**Predictive**

**Modeling**

***1. Linear Regression***

***2. Logistic Regression***

***3. LDA***

**NAME: SUDHARSHANA.S**

**BATCH: PGPDSBA Oct\_C’21**

**DATE: 9/4/2022**

**CONTENTS**

**Introduction 1:**

**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.**

* 1. **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 an ordinal variables and take actions accordingly. Explain why you are combining these sub levels with appropriate reasoning.**
  2. **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.**
  3. **Inference: Basis on these predictions, what are the business insights and recommendations.**

Introduction 2:

**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.**

**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).**

**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.**

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

**Introduction 1:**

The company Gem Stones co ltd, which is a cubic zirconia manufacturer. The company is earning different profits on different prize slots. I must 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 to have better profit share.

The data dictionary of the data set is as follows:

|  |  |
| --- | --- |
| **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. |

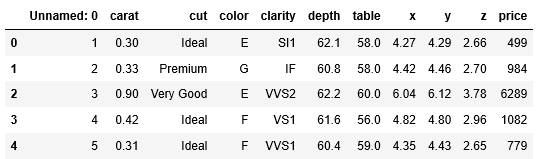
**1.1. Read the data and do exploratory data analysis. Describe the data briefly. (Check the null values, Data types, shape, EDA, duplicate values). Perform Univariate and Bivariate Analysis.**

**Step1:** import the necessary libraries**.**

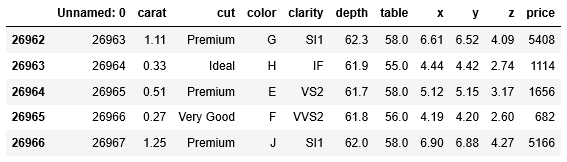
**Step2:** Read the data using the read\_csv function in pandas library.

**Step3:** check for the top and the bottom 5 rows of the dataset using the head and tail function respectively.

**Sample dataset**



**Table 1: Top 5 rows of the dataset**



**Table 2: Bottom 5 rows of the dataset**

* The data looks good on observing the above rows.
* There are 11 variables and 26967 entries in the dataset.
* Price have more weightage when compared to other variables.
* Among the 11 variables unnamed :0 can be dropped as it gives no information.

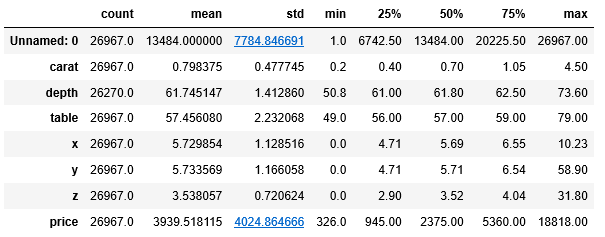
**Step 4:** check the data types and the missing values using the info function.

|  |  |
| --- | --- |
| |  | | --- | |  | |

**Table 3: Data types and missing values**

* There are 26967 rows and 11 columns in the given dataset.
* There are 3 object type variable,2 int type variable and 6 float type variables.
* There are missing values in the depth variable.

**Step 5**: check the descriptive statistics of the dataset using the describe function.



**Table 4: Descriptive statistics**

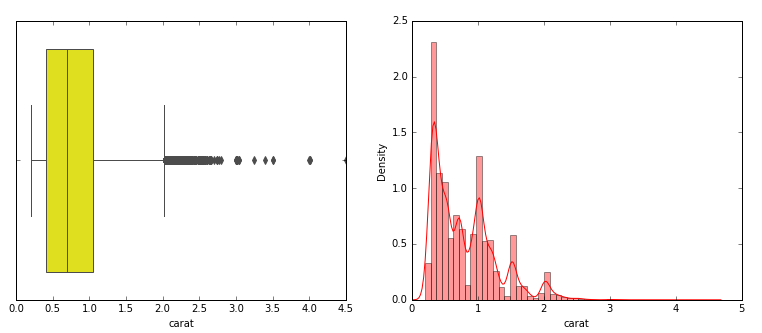
* Based on the descriptive statistics the data looks good.
* We see that for most of the variables mean and the median are nearly equal.
* Standard deviation is high for the spending variable.
* Unnamed: 0 variable can be neglected as it does not give any information.

**Step 6:** On checking for duplicates I found that there are no duplicates in the dataset.

**Step 7**: Drop the unnamed: 0 variable from the dataset. There were no duplicates in the dataset before dropping unnamed: 0 but after dropping unnamed there was 34 duplicates since we are proceeding further in model building without unnamed: 0 we can drop the duplicates.

**UNIVARIATE ANALYSIS:**

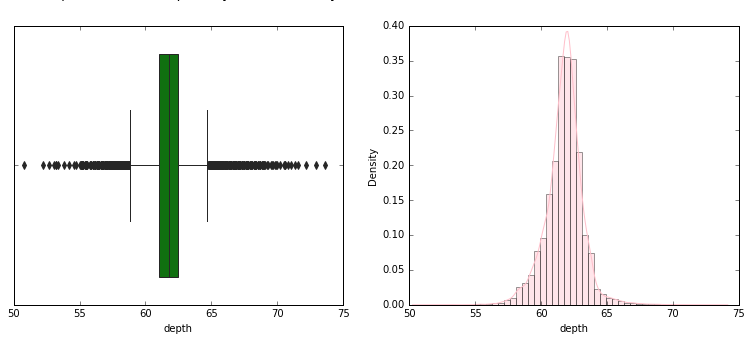
**Carat variable:**



**Fig 1: Box plot and Distplot for carat able**

* Minimum carat: 0.2
* Maximum carat: 4.50
* Median value: 0.70
* Outliers: present
* Skewness: right skewed.

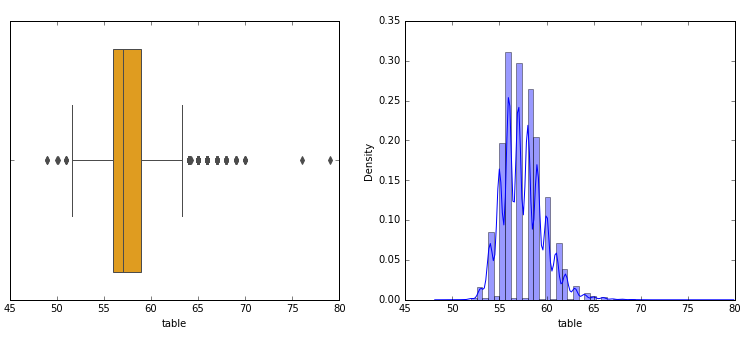
**Depth Variable:**



**Fig 2: Box plot and Distplot of depth variable**

* Minimum depth: 50.8
* Maximum depth: 73.60
* Median: 61.80
* Outliers: present
* Skewness: slightly left skewed.

