

Predictive Modelling

Project Report

Siddhant Raj Singh

PGP

-

DSBA Online

10th October, 2022

Table of Contents

Contents

Problem 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. 5](#_Toc32307)

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. [19](#_Toc32317)

[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. 21](#_Toc32318)

1.4 Inference: Basis on these predictions, what are the business insights and recommendations.... [29](#_Toc32319)

Problem 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. 31](#_Toc32307)

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)..... [43](#_Toc32317)

[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.. 48](#_Toc32318)

[2.4 Inference: Basis on these predictions, what are the insights and recommendations.. 63](#_Toc32318)

List of Figures

[Figure 1 Sample Data 6](#_Toc117054264)

[Figure 2 Sample Data Description 7](#_Toc117054265)

[Figure 3 Box Plot Cubic Zirconia 9](#_Toc117054266)

[Figure 4 Distribution Plot Cubic Zirconia 10](#_Toc117054267)

[Figure 5 Pair Plot Cubic Zirconia 12](#_Toc117054268)

[Figure 6 Heat Map Cubic Zirconia 13](#_Toc117054269)

[Figure 7 Cut Count Plot 14](#_Toc117054270)

[Figure 8 Cut vs Price Bar Plot 15](#_Toc117054271)

[Figure 9 Cut vs Price Box Plot 16](#_Toc117054272)

[Figure 10 Clarity Count Plot 16](#_Toc117054273)

[Figure 11 Clarity vs Price Bar Graph 17](#_Toc117054274)

[Figure 12 Clarity vs Price Box Plot 18](#_Toc117054275)

[Figure 13 Box Plot Outlier Treatment Zirconia 20](#_Toc117054276)

[Figure 14 Sub Level Combining Zirconia 21](#_Toc117054277)

[Figure 15 Data Encoding Cubic Zirconia 22](#_Toc117054278)

[Figure 16 Scaled Data Cubic Zirconia 22](#_Toc117054279)

[Figure 17 OLS Regression Results 25](#_Toc117054280)

[Figure 18 Initial VIF Values 26](#_Toc117054281)

[Figure 19 OLS Regression Result Final 27](#_Toc117054282)

[Figure 20 VIF Final 28](#_Toc117054283)

[Figure 21 Actual vs Predicted Cubic Zirconia 28](#_Toc117054284)

[Figure 22 y\_pred vs y\_actual Cubic Zirconia 30](#_Toc117054285)

[Figure 23 Sample Dataset Holiday Package 32](#_Toc117054286)

[Figure 24 Sample Data Description Holidayp 33](#_Toc117054287)

[Figure 25 Box Plot Holiday Package 34](#_Toc117054288)

[Figure 26 Distribution Plot Holiday Package 35](#_Toc117054289)

[Figure 27 Holiday Package and foreign Countplot 36](#_Toc117054290)

[Figure 28 Holiday\_Package vs Salary 37](#_Toc117054291)

[Figure 29 Foreign vs Salary 38](#_Toc117054292)

[Figure 30 Holiday Package vs Age 38](#_Toc117054293)

[Figure 31 Holiday Package vs educ 39](#_Toc117054294)

[Figure 32 Pair Plot Holiday Package 40](#_Toc117054295)

[Figure 33 Heat Map Holiday Package 41](#_Toc117054296)

[Figure 34 Boxplot outlier treatment Holiday Package 43](#_Toc117054297)

[Figure 35 Data Encoding Holiday Package 44](#_Toc117054298)

[Figure 36 Predicted Classes LR 45](#_Toc117054299)

[Figure 37 Predicted Classes LR Grid 46](#_Toc117054300)

[Figure 38 Predicted Classes LDA 47](#_Toc117054301)

[Figure 39 Confusion Matrix Training Data LR 48](#_Toc117054302)

[Figure 40 Confusion Matrix Test Data LR 49](#_Toc117054303)

[Figure 41 Classification Report Training Data LR 50](#_Toc117054304)

[Figure 42 Classification Report Test Data LR 50](#_Toc117054305)

[Figure 43 auc\_roc Curve training data LR 51](#_Toc117054306)

[Figure 44 auc\_roc curve test data LR 52](#_Toc117054307)

[Figure 45 Confusion Matrix Training Data LR Grid 53](#_Toc117054308)

[Figure 46 Confusion Matrix Test Data LR Grid 53](#_Toc117054309)

[Figure 47 Classification Report Train data LR Grid 54](#_Toc117054310)

[Figure 48 Classification Report Test data LR Grid 54](#_Toc117054311)

[Figure 49 auc\_roc curve Train Data LR Grid 55](#_Toc117054312)

[Figure 50 auc\_roc curve Test Data LR Grid 56](#_Toc117054313)

[Figure 51 Confusion Matrix Train data LDA 57](#_Toc117054314)

[Figure 52 Confusion Matrix Test Data LDA 58](#_Toc117054315)

[Figure 53 Class Report Train Data LDA 59](#_Toc117054316)

[Figure 54 Class Report Test Data LDA 59](#_Toc117054317)

[Figure 55 auc\_roc curve Train Data LDA 60](#_Toc117054318)

[Figure 56 auc\_roc curve Test Data LDA 61](#_Toc117054319)

# **Problem 1:**

# 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.

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.

