Business Report

SMDM Project Business Report DSBA

***Sanjay Srinivasan***

***PGP-DSBA Online***

***JULY’ 21 Batch***

***Date: 19-12-2021***

# *INDEX*

|  |  |  |
| --- | --- | --- |
| ***S. No*** | ***Contents*** | ***Page* No** |
| **1.** | **Problem - 1** | 5 |
|  | Summary | 5 |
|  | Introduction | 5 |
|  | Data Description | 5 |
|  | Sample Dataset | 5 |
|  | Exploratory Data Analysis | 6 |
|  | Checking for missing values in the dataset | 6 |
|  | 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. | 6 |
|  | 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. | 9 |
|  | 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. | 11 |
|  | 4) Inference: Basis on these predictions, what are the business insights and recommendations | 15 |
|  |  |  |
| **2.** | **Problem - 2** | 16 |
|  | Summary | 16 |
|  | Introduction | 16 |
|  | Data Description | 16 |
|  | Sample Dataset | 16 |
|  | Exploratory Data Analysis | 17 |
|  | Checking for missing values in the dataset | 17 |
|  | 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. | 17 |
|  | 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). | 22 |
|  | 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. | 23 |
|  | 4) Inference: Basis on these predictions, what are the insights and recommendations. | 30 |

# *List Of Tables*

|  |  |  |
| --- | --- | --- |
| ***S.No*** | ***Content*** | ***Page No*** |
| 1.1 | Dataset sample | 5 |
| 1.2 | Datatypes of the variable | 6 |
| 1.3 | Check null values | 6 |
| 1.4 | Null values in x,y,z variable | 10 |
| 1.5 | After null values treatment | 10 |
| 1.6 | After scaling | 10 |
| 1.7 | Sample Encoded data | 11 |
| 1.8 | sample Dataframe without the TARGET variable | 11 |
| 1.9 | sample Dataframe with the TARGET variable | 11 |
| 2.1 | Dataset Sample | 16 |
| 2.2 | Datatypes of the variable | 17 |
| 2.3 | Check null values | 17 |
| 2.4 | Null values | 17 |
| 2.5 | Sample Dataset after dropping Unnamed :0 column. | 22 |
| 2.6 | Sample dataframe after Encoding | 22 |
| 2.7 | Train dataframe | 23 |
| 2.8 | Comparing Logistic Regression and Linear Discriminant Analysis results | 29 |

***List Of Figures***

|  |  |  |
| --- | --- | --- |
| ***S.No*** | ***Content*** | ***Page No*** |
| 1.1 | diamond colour count | 6 |
| 1.2 | Univariate Analysis | 7 |
| 1.3 | Jointplot for price vs. carat using Bivariate Analysis | 7 |
| 1.4 | Bivariate Analysis | 8 |
| 1.5 | Multivariate analysis of pairplot | 8 |
| 1.6 | Multivariate analysis for correlation | 9 |
| 1.7 | Multivariate analysis of plotting correlation in heatmap | 9 |
| 1.8 | Null values count | 9 |
| 1.9 | Before Scaling | 10 |
| 1.10 | After Scaling | 11 |
| 1.11 | Coefficient of independent variable | 12 |
| 1.12 | Intercept of our model | 12 |
| 1.13 | R – Square Value for Training data | 12 |
| 1.14 | R – Square Value for Testing data | 12 |
| 1.15 | Mean squared error for training data | 12 |
| 1.16 | Mean squared error for testing data | 12 |
| 1.17 | Variance Inflation Factor of our model | 12 |
| 1.18 | Coefficient of independent variable | 13 |
| 1.19 | R – Square Value for Training and Testing data | 13 |
| 1.20 | Intercept of our model | 13 |
| 1.21 | Mean squared error for Training data | 13 |
| 1.22 | Mean squared error for Testing data | 13 |
| 1.23 | Coefficient of independent variable | 13 |
| 1.24 | OLS Summary | 14 |
| 1.25 | OLS Summary Mean squared error | 14 |
| 1.26 | predicted Output vs. testing data | 14 |
| 1.27 | predicted Output vs. testing data Linear equation | 15 |
| 2.1 | Categorical value count | 18 |
| 2.2 | Number of duplicate rows | 18 |
| 2.3 | Univariate analysis | 18 |
| 2.4 | Univariate analysis for Holiday package | 19 |
| 2.5 | Univariate analysis for Foreign | 19 |
| 2.6 | Bivariate analysis Salary vs. Educ | 20 |
| 2.7 | Bivariate analysis | 20 |
| 2.8 | Multivariate analysis of pairplot | 21 |
| 2.9 | Multivariate analysis heatmap | 21 |
| 2.10 | Before Treating Outlier | 22 |
| 2.11 | After Treating Outlier | 22 |
| 2.12 | Parameters for GridsearchCV in Logistic Regression | 23 |
| 2.13 | Best parameter for Logistic Regression | 23 |
| 2.14 | Best estimator for Logistic Regression | 23 |
| 2.15 | Predicted values from the train dataset of Logistic Regression model | 24 |
| 2.16 | confusion matrix from Train data of Logistic Regression | 24 |
| 2.17 | confusion matrix from test data of Logistic Regression | 24 |
| 2.18 | Classification Report from train data of Logistic Regression | 24 |
| 2.19 | Classification Report from test data of Logistic Regression | 24 |
| 2.20 | AUC and ROC curve train data of Logistic Regression | 25 |
| 2.21 | AUC and ROC curve test data of Logistic Regression | 25 |
| 2.22 | Predicted values from the train dataset of LDA | 26 |
| 2.23 | confusion matrix from Train data of LDA Model | 26 |
| 2.24 | confusion matrix from test data of LDA Model | 26 |
| 2.25 | Classification Report from train data of LDA Model | 26 |
| 2.26 | Classification Report from test data of LDA Model | 26 |
| 2.27 | AUC and ROC curve train and test data of LDA model | 27 |
| 2.28 | Accuracy,F1-Score and Confusion Matrix of LDA model at different cut off value | 27 - 29 |

# Problem - 1

*Summary*

The data is gathered from the company Gem Stones co ltd, which deals in distinguish between higher profitable stones and lower profitable stones to make better profitable stones. 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.

*Introduction*

The purpose of this exercise is to explore the dataset and make the price predictions for the diamonds, based on the higher and lower profitable stones.