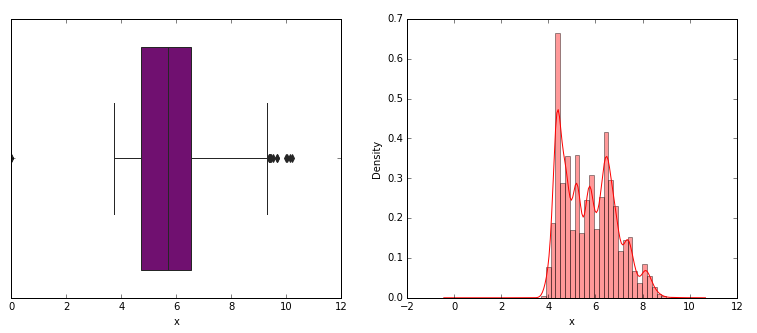
**Table variable:**



**Fig 3: Box plot and distplot for table variable**

* Minimum table: 49.0
* Maximum table: 79.0
* Median: 57.0
* Outliers: present
* Skewness: slightly right skewed.

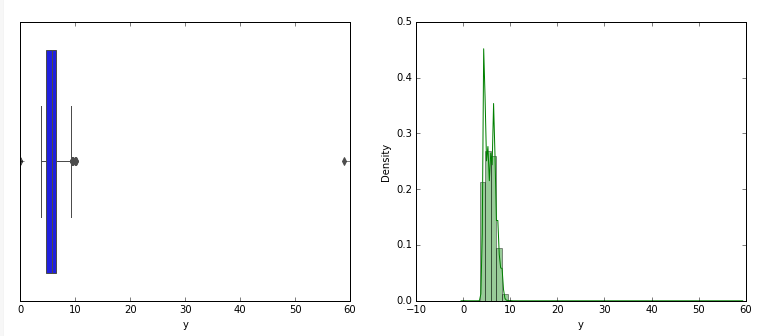
**X variable:**



**Fig 4: Box plot and distplot of X variable**

* Minimum X: 0
* Maximum X: 10.23
* Median value :5.69
* Outliers: present
* Skewness: slightly right skewed

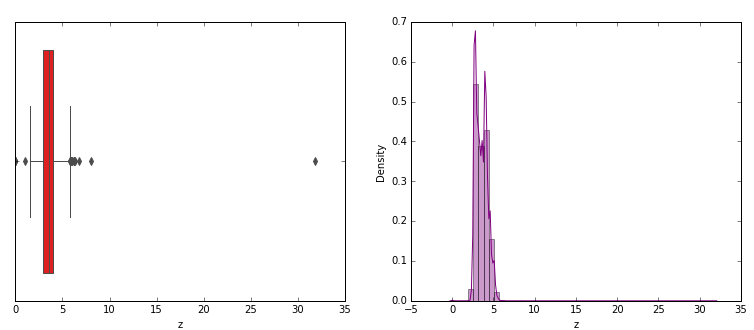
**Y variable:**



**Fig 5: Box plot and distplot of Y variable**

* Minimum Y: 0
* Maximum Y :58.90
* Median: 5.71
* Outliers: present
* Skewness: right skewed

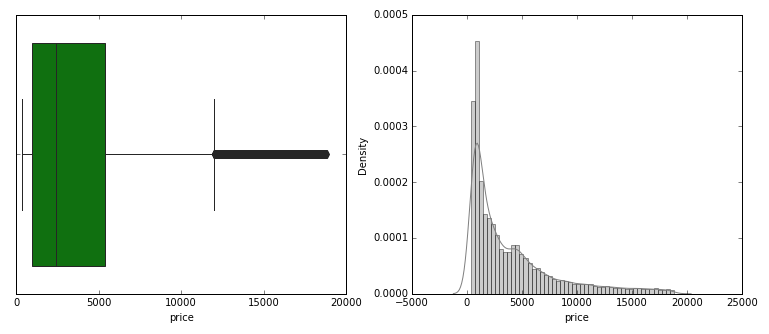
**Z variable:**



**Fig 6: Boxplot and distplot of Z variable**

* Minimum Z: 0
* Maximum Z: 31.80
* Median:3.52
* Outliers: present
* Skewness: right skewed.

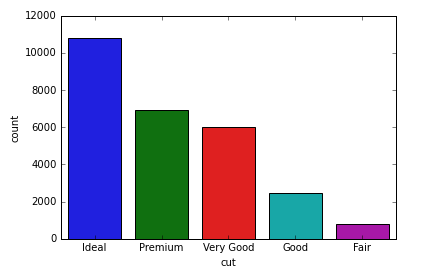
**Price variable:**



**Fig 7: boxplot and distplot of price variable**

* Minimum price: 326.0
* Maximum price: 18818.0
* Median: 2375
* Outliers: present
* Skewness: right skewed.

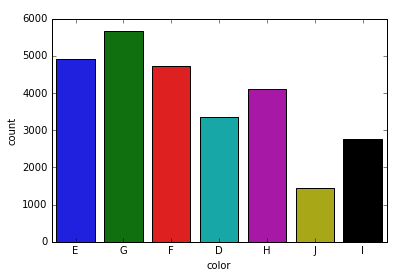
**Cut variable:**



**Fig 8: count plot of cut variable**

* There are 5 categories of cut and their corresponding count is given below.
* **Ideal 10805**
* **Premium 6880**
* **Very Good 6027**
* **Good 2434**
* **Fair 779**

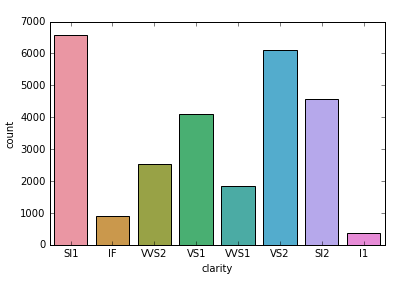
**Color variable:**



**Fig 9: count plot of color variable**

* There are 7 sublevels under ordinal level color their corresponding count is given below.
* **G 5650**
* **E 4916**
* **F 4722**
* **H 4091**
* **D 3341**
* **I 2765**
* **J 1440**

**Clarity variable:**



**Fig 10: count plot of clarity variable**

* There are 8 sublevels under ordinal level clarity their corresponding count is given below.
* **SI1 6564**
* **VS2 6092**
* **SI2 4561**
* **VS1 4086**
* **VVS2 2530**
* **VVS1 1839**
* **IF 891**
* **I1 362**

**Step 8:** Checking for the skewness of the variables.

|  |
| --- |
|  |

**Table 5: skewness of the variables**

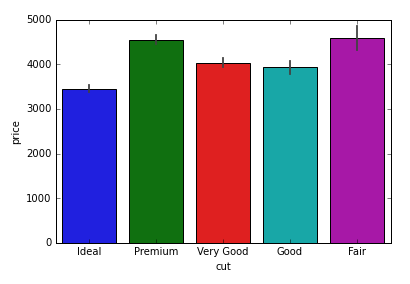
In statistics skewness is the measure of the asymmetry of the probability distribution of the random variables about its mean.skewness can be either positive or negative .if the skewness is zero the datas are perfectly symmetric.