# Data Description

|  |  |
| --- | --- |
| **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. |
| Colour | Colour of the cubic zirconia. With D being the best and J the worst. |
| Clarity | Clarity refers to the absence of the Inclusions and Blemishes. (In order from Best to Worst 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

Description automatically generated

Figure Sample Data

* Shape of the data is (26967, 10) i.e. 10 columns and 26967 rows.
* The sample of the dataset i.e. the first five rows appear to be perfect.
* After checking and gathering more information about the data, we find that there are total 26967 entries in the dataset and columns are of data type float64, object and int64.
* The column ‘Unnamed: 0’ can be removed from the data set as it’s an index row and will contribute nothing to our model.

# 

# Exploratory Data Analysis

## 

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

# Column Non-Null Count Dtype

--- ------ ------------------ ---------

0 carat 26967 non-null float64

1 cut 26967 non-null object

2 colour 26967 non-null object

3 clarity 26967 non-null object

4 depth 26270 non-null float64

5 table 26967 non-null float64

6 x 26967 non-null float64

7 y 26967 non-null float64

8 z 26967 non-null float64

9 price 26967 non-null int64

* There are total 26967 rows and 10 columns in the dataset. All the columns are of float64 or object or int64 type.
* We have both Categorical and Continuous data.
* There are few null values in the data set as observed from the above table.