*Data 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. |

*Sample of the dataset:*

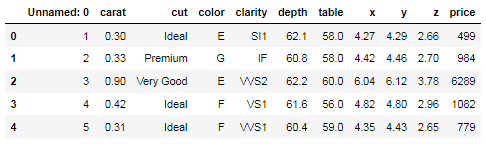


Table 1.1 Dataset Sample

*Exploratory Data Analysis*

*Let us check the types of variables in the data frame.*

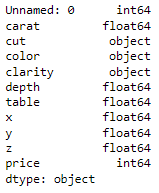


Table- 1.2. Datatypes of the variable

There are total 26967 rows and 11 columns in the dataset. 6 columns are of float64 type , 3 columns are object and 2 columns are int64

*Check for missing values in the dataset:*

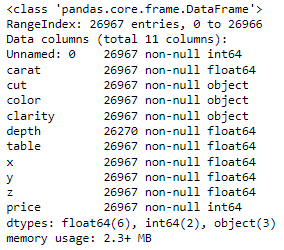


Table- 1.3. Check null values

**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.**

***Uni-Variate Analysis:***

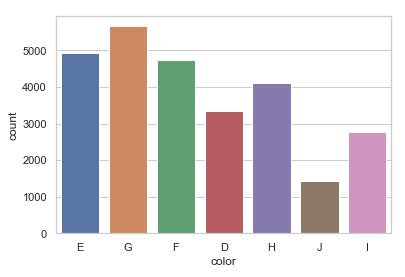


Fig – 1.1 diamond colour count

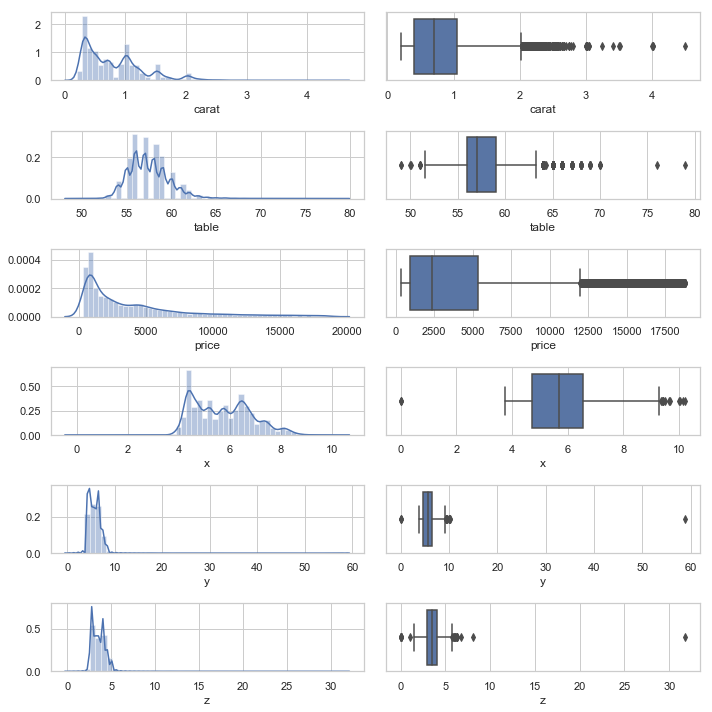


Fig – 1.2 Univariate Analysis

From the above chart (displot and boxplot), there are outliers present in the data.