* The skewness below -1 and above 1 are said to be highly skewed.
* The skewness between -1 and -0.5 or 0.5 and 1 are moderately skewed.
* The skewness in between -0.5 and 0.5 is approximately symmetric.
* **From the above table and the above criteria we can say that most of the variables are highly skewed. The least skewed is depth which seems to be almost symmetric.**

**BIVARIATE ANALYSIS**

**Lets plot the response of the categorical variable with respect to the target variable.**

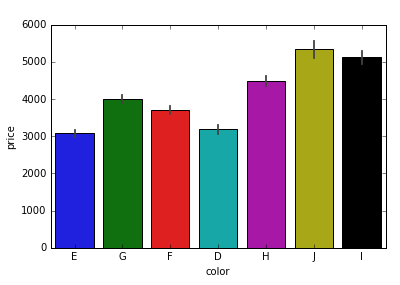
**Cut vs price :**



**Fig 11: bargraph of cut vs price**

* Ideal is having the least price
* Premium and fair is having the highest price.
* Good and the very good cut shares almost equal price.

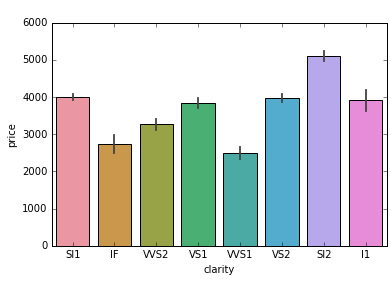
**Color vs price:**



**Fig 12: bargraph of color vs price**

* D and E is having the least price
* J and I is having the highest price.

**Clarity vs price:**

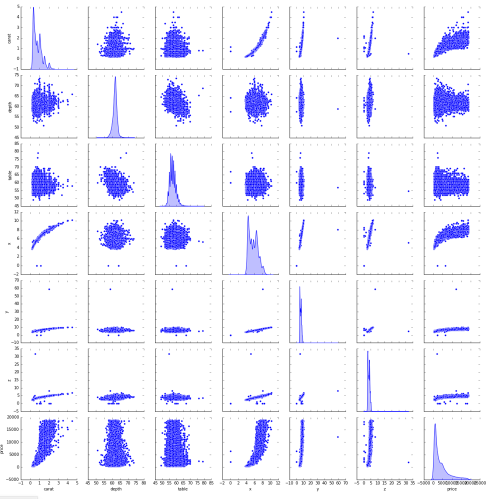


**Fig 13: bargraph of clarity vs price**

* VVS1 is having the least price
* VS1 AND VS2 shares the equal price
* SI2 is having the highest price.

**The response of the continuous variable with the target variable can be seen in the pair plot.**

**MULTIVARIATE ANALYSIS**



**Fig 14: pair plot of the variables**



**Fig 15: Heat map of the variables**

**Strong positive correlation between**

* Carat and price
* Carat and X
* X and y
* X, Y and Z
* From the heat map we can see that most of the variables are highly correlated to each other which gives rise to multicollinearity issue.
* We can see that depth and table is having negative correlation.

**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 an ordinal variables and take actions accordingly. Explain why you are combining these sub levels with appropriate reasoning.**

**Imputing null values:**

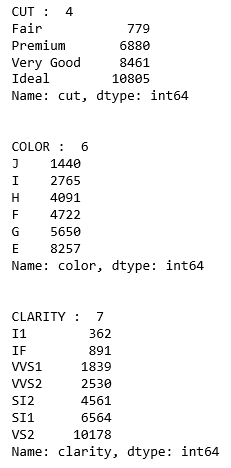
* From the info we can see that there are null values in the depth variable of the dataset.
* We have 697 null values in the depth variable, this can’t be dropped because the data might be lost hence, we are imputing it.
* I impute the missing values with the median of the depth variable, I did so because depth variable is having more outliers.
* Usually mean is affected by the outliers present so median might be the good choice for imputing the null values of the continuous variable.

**Values equal to zero:**

* From the descriptive statistics of the dataset, we can see that the minimum values of the variables X, Y, Z is equal to zero.
* This cannot occur because the length, width and height will not have zero as minimum hence this is wrong.
* There are 8 rows in the dataset with either X, Y, Z with value equal to zero.
* Hence, we can drop these rows as there are only few rows.

**Combining the sublevels of ordinal levels:**

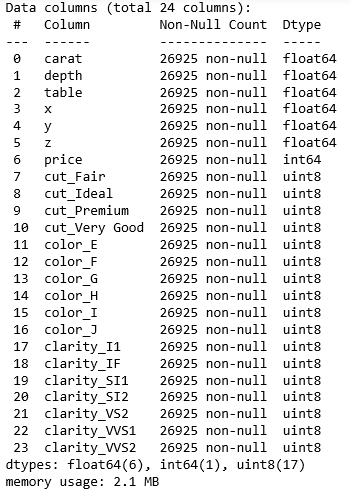
* I have combined the sublevels good and very good in the ordinal level cut
* I have combined the sublevel D and E in the ordinal level color.
* I have combined the sublevels VS1 and VS2 in the ordinal level clarity.
* I have made all the combination of the sublevels in each ordinal level by having price as my combining factor I did so because price is the target variable here.
* Here, good, and very good is having almost the same price.
* D and E is having almost the same price
* VS1 and VS2 is having almost the same price.
* The cut, color, clarity variables and their response with the price can be seen in fig: 11, 12, 13.
* Below given are the levels after combining the sub levels.



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.

**Step 1:** Encode the data having string values for modelling.

**Step 2:** Get the info of the encoded data.



* From the above we can see that all the object variables are encoded.
* There are total of 24 variables in the dataset now.

**Step 3:** copy the predictor or independent variable into x data frame

And copy the target or dependent variable into y data frame.

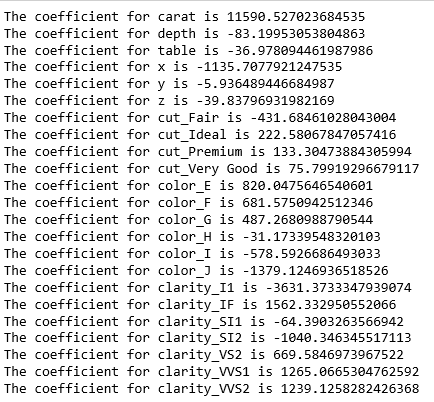
**Step 4**: split x and y into training and the test set in the ratio of 70:30 using train\_test split from sk\_learn model selection.

