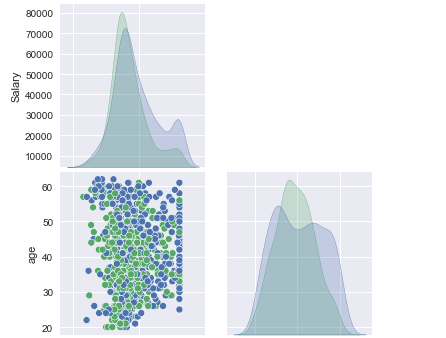
**Figure no. 24**

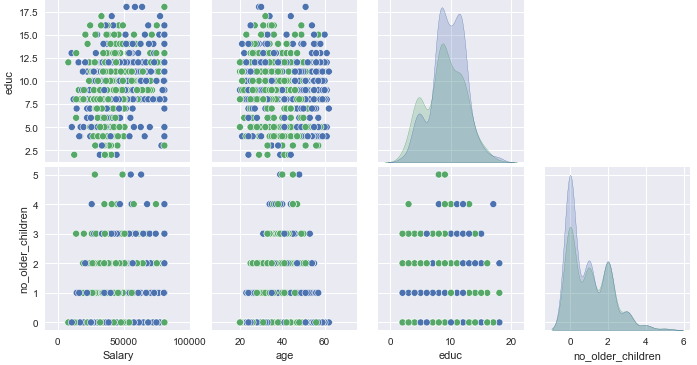
****

**Figure no. 25**

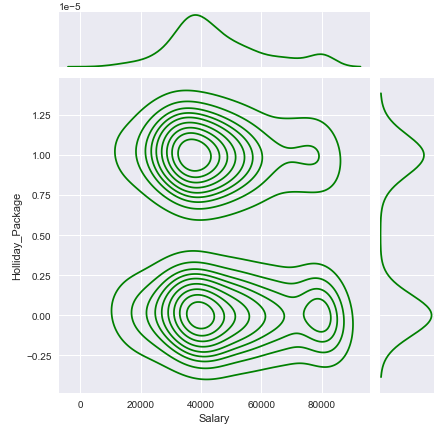
****

**Figure no. 26**

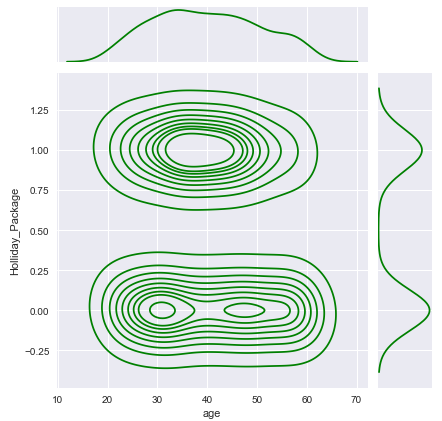
****

****

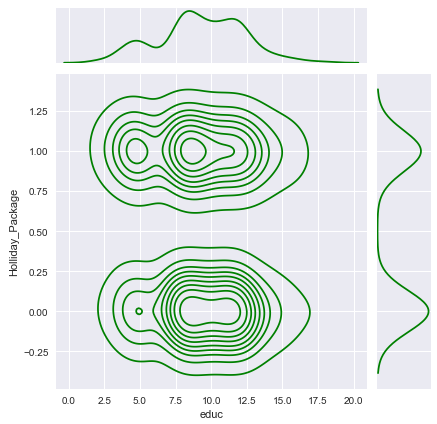
**Figure no. 27**

****

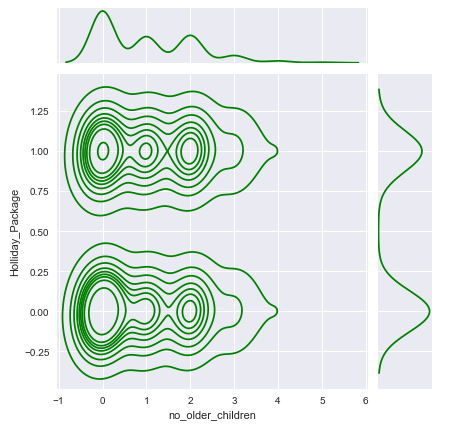
**Figure no. 28**

****

**Figure no. 29**

****

**Figure no. 30**

****

**Figure no. 31**

* From the plots, we can infer that there is no significant correlation between salary, age, years of education and no. of older children to an employee opting for a holiday package or not.
* Among those who opted for a holiday package, their salary is more concentrated between the 30,000 to 50,000 range. However, this does not indicate any relation towards opting for a holiday package or not.
* Similarly, age does not have any significant correlation with an employee’s decision to opt for a holiday package. But there is a higher density in the age group of 30-50 years who opted for a holiday package.
* The years of formal education variable is found to be a poor predictor in deciding whether an employee opts for a holiday package or not.
* The no. of older children of an employee has also turned out to be a poor predictor as the distribution on the status of holiday package is completely overlapping.