## Describing the Data:

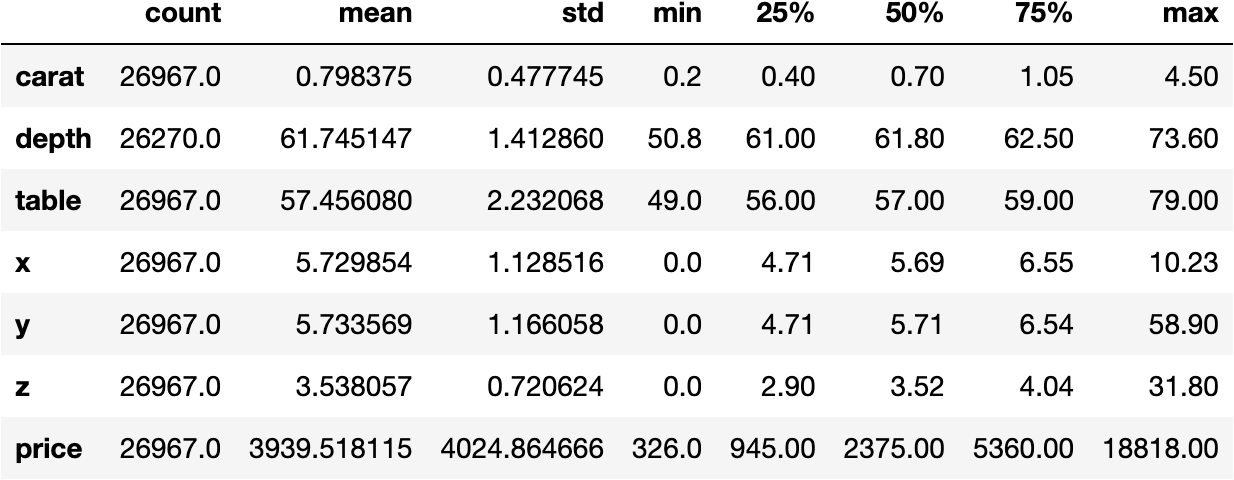


Figure Sample Data Description

* The variable ‘carat’ tells us the weight of the cubic zirconia. From the above table we can conclude that the average weight of the cubic zirconia is 0.798375 with a standard deviation of 0.477745. The median for the same is 0.70 which is less than the mean. Therefore we can say that the variable ‘carat’ is slightly skewed to the right. The maximum and minimum weight of cubic zirconia is 0.2 and 4.50 respectively.
* The variable ‘depth’ tells us the height of cubic zirconia, measured from the Culet to the table, divided by its average Girdle Diameter. From the above table we can conclude that the average depth of the cubic zirconia is 61.745 with a standard deviation of 1.412860. The median for the same is 61.80 which is slightly greater than the average. The difference between average and median is very less. We might say that the variable ‘depth’ is approximately normally distributed. However, the variable ‘depth’ is slightly skewed to the left. The maximum and minimum depth of cubic zirconia is 73.60 and 50.8 respectively.
* The variable ‘table’ tells us the width of the cubic zirconia’s Table which is expressed as a Percentage of its Average Diameter. A diamond’s table is the facet which can be seen when the stone is viewed face up. From the above table we can conclude that the average Table width is 57.4560% with a standard deviation of 2.232. The median for the same is 57.00% which is slightly is less than the mean. Therefore we can say that the variable ‘table’ is slightly skewed to the right.
* The variable ‘x’ tell us the Length of the cubic zirconia in mm. From the above table we can conclude that the average length of the cubic zirconia is 5.729mm with a standard deviation of 1.128mm. The median for the same is 5.69mm which is less than the mean. Therefore we can say that the variable ‘x’ is slightly skewed to the right. The maximum and minimum length of cubic zirconia is 10.23 mm and 0.0 mm respectively.
* The variable ‘y’ tell us the Width of the cubic zirconia in mm. From the above table we can conclude that the average width of the cubic zirconia is 5.733mm with a standard deviation of 1.166mm. The median for the same is 5.71mm which is slightly less than the mean. Therefore we can say that the variable ‘y’ is slightly skewed to the right. The maximum and minimum width of cubic zirconia is 58.90 mm and 0.0 mm respectively.
* The variable ‘z’ tell us the Height of the cubic zirconia in mm. From the above table we can conclude that the average height of the cubic zirconia is 3.538 mm with a standard deviation of 0.72 mm. The median for the same is 3.52 mm which is slightly less than the mean. Therefore we can say that the variable ‘z’ is slightly skewed to the right. The maximum and minimum height of cubic zirconia is 31.80 mm and 0.0 mm respectively.
* The variable ‘price’ tell us the price of the cubic zirconia. From the above table we can conclude that the average price of the cubic zirconia is 3939.518 with a standard deviation of 4024.864. The median for the same is 2375 which is less than the mean. Therefore we can say that the variable ‘price’ is highly skewed to the right. The maximum and minimum price of cubic zirconia is 18818.00 and 326.00 respectively.

## Check for faulty values in the dataset:

* The minimum value of the column x, y and z is 0. Since x, y and z are length, width and height of the cubic zirconia respectively we can say that 0 is a faulty value and thus should be removed from the dataset.
* After removing the faulty values for the column ‘x’, ‘y’ and ‘z’, we get the following shape of the data:
  + **(26958, 10)**

## Check for duplicate values in the dataset:

* We have 33 duplicate values in our dataset. This may also contaminate the training data with the test data or vice versa.
* We will proceed by dropping the duplicate values. This will result for the model to better generalize to the full dataset.
* Following is the shape of the data after removing the duplicates:
  + **(26925, 10)**

Univariate Analysis

Box Plot

We can see, there are lot of features which have outliers. So we might need to treat those before building model.

Distribution Plot

Chart

Description automatically generated

Figure Distribution Plot Cubic Zirconia

* For the variable ‘carat’ we can see that there could be chance of multi modes in the dataset. The dist. plot shows the distribution of data from 0.2 to 2.3 approximately. We can clearly see from the distribution plot that the variable is positively skewed.
* For the variable ‘depth’ we can see that there is single mode in the dataset. The dist. plot shows the distribution of data from 57 to 65 approximately. We can clearly see from the distribution plot that the variable is almost normally distributed.
* For the variable ‘table’ we can see that there could be chance of multi modes in the dataset. The dist. plot shows the distribution of data from 53 to 64. We can clearly see from the distribution plot that the variable is positively skewed.
* For the variable ‘x’ we can see that there could be chance of multi modes in the dataset. The dist. plot shows the distribution of data from 3.8 to 9.0 approximately. We can clearly see from the distribution plot that the variable is positively skewed.
* For the variable ‘y’ we can see that there could be chance of multi modes in the dataset. The dist. plot shows the distribution of data from 5 to 8 approximately. We can clearly see from the distribution plot that the variable is positively skewed.
* For the variable ‘z’ we can see that there could be chance of multi modes in the dataset. The dist. plot shows the distribution of data from 3 to 8 approximately. We can clearly see from the distribution plot that the variable is positively skewed.
* For the variable ‘price’ we can see that there could be chance of multi modes in the dataset. The dist. plot shows the distribution of data from 500 to 12000 approximately. We can clearly see from the distribution plot that the variable is positively skewed.

## Bivariate Analysis

## Pair Plot

A picture containing diagram

Description automatically generated

Figure Pair Plot Cubic Zirconia

* The above Pairplot shows the relationship between the variables present in the Data Frame

using scatter plot and distribution of the variable using Histogram.

* There is a strong positive correlation between the variable ‘carat’ and ‘x’. This means larger

length of cubic zirconia will result in higher carat weight.

* There is a strong positive correlation between the variable ‘carat’ and ‘y’. This means larger

width of cubic zirconia will result in higher carat weight.