**Step 5:**

* Fit the data in the linear regression model from sk\_learn linear model.

**Step 6**: Get the coefficients of the variables in the model.

* The variable coefficient tells how the target variable varies with respect to the independent variable.
* The target variable changes with respect to the range of the coefficient of the variable, for example from the below coefficient value we can say that for unit change in carat the price increases by 11590.5270 times.



**Step 7:** Get the intercept of the model.

* The intercept (sometimes called the “constant”) in a regression model represents the mean value of the response variable when all the predictor variables in the model are equal to zero.
* **The intercept for our model is 7895.921294608928.**

**Step 8: get the R squared score.**

* R-Squared is a statistical measure of fit that indicates how much variation of a dependent variable is explained by the independent variable(s) in a regression model.
* **R-squared on training data is 0.91944**
* **R-squared on testing data is 0.9232**

**Step 9: get the adjusted R squared**

* Adjusted R2 is a corrected goodness-of-fit (model accuracy) measure for linear models. It identifies the percentage of variance in the target field that is explained by the input or inputs. R2 tends to optimistically estimate the fit of the linear regression.
* The Adjusted R-squared considers the number of independent variables used for predicting the target variable. In doing so, we can determine whether adding new variables to the model increases the model fit.
* **Adjusted R-squared on training data is 0.91944**
* **Adjusted R-squared on testing data is 0.9232**

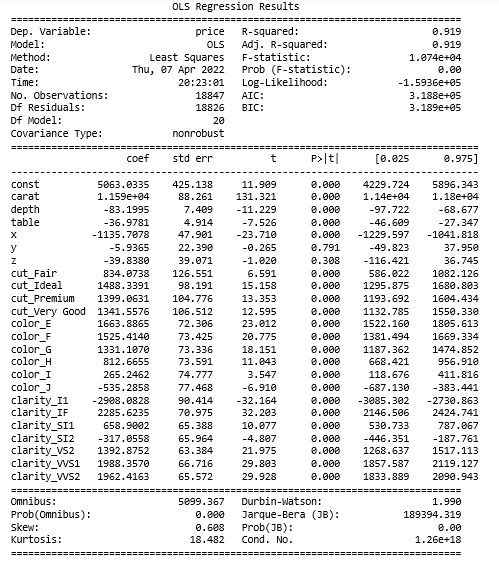
**Step 10: Get the RMSE**

* The most common metric for evaluating linear regression model performance is called root mean squared error, or RMSE. The basic idea is to **measure how bad/erroneous the model's predictions are when compared to actual observed values**. So, a high RMSE is “bad” and a low RMSE is “good”.
* RMSE on training data is 1137.57
* RMSE on testing data is 1121.94
* Here the range of the target variable is 326 to 18818, hence we can normalise RMSE for better understanding because the RMSE value changes based on the target range.
* Normalized RMSE = RMSE / (max value – min value)
* Normalised RMSE on training data is 0.0614
* Normalised RMSE on testing data is 0.0606

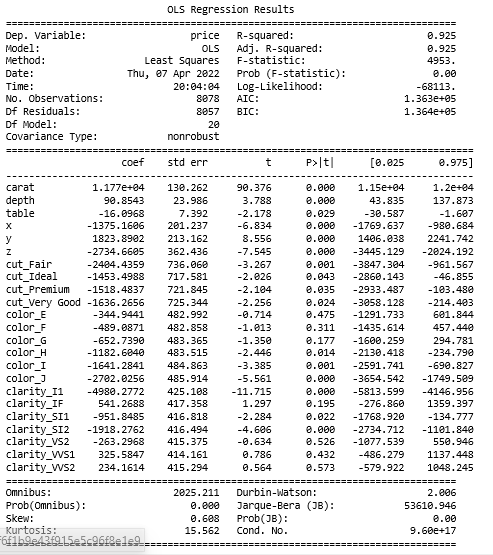
**Step 11: Fit the training data and the test data in the stats\_model OLS model separately.**

* OLS or Ordinary Least Squares is a method used in Linear Regression for estimating the unknown parameters by creating a model which will minimize the sum of the squared errors between the observed data and the predicted one.
* We get the values of the parameters separately for the train and the test set.
* I have added a constant to get the constant value in the model.

**Step 12:** Get the OLS regression models of the training and the test set separately.

**Training data:**

**Testing data:**



* **Standard Errors assume that the covariance matrix of the errors is correctly specified.**
* **The smallest eigenvalue of test data is 6.31e-29. This might indicate that there are strong multicollinearity problems or that the design matrix is singular.**
* **The smallest eigenvalue of train data is 8.53e-29. This might indicate that there are strong multicollinearity problems or that the design matrix is singular.**

**Step 13:** Get the intercept of the model.

* The intercept (sometimes called the “constant”) in a regression model represents the mean value of the response variable when all the predictor variables in the model are equal to zero.
* **The intercept for our model is 5063.0335.**

**Step 14: get the R squared score.**

* R-Squared is a statistical measure of fit that indicates how much variation of a dependent variable is explained by the independent variable(s) in a regression model.
* **R-squared on training data is 0.91944**
* **R-squared on testing data is 0.925**

**Step 15: get the adjusted R squared**

* Adjusted R2 is a corrected goodness-of-fit (model accuracy) measure for linear models. It identifies the percentage of variance in the target field that is explained by the input or inputs. R2 tends to optimistically estimate the fit of the linear regression.
* The Adjusted R-squared considers the number of independent variables used for predicting the target variable. In doing so, we can determine whether adding new variables to the model increases the model fit.
* **Adjusted R-squared on training data is 0.91944**
* **Adjusted R-squared on testing data is 0.925**

**Step 16: Get the RMSE**

* The most common metric for evaluating linear regression model performance is called root mean squared error, or RMSE. The basic idea is to **measure how bad/erroneous the model's predictions are when compared to actual observed values**. So, a high RMSE is “bad” and a low RMSE is “good”.
* RMSE on training data is 1137.57
* RMSE on testing data is 1121.94
* Here the range of the target variable is 326 to 18818, hence we can normalise RMSE for better understanding because the RMSE value changes based on the target range.
* Normalized RMSE = RMSE / (max value – min value)
* Normalised RMSE on training data is 0.0614
* Normalised RMSE on testing data is 0.0606

**Step 17: std err**reflects the level of accuracy of the coefficients. The lower it is, the higher is the level of accuracy

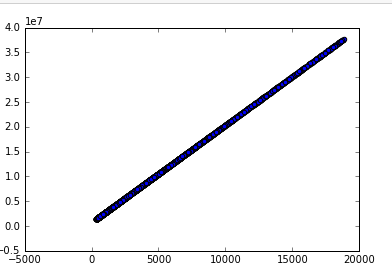
**Step 18: P >|t|** is your *p-value*. A p-value of less than 0.05 is statistically significant.