***Bi – variate Analysis:***

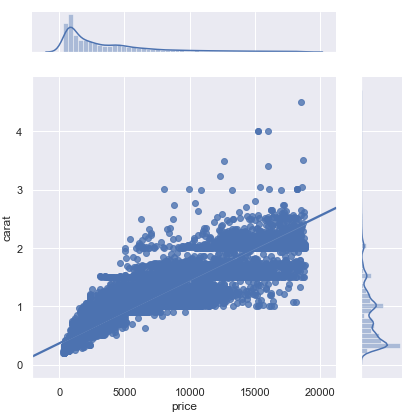


Fig – 1.3 Jointplot for price vs. carat using Bivariate Analysis

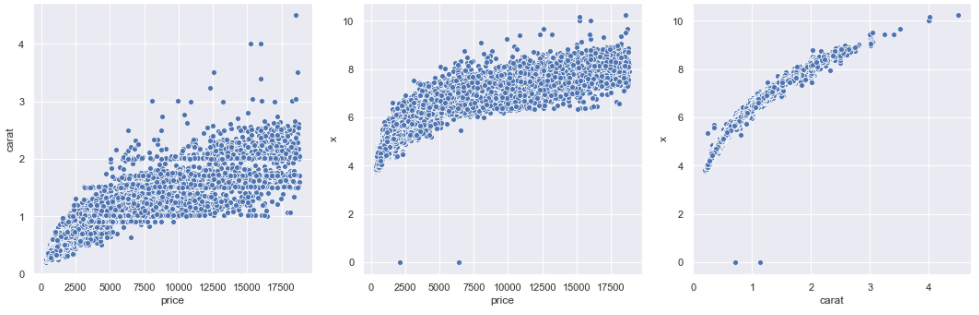


Fig – 1.4 Bivariate Analysis

### From the scatterplot, we can infer that as the 'price' and ‘carat’ increases, the ‘carat’, 'x' increases.

***Multi – variate Analysis:***

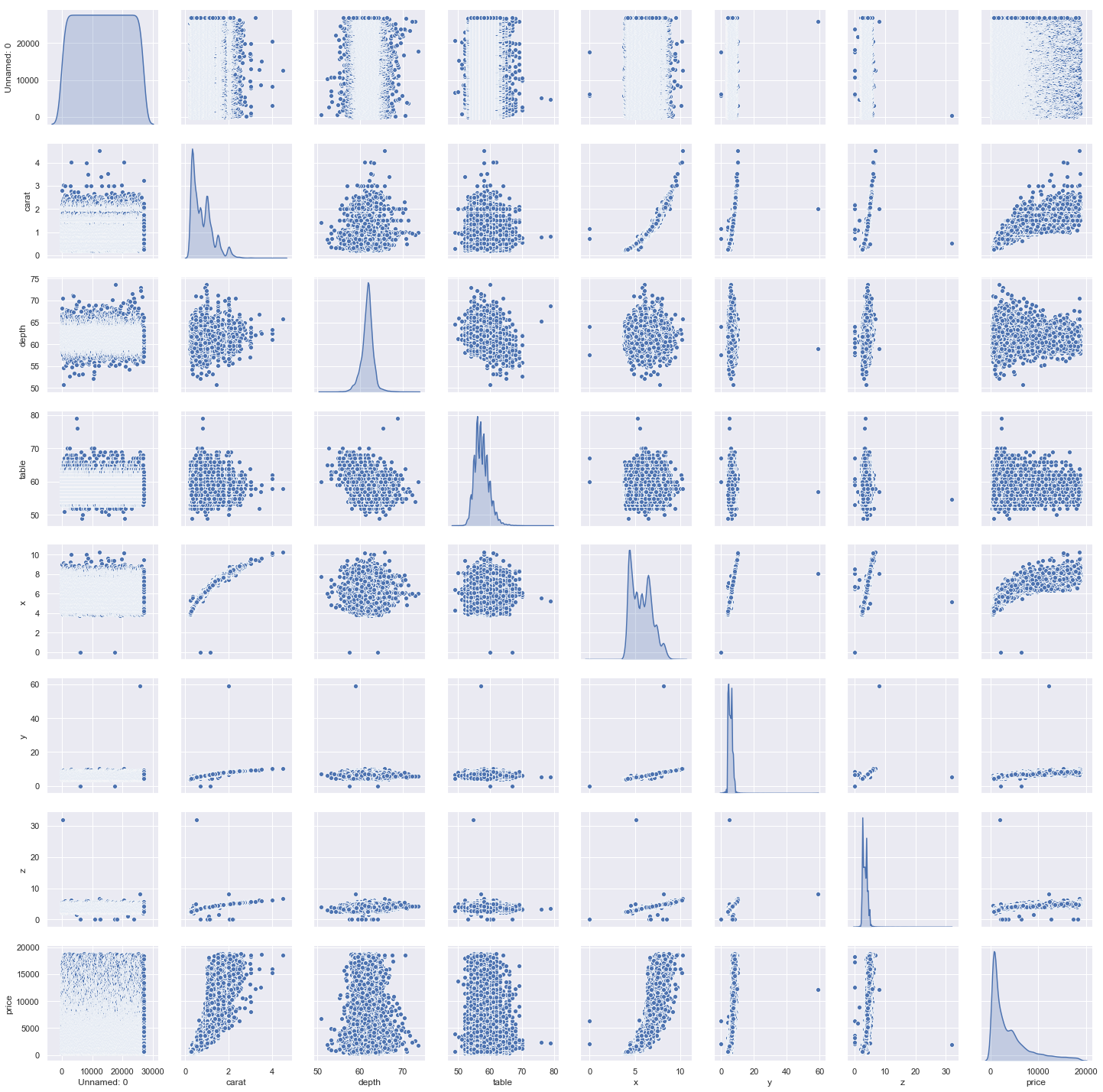


Fig – 1.5 Multivariate analysis of pairplot

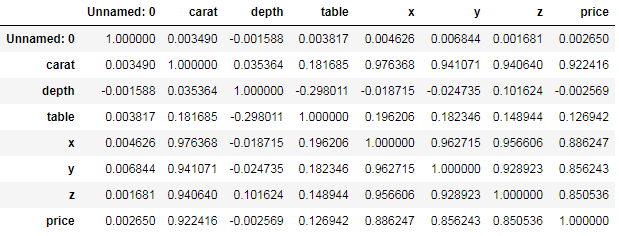


Fig – 1.6 Multivariate analysis for correlation

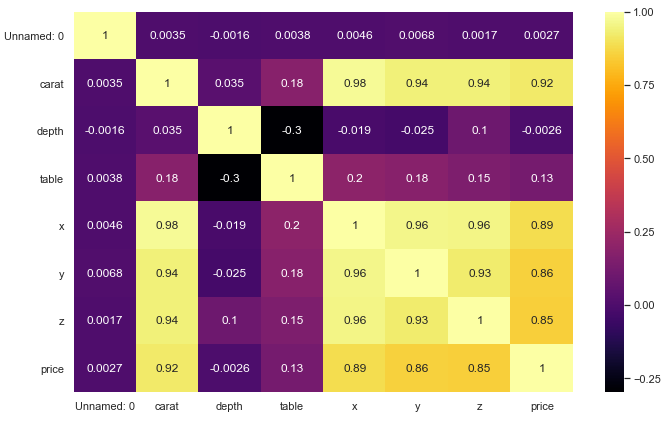


Fig – 1.7 Multivariate analysis of plotting correlation in heatmap

### 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.

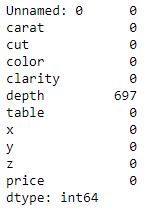


Fig – 1.8 Null values count

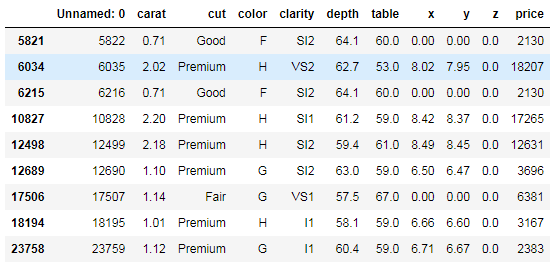


Table – 1.4 Null values in x,y,z variable

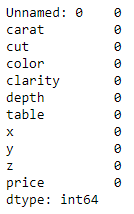


Table – 1.5 After null values treatment

***Before Scaling and treating outliers:***

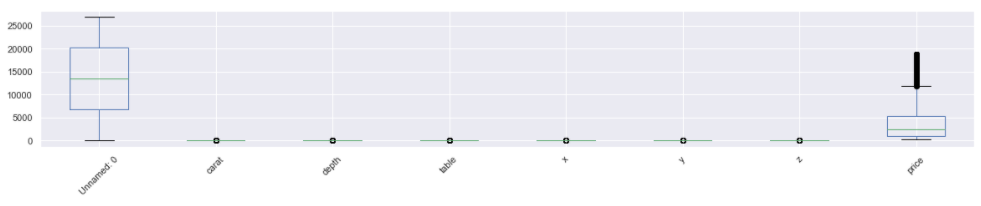