* There is a strong positive correlation between the variable ‘carat’ and ‘z’. This means larger

height of cubic zirconia will result in higher carat weight.

* There is a strong positive correlation between the variable ‘carat’ and ‘price’. This means larger

the carat weight of cubic zirconia, higher the price for which it is sold.

* There is a very strong negative correlation between the variable ‘depth’ and ‘price’. The ‘depth’

of a diamond is its height (in millimetres) measured from the culet (bottom tip) to the table (flat, top surface). From the graph we can conclude that higher the depth of the diamond, lower the price for which it is sold.

* There is very strong negative correlation between the variable ‘depth’ and ‘table’. The ‘depth’

of a diamond is its height (in millimetres) measured from the culet (bottom tip) to the table (flat, top surface).

The ‘table’ of the diamond is the facet which can be seen when the stone is viewed face up. From the graph we can conclude that larger the ‘depth’ of the diamond means smaller the ‘table’ of the diamond.

* The variable ‘x’ is having a very strong positive correlation with variables ‘y’ and ‘z’. This means that larger the length of cubic zirconia, larger the width and height of the same and vice-versa (since variables ‘x’, ‘y’ and ‘z’ are strongly correlated with each other).

## Heat Map

Graphical user interface, Teams

Description automatically generated with medium confidence

Figure Heat Map Cubic Zirconia

From the above Heat Map we can observe the following points:

* ‘carat’ and ‘price’ are strongly correlated (0.92).
* ‘carat’ and ‘z’ are strongly correlated (0.95).
* ‘carat’ and ‘y’ are strongly correlated (0.94)
* ‘carat’ and ‘x’ are strongly correlated (0.98).
* ‘depth’ and ‘price’ are somewhat negatively correlated (-0.0027).
* ‘depth’ and ‘table’ are strongly negatively correlated (-0.3)
* ‘x’ , ‘y’ and ‘z’ are strongly correlated with each other.

## Univariate and Bivariate Analysis for Categorical Variables

Cut

Chart, bar chart

Description automatically generated

Figure Cut Count Plot

* The variable ‘cut’ refers to the cut quality of the cubic zirconia. Quality is increasing order Fair, Good, Very Good, Premium, Ideal.
* The most preferred cut is the ideal cut for the cubic zirconia as it is having the highest count. The least preferred cut for the same is Fair as it is having the lowest count.

Chart, bar chart

Description automatically generated

Figure Cut vs Price Bar Plot

* The cut type ideal is having the lowest price on an average and therefore it is the most preferred cut as seen from the above count plot.
* The cut type Fair is having the highest price on an average and therefore it is the least preferred cut as seen from the above count plot.
* The cut type premium is also having a high price on an average.

Chart, box and whisker chart

Description automatically generated

Figure Cut vs Price Box Plot

* From the bivariate analysis using boxplot, we can see that there are outliers present across all levels of quality. The median of premium cut is the highest and the ideal cut is the lowest. The lowest quartile for all the five cut quality levels are the same while the premium cut has the highest third quartile.

Clarity

Chart, bar chart

Description automatically generated

Figure Clarity Count Plot

* Diamond clarity refers to the purity and rarity of the stone, and the degree to which it presents blemishes and inclusions. The best quality of the diamond is ‘FL’ meaning Flawless followed by ‘IF’ meaning Internally Flawless. We don’t have Flawless data in our data set. The highest level of Clarity we have in our dataset is Internally Flawless. The lowest quality of Clarity that we have in our dataset is ‘l1’ which means Included.
* From the above graph we can see that the highest count of cubic zirconia is ‘Sl1’ i.e. Slightly Included followed by ‘VS2’ which means Very Slightly Included. The least count is of ‘l1’ i.e. Included. This is the lowest quality of clarity.

Chart, bar chart

Description automatically generated

Figure Clarity vs Price Bar Graph

* The clarity type ‘Sl2’ is having the highest price on an average followed by ‘Sl1’. ‘Sl1’ and ‘Sl2’ are Slightly Included clarity for cubic zirconia. ‘IF’ i.e. Internally Flawless type clarity is having pretty low average price despite being the highest quality of diamond in our dataset.