We test if the true value of the coefficient is equal to zero (no relationship). The statistical test for this is called Hypothesis testing.

* A low P-value (< 0.05) means that the coefficient is likely not to equal zero.
* A high P-value (> 0.05) means that we cannot conclude that the explanatory variable affects the dependent variable.
* A high P-value is also called an insignificant P-value.

**Step 19: Confidence Interval** represents the range in which our coefficients are likely to fall (with a likelihood of 95%)

**Step 20:** plot the regression line for the predicted and the actual price.



**Fig 16: Regression line of the model**

* From the above regression models and their performance we can say that both the models performed well in capturing the degree to which the independent variables are influencing the dependent variable.
* Based on the RMSE and Adjusted R squared value we can say that both the models performed equaly good.
* From the plot of regression line we can say that it is the line that reduces the distance between the actual score from the predicted score, model performance is good.

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

* The price of the cubic zirconia depends mostly on the carat weight of the cubic zirconia more is the carat weight the costlier it is.
* The fair cut cubic zirconia is having the direct relation with price if the cubic zirconia is fair cut, then they are costlier.

1. The colour E cubic zirconia is having the higher price when compared to other colours.
2. Clarity IF, VVS1, VVS2 have direct relation with price. cubic zirconia is costlier for these types of clarity.
3. The depth, table, height, width, length all this dimension is indirectly related to price which mean more the above dimensions value cheap is the cubic zirconia.
4. From the above we can say that a cubic zirconia with high carat weight fairly cut with colour E and clarity IF and with the dimension depth, table, height, width, length kept minimum can have the highest price.
5. If the cubic zirconia manufactured fulfils the above criteria, then the profit can be maximized.

8. The 5 best attributes that are more important are:

* Carat weight
* Clarity (IF)
* Cut (ideal)
* Colour (E)
* Dimension (X)

**Introduction 2:**

We are hired by a tour and travel agency which deals in selling holiday packages. We are provided details of 872 employees of a company. Among these employees, some opted for the package, and some didn't. we must help the company in predicting whether an employee will opt for the package or not based on the information given in the data set.

The data dictionary of the dataset is as follows:

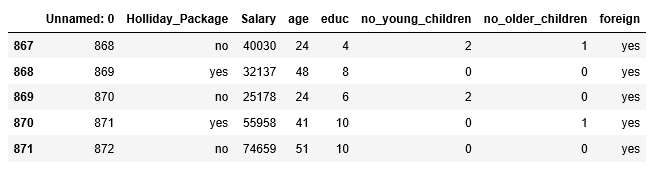
|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| Holiday\_Package | Opted for Holiday Package yes/no? |
| Salary | Employee salary |
| age | Age in years |
| edu | Years of formal education |
| no\_young\_children | The number of young children (younger than 7 years) |
| no\_older\_children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

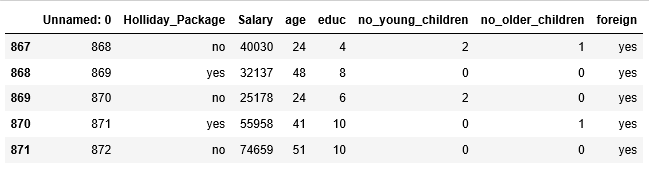
**Step1:** import the necessary libraries**.**

**Step2:** Read the data using the read\_csv function in pandas library.

**Step3:** check for the top and the bottom 5 rows of the dataset using the head and tail function respectively.



**Table 6: Top five rows of the dataset**



**Table 7: Bottom five rows of dataset**

* The data looks good on seeing the head and tail of the data.
* There are 8 variables in the dataset.
* Here Holliday\_package is the dependent variable, and the rest are independent variable.

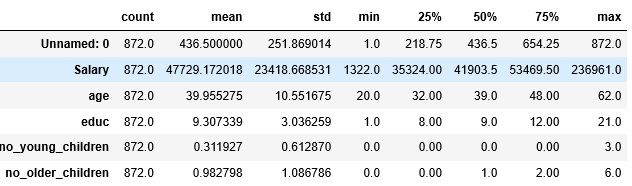
**Step 4:** check the data types and the missing values using the info function.

|  |
| --- |
|  |

**Table 8: Data type and missing values**

* Holliday\_package and foreign are categorical variables
* Rest are numeric variables.
* 872 records, no missing values.
* 7 independent variable and one target variable Holliday\_package.

**Step 5:** To check the descriptive statistics of the dataset.



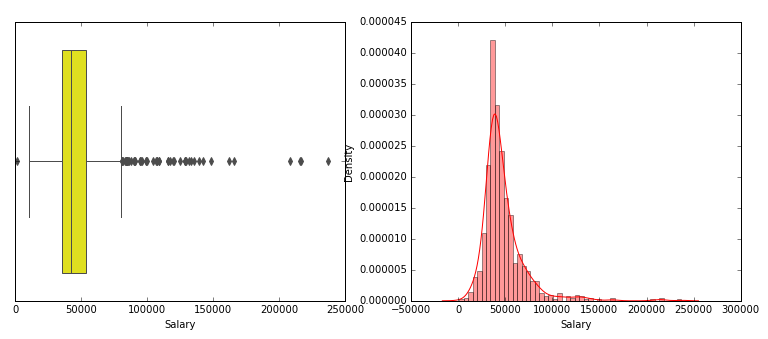
**Table 9: Descriptive statistics of dataset**

* Salary is having more weightage among the other variables.
* Unnamed: 0 can be removed as it doesn’t give any useful information.

**Step 6:** On checking for duplicates I found that there are no duplicates in the dataset.

**UNIVARIATE ANALYSIS**

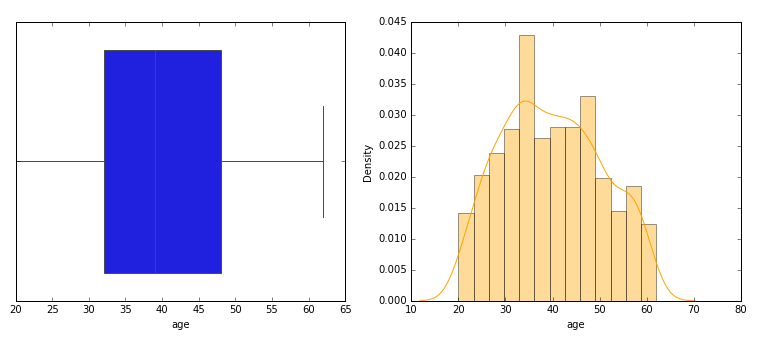
**Salary variable:**



**Fig 17: Box plot and distplot of salary variable**

* Minimum salary: 1322.0
* Maximum salary: 236961.0
* Median :41903.5
* Outliers: present
* Skewness: Right skewed.