Fig – 1.9 Before Scaling

***After Scaling:***

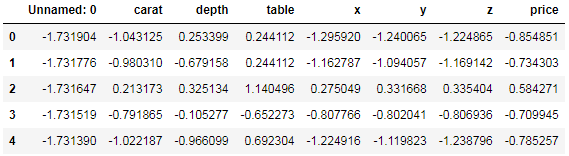


Table – 1.6 After scaling

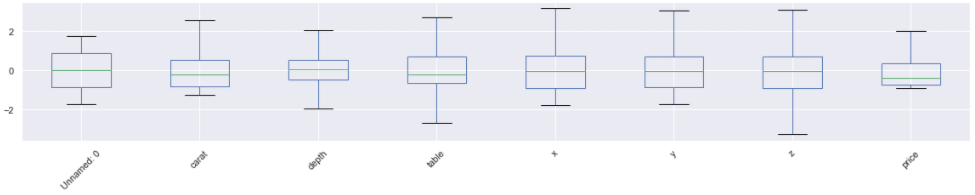


Fig – 1.10 After Scaling

Yes, Scaling needs to be done as the values of the variables are different. price, carat, x, y, z, depth, table are in different values and this may get more weightage. The plot of the data prior and after scaling. Scaling will have all the values in the relative same range. I have used z-score to standardised the data to relative same scale -3 to +3

### 1.3 Encode the data (having string values) for Modelling. Data Split: Split the data into test and train (70:30). Apply Linear regression. Performance Metrics: Check the performance of Predictions on Train and Test sets using Rsquare, RMSE

***Encoding the data:***

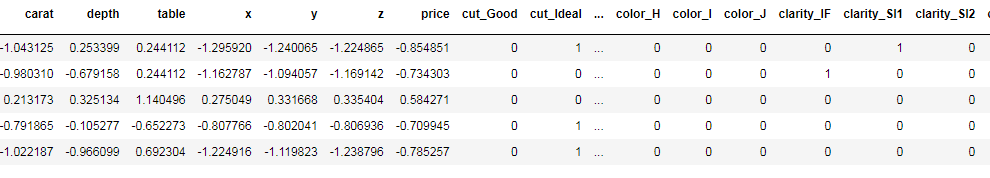


Table – 1.7 Sample Encoded data

From the above dataframe, we can infer that the data are encoded using get dummies encoding method.