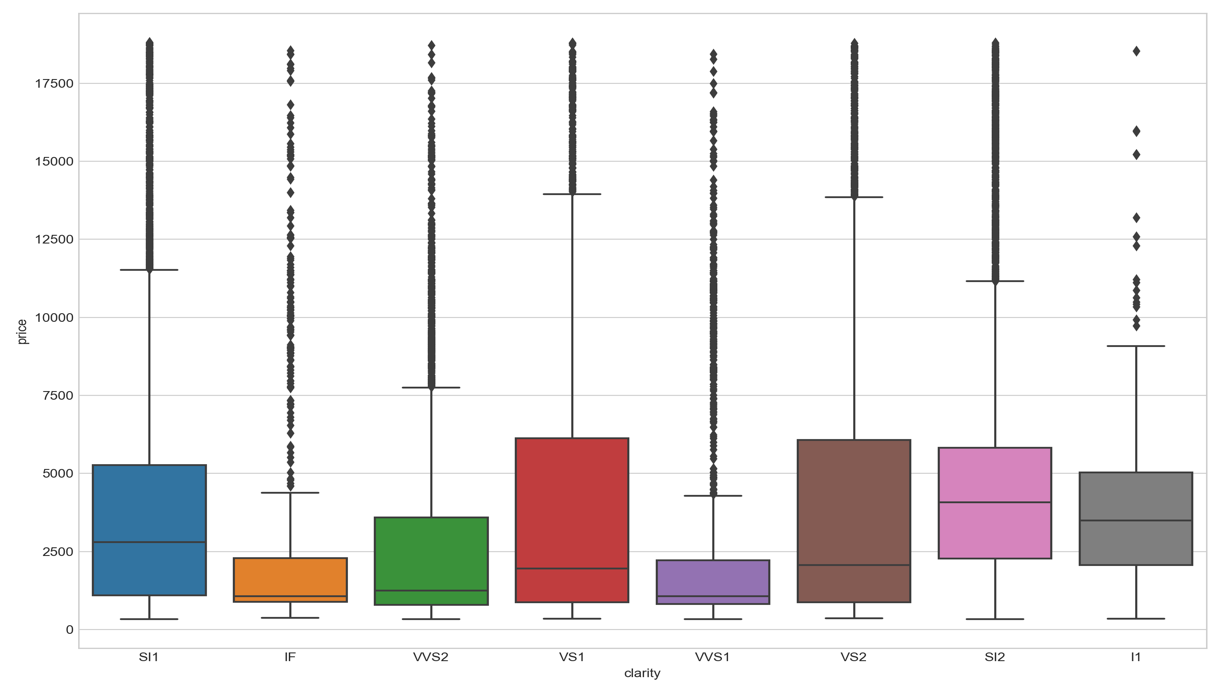


Figure Clarity vs Price Box Plot

* From the bivariate analysis using boxplot, we can see that there are outliers present across all clarity categories. The median of ‘SI2’ is the highest and the “IF’ has the lowest median. The lowest quartile for all the five cut quality levels are the same while the “VS2’ has the highest third quartile. “VVS1” has the highest number of outliers among all the colour levels.

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.

Checking for Null Values

|  |  |
| --- | --- |
| **Columns** | **Null Values** |
| carat | 0 |
| cut | 0 |
| color | 0 |
| clarity | 0 |
| depth | 697 |
| table | 0 |
| x | 0 |
| y | 0 |
| z | 0 |
| price | 0 |

* We have 697 null values for the column 'depth'. We will impute these null values with their median as there are outliers present in the dataset (Means are affected by outliers).
* After Imputing the Null Values with median we get the following results:

|  |  |
| --- | --- |
| **Columns** | **Null Values** |
| carat | 0 |
| cut | 0 |
| color | 0 |
| clarity | 0 |
| depth | 0 |
| table | 0 |
| x | 0 |
| y | 0 |
| z | 0 |
| price | 0 |

**NOTE: As we have seen from the Boxplots that we had outliers present across all the columns of the dataset, we will proceed by treating the same because treating outliers sometimes results in the models having better performance.**

Box Plot after Outlier Treatment

Chart, box and whisker chart

Description automatically generated

Figure Box Plot Outlier Treatment Zirconia

Checking for values which are equal to zero

**NOTE: We have already checked for the values which are equal to zero while performing initial diagnosis on the dataset. The minimum value of the column x, y and z was 0. They have no meaning and we had to change them or drop them. Since x, y and z are length, width and height of the cubic zirconia respectively we can say that 0 was a faulty value and thus we have removed them from the dataset.**

Combining the Sub-Levels of ordinal values

* For the column 'clarity' we will be combining few grades.
* 'clarity' of type 'SI1' and 'SI2' belong to the same grade without having much difference between them i.e. they belong to the grade 'Slightly Included'. So we will combine the sub-levels of ordinal variable 'SI1' and 'SI2' into one single grade 'Bad'
* 'clarity' of type 'VS1' and 'VS2' belong to the same grade without having much difference between them i.e. they belong to the grade 'Very Slightly Included'. So we will combine the sub-levels of ordinal variable 'VS1' and 'VS2' into one single grade 'Good'.
* 'clarity' of type 'VVS1' and 'VVS2' belong to the same grade without having much difference between them i.e. they belong to the grade 'Very Very Slightly Included'. So we will combine the sub-levels of ordinal variable 'VVS1' and 'VVS2' into one single grade 'VGood'

After combining the sub-levels of ordinal values we get the following data frame:

Text, application

Description automatically generated with medium confidence

Figure Sub Level Combining Zirconia

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.