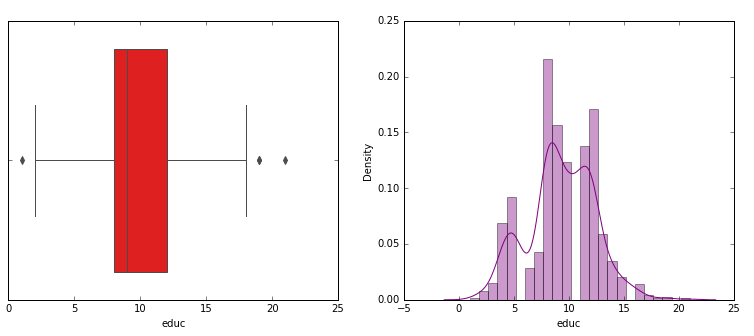
**Age variable:**



**Fig 18: Boxplot and distplot of age variable**

* Minimum age: 20
* Maximum age: 62
* Median:39
* Outliers: absent
* Skewness: almost normal

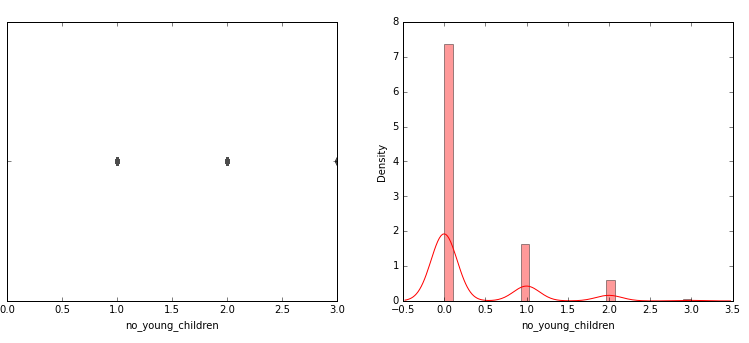
**Educ variable:**



**Fig 19: Boxplot and distplot of educ variable**

* **Minimum educ: 1.0**
* **Maximum educ: 21**
* **Median: 9.0**
* **Outliers: present**
* **Skewness: right skewed**

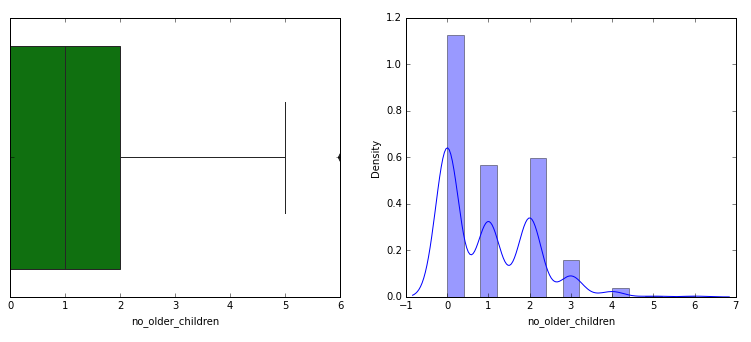
**No\_young\_children variable:**



**Fig 20: Boxplot and distplot of no\_young\_children variable**

* **Minimum no\_young\_children: 0.0**
* **Maximum no\_young\_children: 3**
* **Median: 0**
* **Outliers: present**
* **Skewness: highly right skewed**

**No\_older\_children variable:**



**Fig 21: Boxplot and distplot of no\_older\_children variable**

* **Minimum no\_older\_children: 0.0**
* **Maximum no\_older\_children: 6**
* **Median: 1**
* **Outliers: present**
* **Skewness: right skewed**

**Skewness of the variables:**

**1.Salary 3.103216**

**2.age 0.146412**

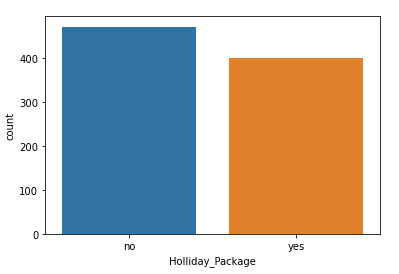
**3.educ -0.045501**

**4.no\_young\_children 1.946515**

**5.no\_older\_children 0.953951**

* From the above values we can say that the variables are highly skewed.
* Educ variable is almost normal.

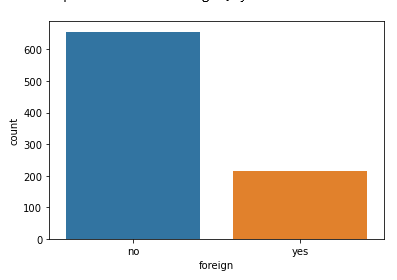
**Holliday\_package variable:**



**Fig 22: count plot of Holliday\_package variable**

* **The people who opted for holiday package is 401**
* **The people who didn’t opt for holiday package is 470**

**Foreign variable:**

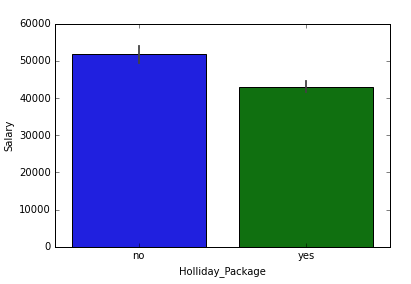


**Fig 23: count plot of foreign variable**

* **There are 216 foreigners and the rest 656 are not foreigners.**

**Bivariate Analysis**

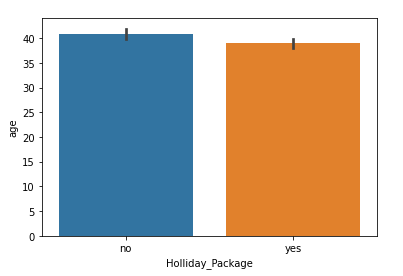
**Salary vs Holliday\_package:**



**Fig 24: barplot of salary vs Holliday\_package**

* Employees with the salary from minimum to median range opt for Holliday package.
* Employees with high salary do not opt for package

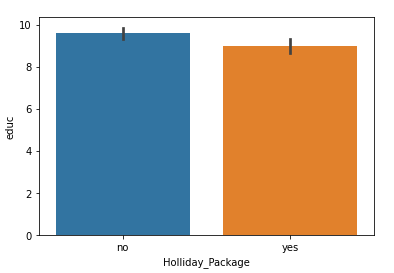
**Age vs Holliday\_package:**



**Fig 25: barplot of age vs holiday package**

* The age group in the range of minimum to median opt for the holiday\_package.
* High age group people that are more than 40 do not opt holiday\_package