***Dataframe after separating the Target variable:***

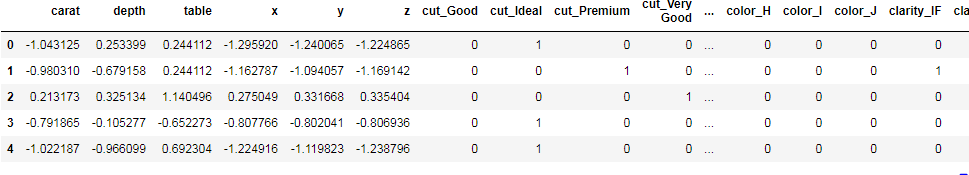


Table – 1.8 sample Dataframe without the TARGET variable



Table – 1.9 sample Dataframe with the TARGET variable

***LINEAR REGRESSION:***

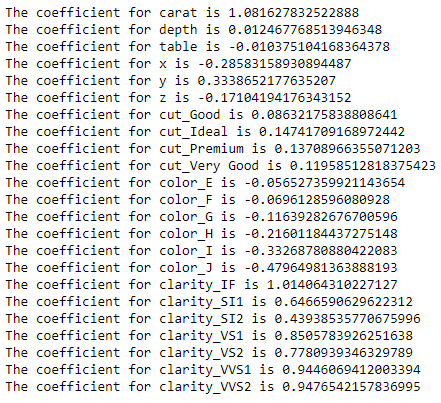


Fig – 1.11 Coefficient of independent variable



Fig – 1.12 Intercept of our model



Fig – 1.13 R – Square Value for Training data



Fig – 1.14 R – Square Value for Testing data



Fig – 1.15 Mean squared error for training data



Fig – 1.16 Mean squared error for Testing data

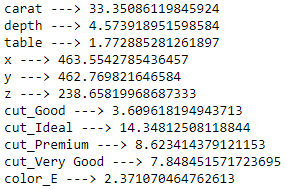


Fig – 1.17 Variance Inflation Factor of our model

***RIDGE REGRESSION:***

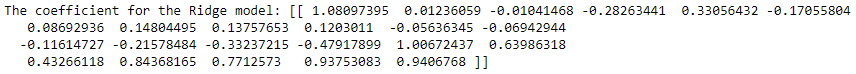


Fig – 1.18 Coefficient of independent variable



Fig – 1.19 R – Square Value for Training and Testing data



Fig – 1.20 Intercept of our model



Fig – 1.21 Mean squared error for Training data



Fig – 1.22 Mean squared error for Testing data

***ORDINARY LEAST SQUARE METHOD :***

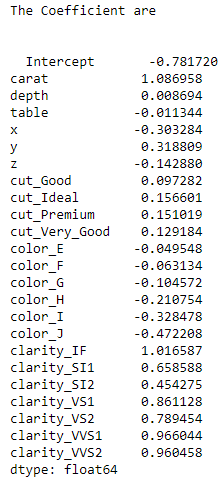


Fig – 1.23 Coefficient of independent variable

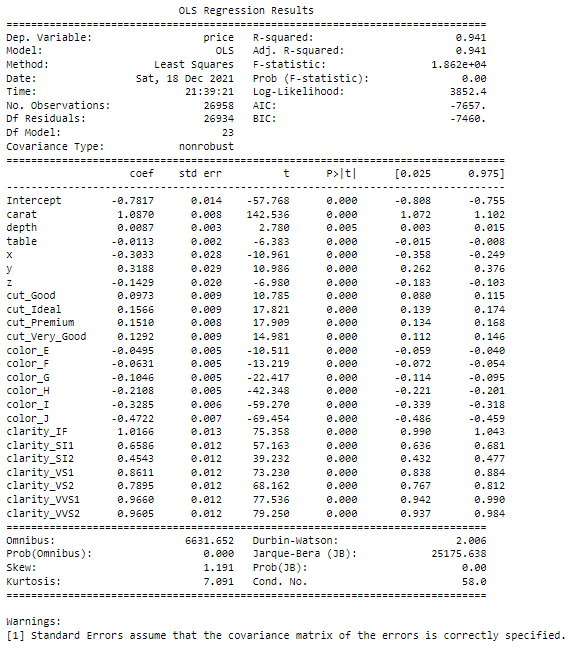


Fig – 1.24 OLS Summary



Fig – 1.25 OLS Summary Mean squared error

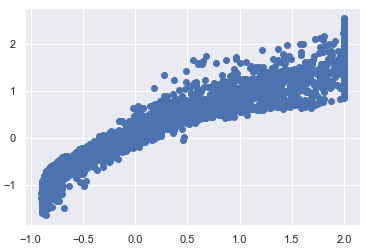


Fig – 1.26 predicted Output vs. testing data

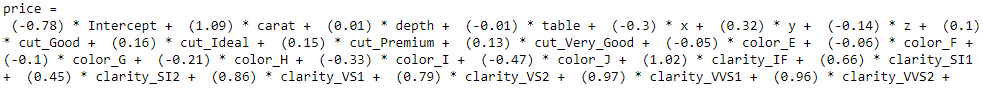


Fig – 1.27 predicted Output vs. testing data Linear equation

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

Based on the R2(R squared) value, The company making out the better profitable share by distinguishing the higher profitable stones and lower profitable stones so that the company can increase the price for higher and lower profitable stones to make higher profits.

The best 5 attributes that are most important from the coefficients are carat, clarity\_IF, clarity\_VVS2, clarity\_VVS1, clarity\_VS1. These are the best attributes that will increase the price of the diamond costliest.

***Business Insights:***

***Train set:***

rsquare: 0.94

adjusted square: 0.94

R - Square is 0.94 which tells the correlation between price of diamonds vs. different independent variable’s explained by 94%

If we see the final model:

price = (-0.78) \* Intercept + (1.09) \* carat + (0.01) \* depth + (-0.01) \* table + (-0.3) \* x + (0.32) \* y + (-0.14) \* z + (0.1) \* cut\_Good + (0.16) \* cut\_Ideal + (0.15) \* cut\_Premium + (0.13) \* cut\_Very\_Good + (-0.05) \* color\_E + (-0.06) \* color\_F + (-0.1) \* color\_G + (-0.21) \* color\_H + (-0.33) \* color\_I + (-0.47) \* color\_J + (1.02) \* clarity\_IF + (0.66) \* clarity\_SI1 + (0.45) \* clarity\_SI2 + (0.86) \* clarity\_VS1 + (0.79) \* clarity\_VS2 + (0.97) \* clarity\_VVS1 + (0.96) \* clarity\_VVS2

Co-efficient of the Carat is highest most, which signifies if there is increase of one unit of carat there will increase of 1.09 in price.

Next most positive effecting independent variable is IF clarity type variable.

Most Negatively effecting parameter is J colour type diamonds ,means a loss of -0.47 will occur with decrease the price of one unit of J colour type diamonds.

***Test set:***

rsquare: 0.941

Adjusted square: 0.941

Finally, Our linear model is good as the r-square difference in train & test dataset is less than 5%.

***Recommendations :***

To Increase the price of the diamond, carat, clarity of the diamond needs to be increased so that the price of the diamond increase which in turn increases profits. The company can sell the diamonds and make higher profitable price from the stone with lower profitable price.

# Problem – 2

*Summary*

The data is gathered from an tour and travel agency which deals in selling holiday packages to sell their packages to employees. 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.

*Introduction*

The purpose of this exercise is to sell tour and travel packages to their employees by predicting the employees would buy tour and travel package using Logistic regression and Linear discriminant analysis (LDA). This dataset consist of 872 rows and 8 columns,

*Data Description*

1. Holiday\_Package: Opted for Holiday Package yes/no?  
2. Salary: Employee salary  
3. Age: Age in years  
4. Edu: Years of formal education  
5. No\_young\_children: The number of young children (younger than 7 years)  
6. No\_older\_children: Number of older children  
7. Foreign: foreigner Yes/No

*Sample of the dataset:*

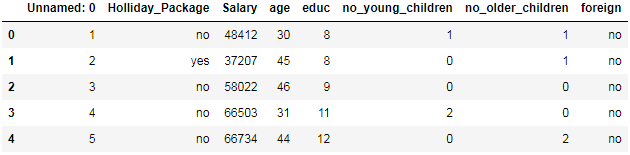


Table 2.1 Dataset Sample

Dataset has 8 variables with Tour and travel package. Based on the travel package, employees will buy the tour and travel packages.

*Exploratory Data Analysis*

*Let us check the types of variables in the data frame.*

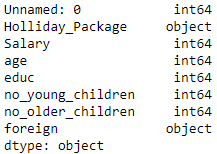


Table 2.2 Datatypes of the variable

There are total 872 rows and 8 columns in the dataset. Out of 8, 2 column is of Object type and rest 6 are of integer data type.

*Check for missing values in the dataset:*

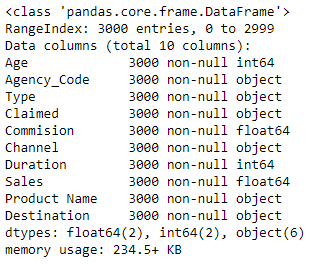


Table 2.3 Check null values

From this, it is clear that there are no null values present in the dataset.

### 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.

There are no null values present in the dataset

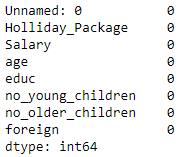


Table 2.4 Null values

The numbers of unique variables are taken from the categorical column.

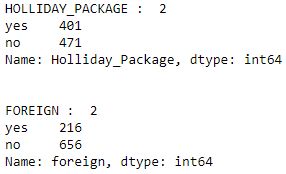


Fig – 2.1 Categorical value count

The numbers of duplicate values are taken from the dataset and duplicate records have been dropped from the dataset.



Fig – 2.2 Number of duplicate rows

#### Univariate Analysis:

Univariate analysis is the simplest form of analysing data. Analyzing each variable in a detailed manner. There are outliers present in all variables except age.