* The columns ‘cut’, ‘color’ and ‘clarity’ are of type object. We have to convert them into 0’s and 1’s for Modelling. We will be Encoding Categorical data into integer format so that data with converted categorical value can be provided to the model to give and improve predictions.
* The columns ‘cut’, ‘color’ and ‘clarity’ are Categorical data of type ordinal. Ordinal data is a categorical data type where the variables have natural, ordered categories. This means that these possible values are ordered.
* We will Encode the data using Sklearn’s – OrdinalEncoder which takes in a parameter categories and gives it some sort of ranking.

We get the following dataset:

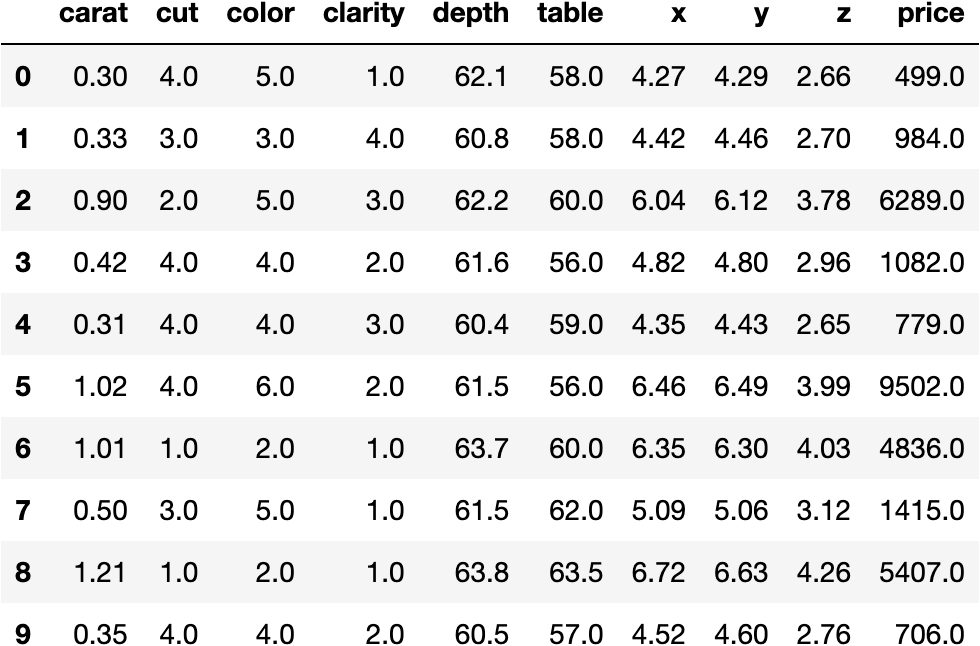


Figure Data Encoding Cubic Zirconia

Scaling the data

* To ensure that the gradient descent moves smoothly towards the minima and that the steps for gradient descent are updated at the same rate for all the features, we scale the data before feeding it to the model. Having features on a similar scale will help the gradient descent converge more quickly towards the minima.
* We will scale the data using Sklearn’s - StandardScaler. StandardScaler follows Standard Normal Distribution (SND). Therefore, it makes mean = 0 and scales the data to unit variance.

After Scaling the data we get the following :