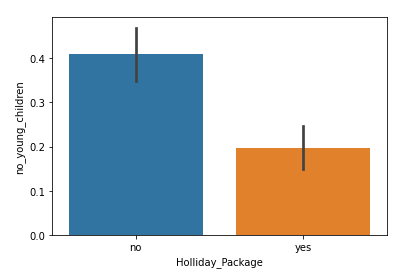
**Educ vs Holliday package:**



**Fig 26: barplot of educ vs holliday\_package**

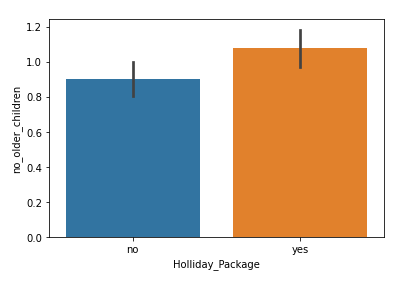
* Employees of every education level shows interest in holiday\_package

**No\_young\_children vs holliday\_package:**



**Fig 27: barplot of no\_young\_children vs holliday\_package**

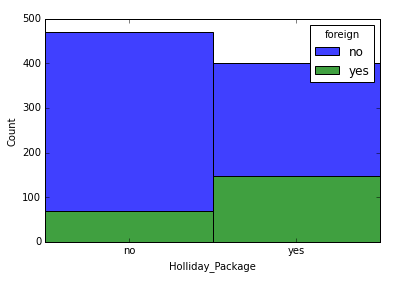
**No\_older\_children vs holliday\_package:**



**Fig 27: barplot of no\_older\_children vs holliday\_package**

* More the age of the children more is their interest on the holiday package.

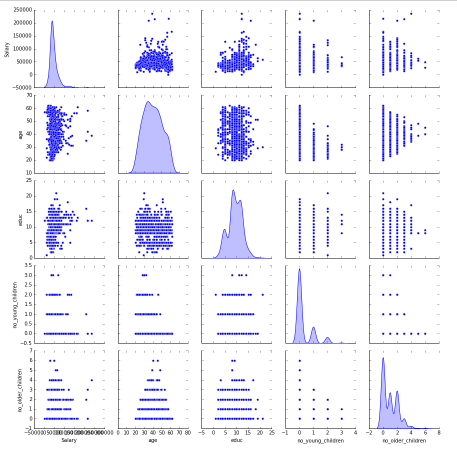
**Foreign vs holliday\_package:**



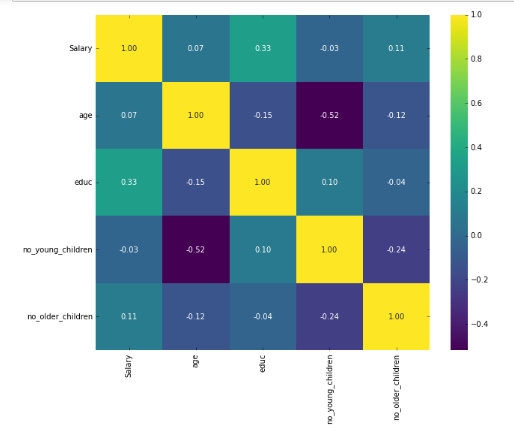
**Fig 28: stack chart of foreign vs holliday package**

* We can see that the employees who are foreigners opt for the holiday package.

**Multivariate Analysis**



**Fig 29: pair plot of continuous variables**

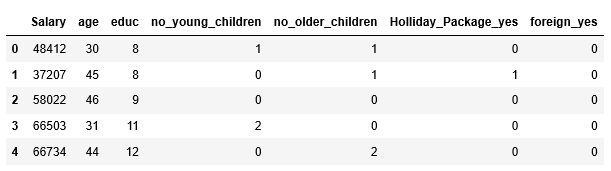


**Fig 30: heat map of the continuous variables**

* None of the variables have strong correlation.

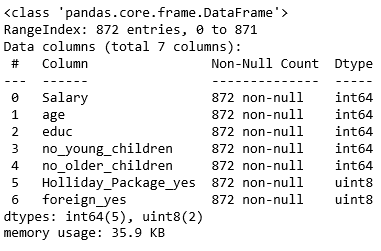
**2.2 Do not scale the data. Encode the data (having string values) for Modelling. Data Split: Split the data into train and test (70:30). Apply logistic Regression and LDA (linear discriminant analysis).**

**Step 1:** Encode the data having string values for modelling.



**Table 10: sample of encoded data**

**Step 2:** Get the info of the encoded data.



* From the above we can see that all the object variables are encoded.

**Step 3:** copy the predictor or independent variable into x data frame

And copy the target or dependent variable into y data frame.

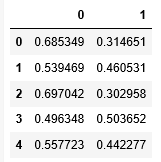
* Here Holliday\_package\_yes is the target variable, I have dropped Holliday\_package\_no after encoding.

**Step 4**: split x and y into training and the test set in the ratio of 70:30 using train\_test split from sk\_learn model selection.

**Step 5:**

* Fit the data in the logistic regression model from sk\_learn linear model.
* Fit the data in the linear discriminant analysis in the sk\_learn discriminant analysis.

**Step 6**: Get the predicted probability of the test data.



**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.**

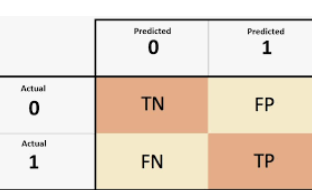
**Accuracy:** Accuracy is defined as the percentage of correct predictions for the test data. It can be calculated easily by dividing the number of correct predictions by the number of total predictions.

**Precision:** The ability of a classification model to identify only the relevant data points. Mathematically, precision the number of true positives divided by the number of true positives plus the number of false positives.

**Recall:** The ability of a model to find all the relevant cases within a data set. Mathematically, we define recall as the number of true positives divided by the number of true positives plus the number of false negatives.

**F score:** In statistical analysis of binary classification, the F-score or F-measure is a measure of a test's accuracy. It is calculated from the precision and recall. It uses harmonic mean for valuation.

**Confusion matrix:** A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.



This is how the confusion matrix is printed in python ,TN(true negative) first and TP(true ositive) finally.

**ROC Curve:** An 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:

1. True Positive Rate

2. False Positive Rate

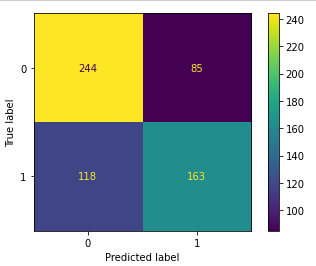
**ROC\_ AUC score:** The area under the ROC curve (AUC) results were considered excellent for AUC values between 0.9-1, good for AUC values between **0.8-0.9**, fair for AUC values between 0.7-0.8, poor for AUC values between 0.6-0.7 and failed for AUC values between 0.5-0.6.