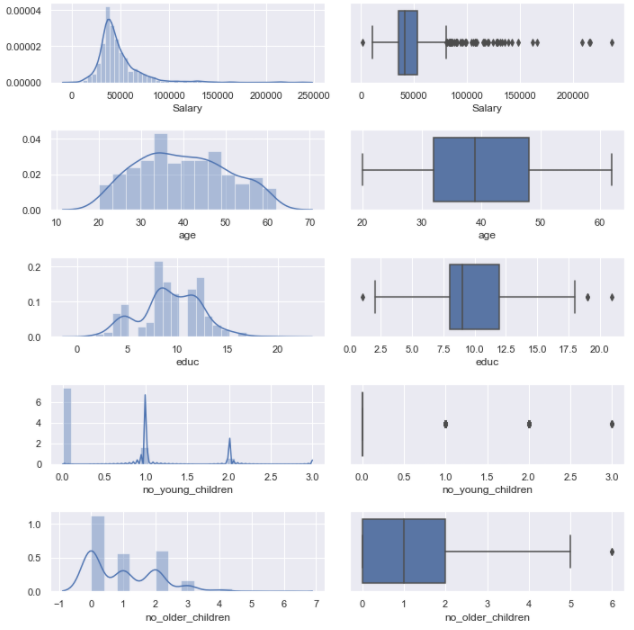


Fig – 2.3 Univariate analysis

The value counts from the Holiday Package variable are, 401 customers have opted for Holiday Package whereas 471 customers have not opted for Holiday Package.

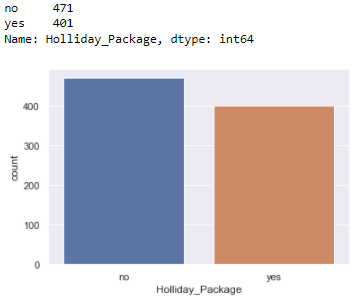


Fig – 2.4 Univariate analysis for Holiday package

The value counts from the Foreign variable are, 216 customers have opted for Foreign whereas 656 customers have not opted for Foreign.

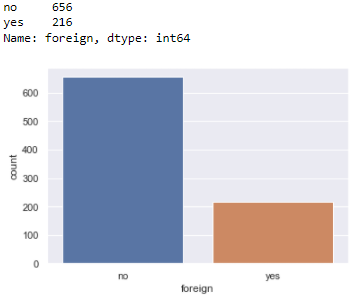


Fig – 2.5 Univariate analysis for Foreign

***Bivariate Analysis:***

Bivariate analysis is the simplest form of analysing data. Analyzing a single variable with another variable in detail.

*Insights:*

The Educ (Years of Education) is plotted against the salary. This Bi-variate plot shows the 20+ years of education earns 60000 whereas, education with 12 years, earns the salary greater than 200000.

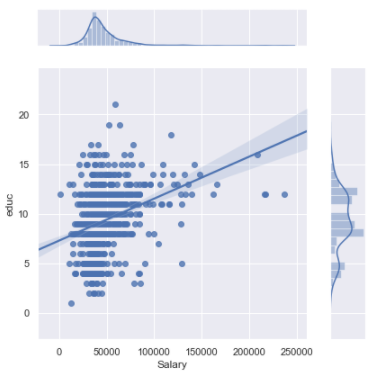


Fig – 2.6 Bivariate analysis Salary vs. Educ

The salary and age in the x-axis is plotted against education and age. The plot 1 denotes, Only few employees with 10-17 years of education earns salary greater than 200000. In 2nd plot, more number of employees gets salary in the range of 50000 - 100000.



Fig – 2.7 Bivariate analysis

**Multivariate Analysis:**

Analysing the data with two or more variable.

*Insights:*

There is no strong correlation observed between few fields.