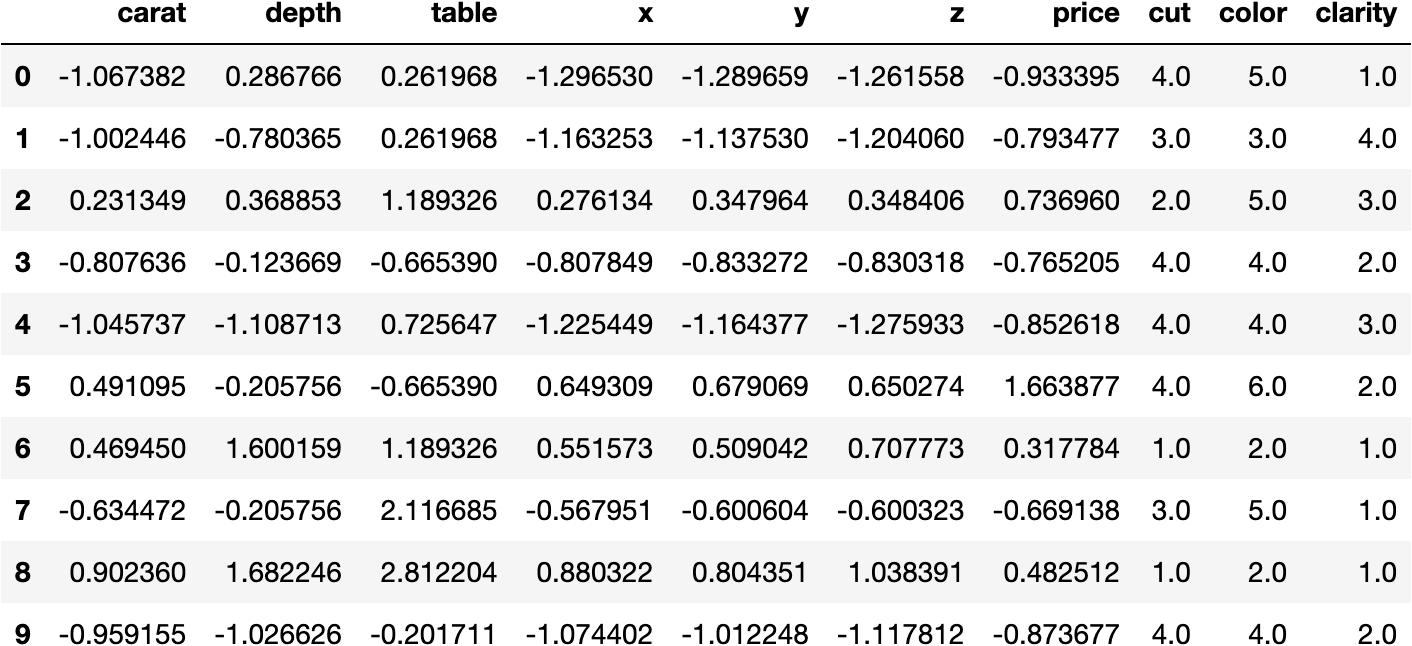


Figure Scaled Data Cubic Zirconia

Test and Train Split

* We will first create ‘X’ and ‘y’ which will contain Features/predictable variables and Target. ‘X’ will contain Features of the dataset and ‘y’ will contain the Target variable.
* To fairly evaluate our model’s performance, we don’t want to evaluate it on the same data it was trained at. Therefore, we will separate out a training set and a test set. There are four components to this – X\_train, X\_test, y\_train, y\_test. We will use Sklearn library in which we have sub-package model selection.

From model selection we will import ‘train\_test\_split’. We will keep the size of train and test in the ration of 70:30.

* We now invoke the linear regression function and find the best fit model on training data. Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.
* We create an instance of Linear Regression model. We have various parameters for the same. These parameters are called the hyperparameters. After creating the model we will train the model on the training dataset using the fit method. We will use this model which has been fit using the training dataset to make predictions of the dataset.

Checking the coefficients for each of the independent attributes

1. **The coefficient for carat is 1.1807803656823492**
2. **The coefficient for depth is 0.012117788522441456**
3. **The coefficient for table is -0.010155468098205767**
4. **The coefficient for x is -0.48377721519228417**
5. **The coefficient for y is 0.5411112530031568**
6. **The coefficient for z is -0.19361184086120325**
7. **The coefficient for cut is 0.03282543465518456**
8. **The coefficient for color is 0.07563010675843157**
9. **The coefficient for clarity is 0.228993074151028**

* ‘coef\_’ gives you an array of weights estimated by linear regression. This is the slope (m1, m2, m3,..).
* Higher slope will result in higher weightage.
* 'carat' is having highest slope and thus highest weightage.
* 'table' is having the lowest slope and lowest weightage.
  + This means more the 'carat' more the price and more the 'table' less the price.
  + The one unit increase in carat increases price by 1.1807803656823492.
  + The one unit increase in cut increases price by 0.03282543465518456.
  + The one unit increase in color increases price by 0.07563010675843157.
  + The one unit increase in clarity increases price by 0.228993074151028.
  + The one unit increase in y increases price by 0.5411112530031568.
  + The one unit increase in depth increases price by 0.012117788522441456.
  + The one unit increase in table decreases price by -0.010155468098205767.
  + The one unit increase in x decreases price by -0.48377721519228417.
  + The one unit increase in z decreases price by -0.19361184086120325.

Intercept for the model

* intercept represents the mean value of the response variable when all of the predictor variables in the model are equal to zero.
* when the other predictor variable are zero i.e. like carat, cut, color, clarity all are zero then the C=-0.762009977. (Y = m1X1 + m2X2+ ..... + mnXn + C + e) that means price is -0.762009977

R square

* R-squared (R2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.
* For our dataset:
  + R square of Training data - 0.9275358910663005
  + R square of Testing data - 0.928081803037983
* R square of 0.92 means only 92% of the variation in the price is explained by predictors in our model
* It is always between 0 and 100%, in which 0% indicates that the model explains none of the variability of the response data around its mean and 100% indicates that the model explains all the variability of the response data around its mean.
* -In the regression model we can see the R-square value on training and test data respectively as 0.9275358910663005 and 0.928081803037983.