**Classification report:** A Classification report is used to measure the quality of predictions from a classification algorithm. The report shows the main classification metrics precision, recall and f1-score on a per-class basis. The metrics are calculated by using true and false positives, true and false negatives

**Performance metrics of logistic regression model:**

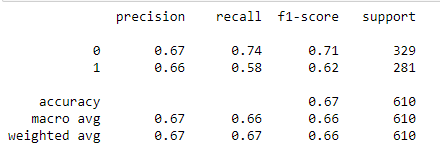
**Train data analysis:**

**Confusion matrix:**



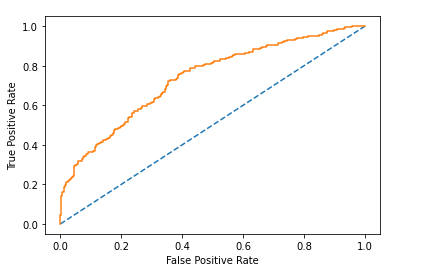
* True negative:244
* False negative:118
* False positive:85
* True positive:163.

**Classification Report:**



* **Precision: 0.66**
* **Recall: 0.58**
* **F1-score: 0.62**
* **Accuracy: 0.67**

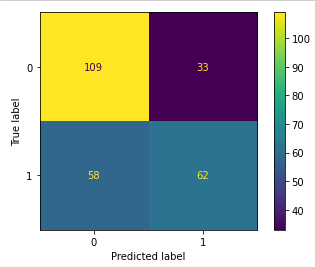
**ROC curve:**



* The curve is appearing to be near the y axis.
* The area under the the curve is relatively moderate which shows that the model is relatively good.
* **ROC\_ AUC score: 0.735**

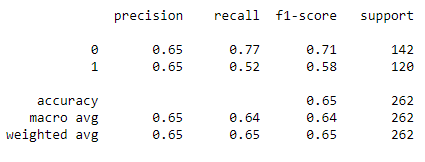
**Test data analysis:**

**Confusion matrix:**



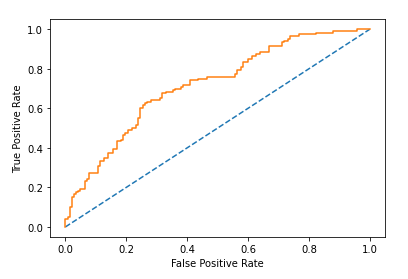
* True negative:109
* False negative:58
* False positive:33
* True positive:62

**Classification Report:**



* **Precision: 0.65**
* **Recall: 0.52**
* **F1-score: 0.58**
* **Accuracy: 0.65**

**ROC curve:**

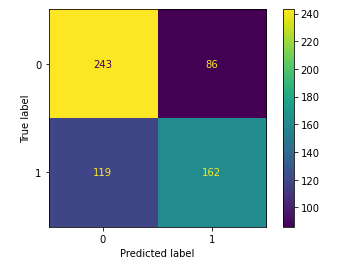


* The curve is appearing to be near the y axis.
* The area under the the curve is moderate which shows that the model is relatively good.
* **ROC\_ AUC score: 0.735**

**Performance metrics of LDA**

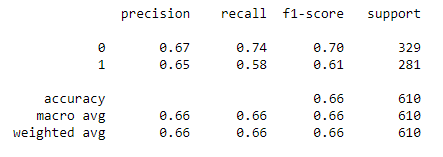
**Train data analysis:**

**Confusion matrix:**



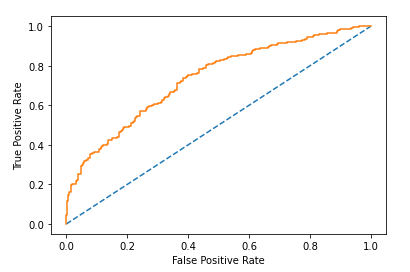
* True negative:243
* False negative:119
* False positive:86
* True positive:162

**Classification Report:**



* **Precision: 0.65**
* **Recall: 0.58**
* **F1-score: 0.61**
* **Accuracy: 0.66**

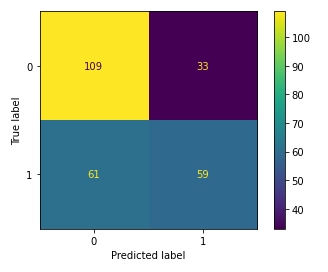
**ROC curve:**



* The curve is appearing to be near the y axis.
* The area under the the curve shows that the model is relatively good.
* **ROC\_ AUC score: 0.733**

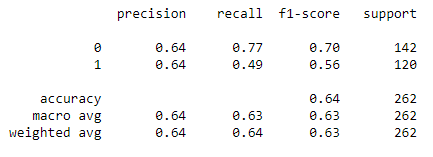
**Test data analysis:**

**Confusion matrix:**



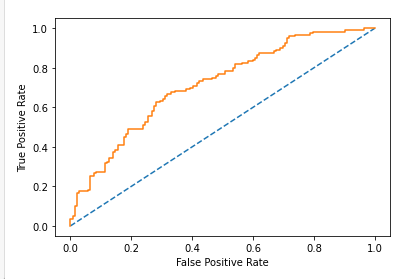
* True negative:109
* False negative:61
* False positive:33
* True positive:59

**Classification Report:**



* **Precision: 0.64**
* **Recall: 0.49**
* **F1-score: 0.56**
* **Accuracy: 0.64**

**ROC curve:**



* The curve is appearing to be near the y axis.
* The area under the the curve shows that the model is relatively good.
* **ROC\_ AUC score: 0.733**
* I consider precision and recall value in this case to choose the best model.
* I choose precision and recall as deciding parameter because it gives the ability of the model to show the relevent data.
* Based on the performance metrices of the test data of logistic regression model and LDA from the above ,we can conclude that logistic regression performs better than LDA .

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

* If the employee is foreigner there is high chance for opting holiday package so some special offers can be given to the domestic employees to encourage them to choose the holiday package.
* If the employee is having young children or infants, then the chance of opting holiday package is less.
* The holiday package can be modified to attract the young children as well.
* The employees having high salary is not opting for holiday package.
* The company can have a look at employees with high salary and concentrate on them to sell their holiday package.
* The age of the employee does not matter in the selection of the holiday package.
* Many new offers to attract domestic employees should be introduced.
* Attractive destination points can be given in holiday package to cover the employees.
* Special attention can be given in kids care and comfort to pull the employees with young children.