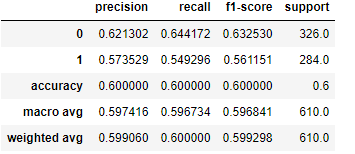
**2.2 Do not scale the data. Encode the data (having string values) for Modelling. Data Split: Split the data into train and test (70:30). Apply Logistic Regression and LDA (linear discriminant analysis).**

* The categorical variables Holliday\_Package and foreign are encoded and converted to integers.
* The target positive case ‘yes’ for Holliday\_Package is 1 and the negative case is 0.
* The data is split into 70:30 ratio in train and test datasets.
* LogisticRegression() method from sklearn package is then used for building the model.
* To optimise the hyper-parameters, GridSearchCV is used.
* LinearDiscriminantAnalysis() from sklearn is used to build the LDA model.

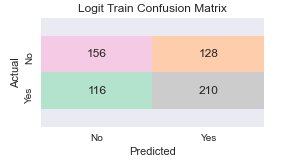
**2.3 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score for each model Final Model: Compare Both the models and write inference which model is best/optimized.**

**LOGISTIC REGRESSION**

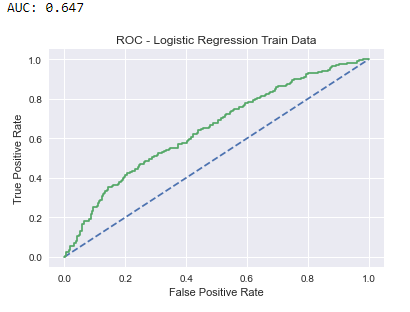
**TRAIN**

****

**Table no. 13: Logit Train Set**

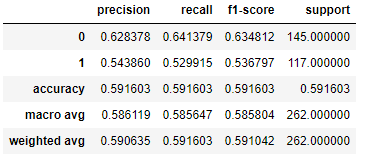
****

**Figure no. 32**

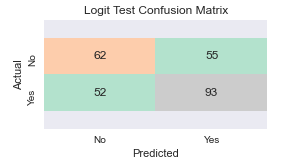
****

**Figure no. 33**

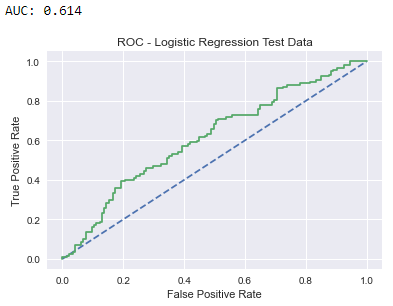
**TEST**

****

**Table no. 14: Logit Test Set**

****

**Figure no. 34**

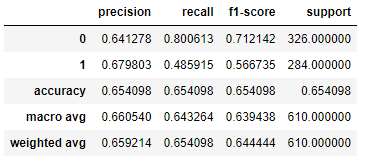
****

**Figure no. 35**

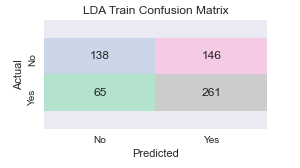
* The Logistic Regression Model was optimised to improve the recall rate for the positive target case.
* While the accuracy and precision dropped a little, the recall rate improved significantly.
* The model produced an accuracy of 60% in train and 59% in test.
* The precision for the positive target case in train is 58%, whereas in test it is 54%.
* The recall rate for the positive target case in train is 55% and in test is 53%. It is considered as the most significant performance measure for this particular business case.
* While the recall rate was optimised, the Area Under Curve (AUC) value also reduced significantly.
* For train the AUC is 0.65 and for test it is 0.61.
* Based on the accuracy values, the model seems to be a right fitting one.