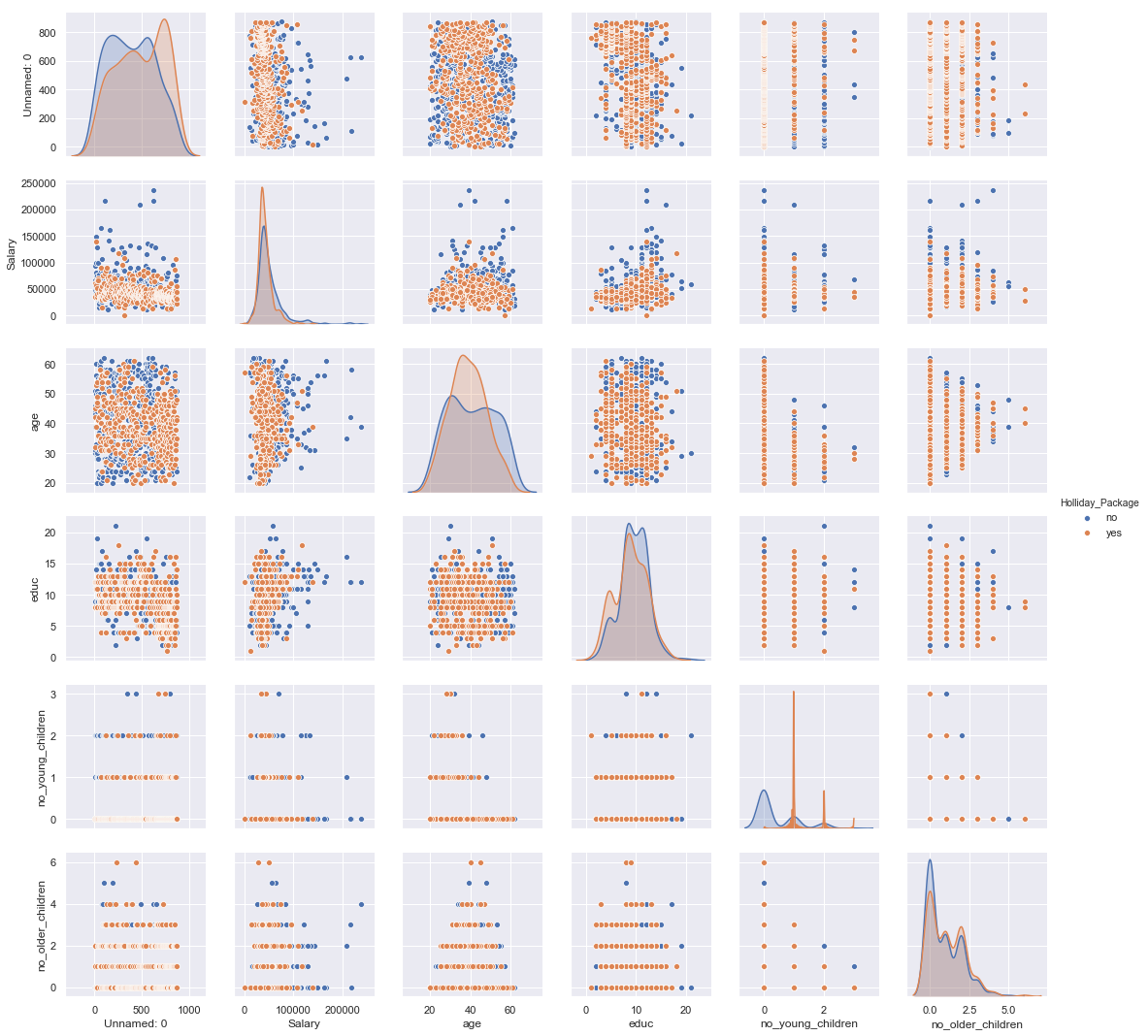
****

Fig – 2.8 Multivariate analysis of pairplot

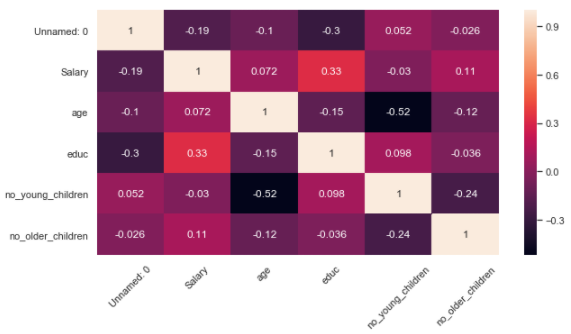


Fig – 2.9 Multivariate analysis heatmap

***Before Treating Outlier :***

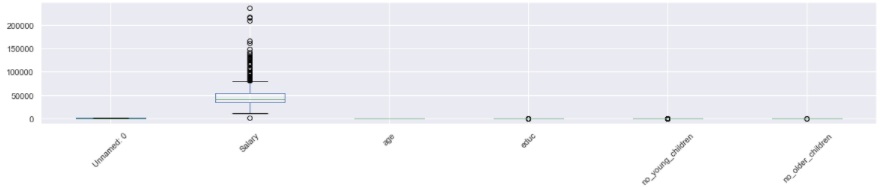


Fig – 2.10 Before Treating Outlier

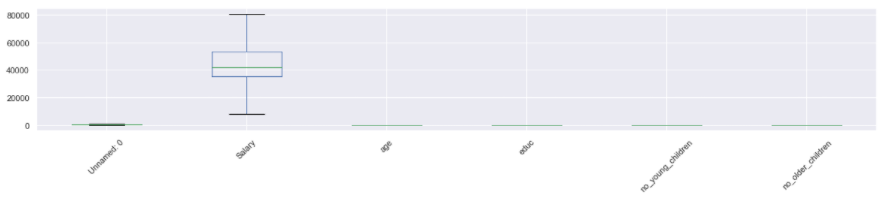


Fig – 2.11 After Treating Outlier

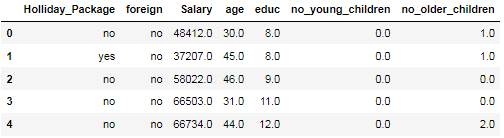


Table – 2.5 Sample Dataset after dropping Unnamed :0 column.

**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 get dummies encoding method is performed for Holiday\_package and Foreign column.

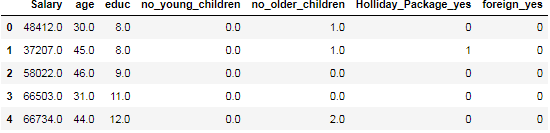


Table 2.6 Sample dataframe after Encoding

The Sample dataframe after removing the Target variable from the original dataframe. The dataframe are split into train and test data from the dataframe. The train data has 70% of the data and test data has 30% of the data from the dataframe.

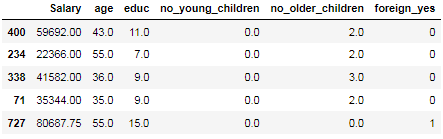


Table 2.7 Train dataframe

### 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:***

The Logistic Regression with the parameters using the grid search CV. The model is fitted into the Logistic Regression using the grid search CV.

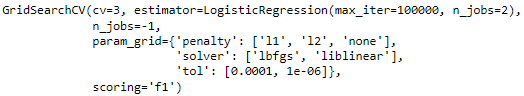


Fig – 2.12 Parameters for GridsearchCV in Logistic Regression

The best parameters are identified from the decision tree algorithm by using the grid search CV.



Fig – 2.13 Best parameter for Logistic Regression



Fig – 2.14 Best estimator for Logistic Regression

The values are predicted from the train data.

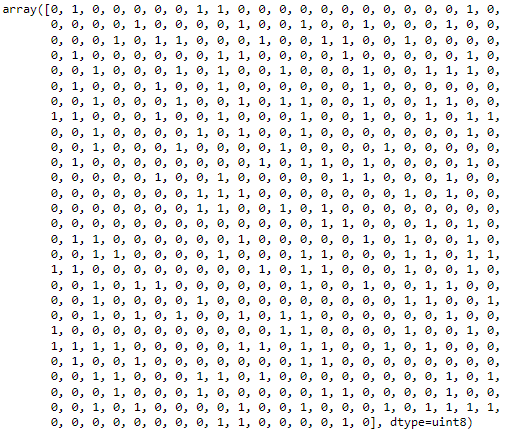


Fig – 2.15 Predicted values from the train dataset of Logistic Regression model

Confusion Matrix is obtained from the train data and test data using Logistic Regression.



Fig 2.16 confusion matrix from Train data of Logistic Regression



Fig 2.17 confusion matrix from test data of Logistic Regression

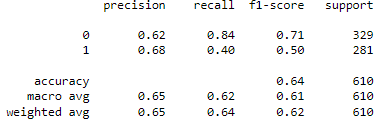


Fig 2.18 Classification Report from train data of Logistic Regression

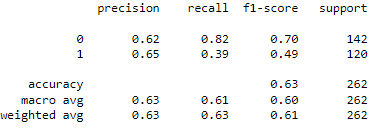


Fig 2.19 Classification Report from test data of Logistic Regression

**ROC curve** (**receiver operating characteristic curve**) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters.

* True Positive Rate
* False Positive Rate

The probability of the Area under the ROC curve for the train data is 66.1%

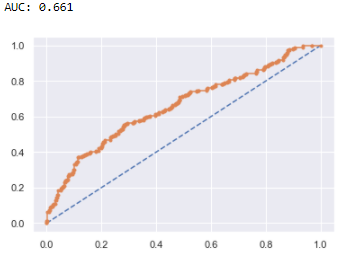


Fig 2.20 AUC and ROC curve train data of Logistic Regression

The probability of the Area under the ROC curve for the train data is 67.3%

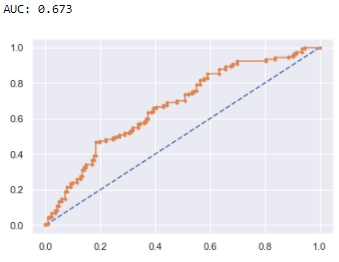


Fig 2.21 AUC and ROC curve test data of Logistic Regression

***Linear Discriminant Analysis:***

The values are predicted from the train data.

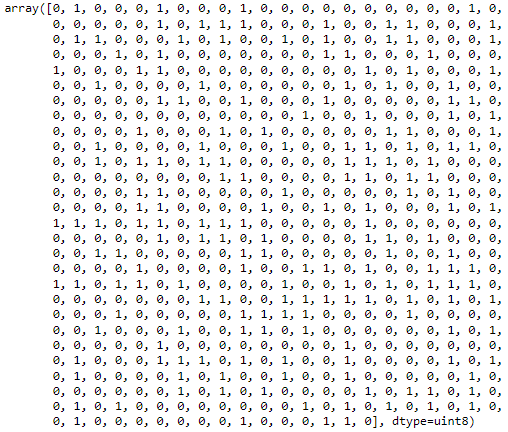


Fig – 2.22 Predicted values from the train dataset of LDA

Confusion Matrix is obtained from the train data and test data using Random Forest Algorithm.



Fig 2.23 confusion matrix from Train data of LDA Model



Fig 2.24 confusion matrix from test data of LDA Model

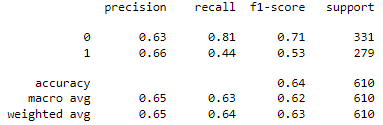


Fig 2.25 Classification Report from train data of LDA Model

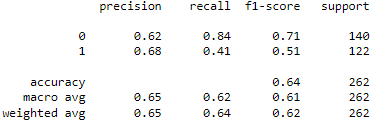


Fig 2.26 Classification Report from test data of LDA Model

**ROC curve** (**receiver operating characteristic curve**) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters.

* True Positive Rate
* False Positive Rate

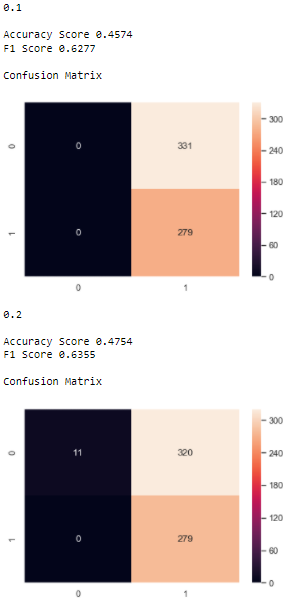
The probability of the Area under the ROC curve for the train data is 66.9%.

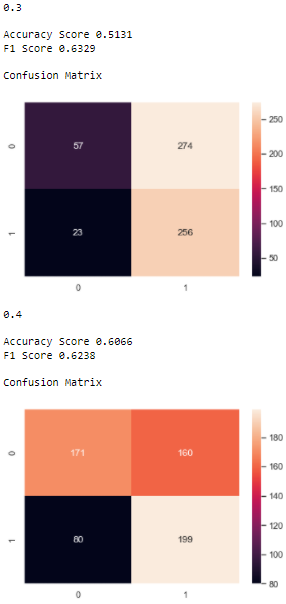
The probability of the Area under the ROC curve for the train data is 65.5%.

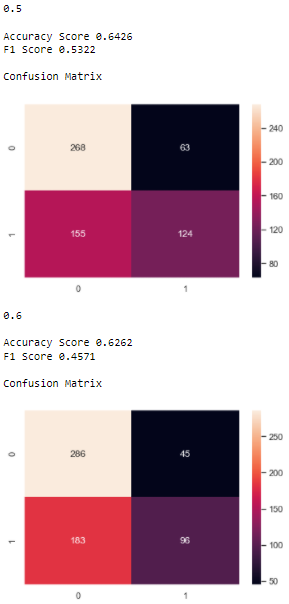


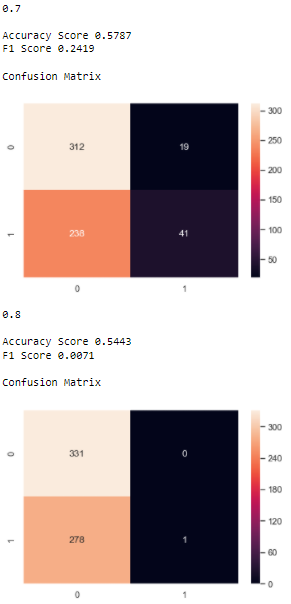
Fig 2.27 AUC and ROC curve train and test data of LDA model

Confusion Matrix, Accuracy and F1 score for different cut off value.









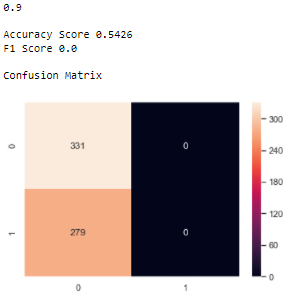


Fig 2.28 Accuracy,F1-Score and Confusion Matrix of LDA model at different cut off value

From this we can infer that, the better result comes at the cut off value of 0.2.

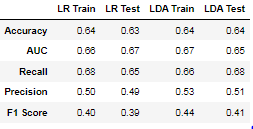


Table 2.8 Comparing Logistic Regression and Linear Discriminant Analysis results

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

***Insights:***

Based on these inference from Logistic Regression(LR) and Linear Discriminant Analysis(LDA), LDA(Linear Discriminant Analysis) gives the better predictions and accurate results for both train and test data.

***Recommendations:***

To increase more holiday packages for the employee

* We can provide complimentary breakfast and dinner for the holiday package.
* Great deals like extra-day stay for the holiday package from the normal trip package.
* Travel package rewards for the employee performance