**LINEAR DISCRIMINANT ANALYSIS**

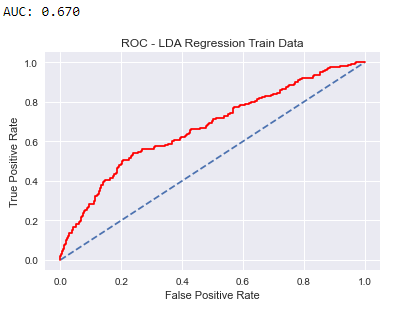
**TRAIN**

****

**Table no. 15: LDA Train Set**

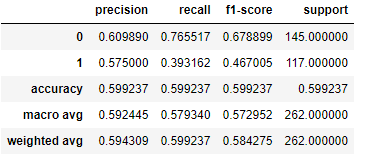
****

**Figure no. 36**

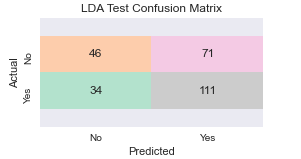
****

**Figure no. 37**

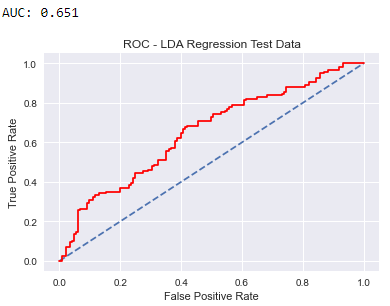
**TEST**

****

**Table no. 16: LDA Test Set**

****

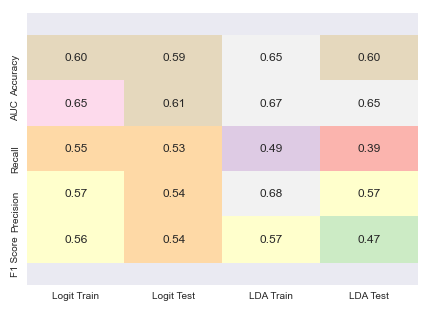
**Figure no. 38**

****

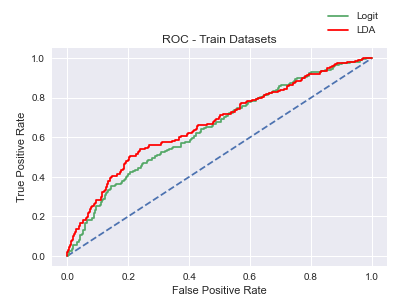
**Figure no. 39**

* The LDA model produced an accuracy of 65% in train and 60% in test.
* The precision for positive target case in train is 68% and in test it is 58%.
* The recall rate for positive target case in train is 49% and in test it is 39%. It is considered as the most significant performance measure for this particular business problem.
* For train the AUC is 0.67 and for test it is 0.65.
* The recall rate came out to be too low and the model is found to be overfitting in terms of accuracy.
* The derived model is not reliable to predict the target cases.

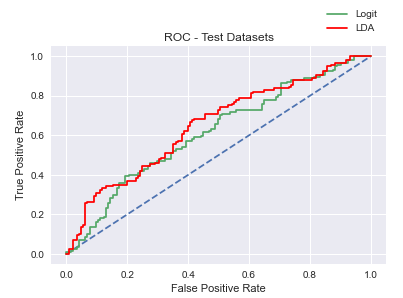
**COMPARISON**

****

**Figure no. 40**

****

**Figure no. 41**

****

**Figure no. 42**

* The accuracy of the LDA model is found to be higher than that of the Logit model, 60% and 59% respectively for the test dataset.
* The precision for the test dataset is also higher in the LDA model than the Logit model, 57% and 54% respectively.
* The AUC score on the test dataset from the LDA model is significantly higher than the Logit model, 0.65 and 0.61 respectively.
* But, considering the business case the ability to recall the positive target case is more important than the other performance measures.
* The recall score is higher from the Logit model at 53% for the test dataset, whereas the recall score from the LDA model is only 39% which is well below par.
* The overall F1 score is also higher in the Logit model at 54% for the test data. It is only 47% for the test data in LDA model.
* The ROC curve shows a better coverage in the LDA model, but based on the recall and F1 score Logit model is chosen as the final model for this business case.

**2.4 Inference: Basis on these predictions, what are the insights and recommendations.**

Following are some of the obtained insights and future recommendations for the given problem:

* None of the given predictor variables could differentiate between the positive and negative target cases, resulting in both the models being unable to give any promising results.
* The predictions are widely hit and miss and not reliable in predicting the target response variable.
* The recall rate of 53% indicates that the model has been able to recollect only half the actual cases of employees opting for a holiday package.
* The travel company can provide tailored packages based on age groups in order to attract employees of different ages to opt for suitable packages.
* Value add-ons could be considered depending on the number of children the employee has and the age of the children.
* Foreign nationals can be further enticed with specific packages based on their interest which may vary from that of locals.
* It is imperative that the travel company identify new features which can positively correlate the pattern of opting for a holiday package or not, for training and helping in building a better model.
* Some examples of such features are gender, marital status, total household income (which may help us calculate disposable income), travel preferences (coastal areas/beaches, hill stations, historical sites, theme parks) etc.

**THE END**