RMSE

* This is the root of the mean of the squared error.
* It is one of the most popular metrics because it’s going to punish the larger errors by punishing them.
* For our data set:
  + RMSE for Training data - 0.26856363030921065
  + RMSE for Testing data - 0.26962717962644167
* One way to assess how well a regression model fits a dataset is to calculate the root mean square error, which is a metric that tells us the average distance between the predicted values from the model and the actual values in the dataset. The lower the RMSE, the better a given model is able to “fit” a dataset.
* For the training and testing data we have RMSE as 0.26856363030921065 and 0.26962717962644167

Linear Regression using statsmodel(OLS)

* The difference between statsmodel and linear model is that in statsmodel we get a complete summary of the model.
* This summary will by default give us values like R-Squared, Adj. R-Squared, F-statistic, p-values and many more.

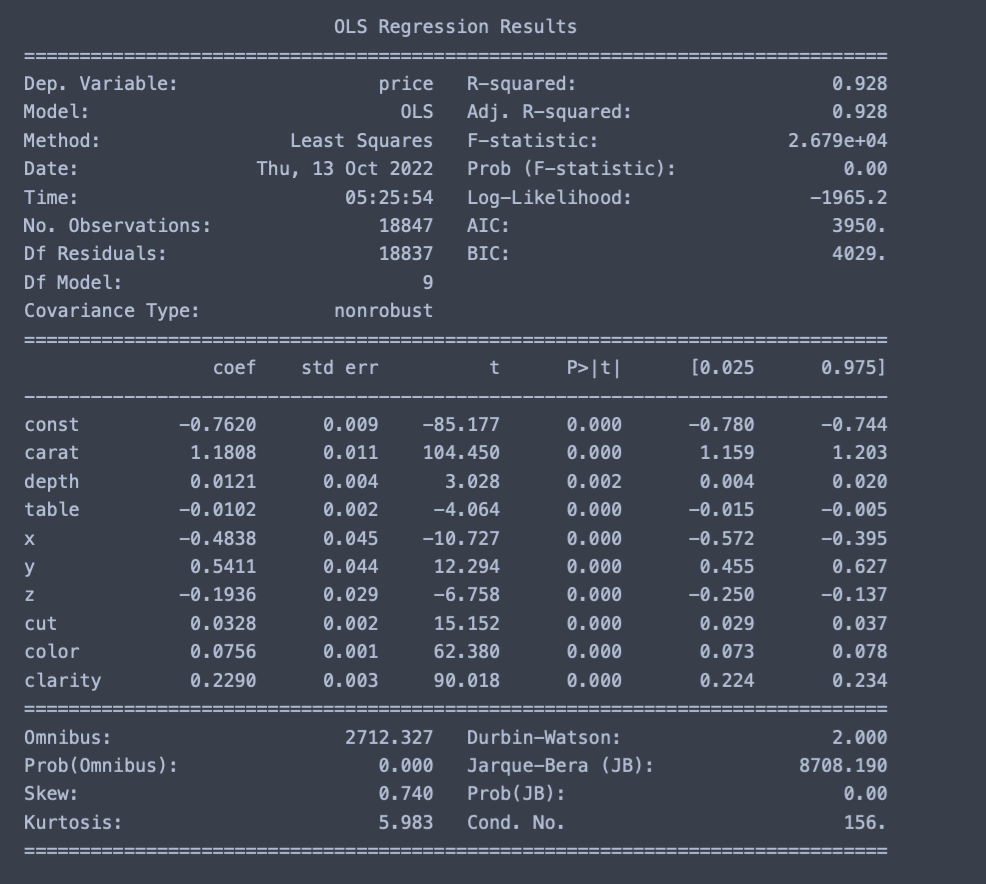


Figure OLS Regression Results

* to begin with 92.8% is our r squared value i.e. 92.8% of the variation in our y variable which is the price of the cubic zirconia is explained by the predictors in our model.
* Adj. R square is a modified version of R square that also accounts for predictors that are not significant in a regression model. The adj. R square shows whether adding additional parameters improve a regression model or not.
* Adding more independent variables to a regression model tends to increase the R-squared value, which would result in adding even more variables for our model. This is called overfitting and can return an unwarranted high R-squared value. Adjusted R-squared is used to determine how reliable the correlation is and how much it is determined by the addition of independent variables.
* Compared to a model with additional input variables, a lower adjusted R-squared indicates that the additional input variables are not adding value to the model.
* Compared to a model with additional input variables, a higher adjusted R-squared indicates that the additional input variables are adding value to the model.
* The Null and Alternate hypothesis of linear regression is as follows:
  + H\_0: There is no relation between x and y
  + H\_A: There is a relation between x and y
* The p-value, or probability value, tells you how likely it is that your data could have occurred under the null hypothesis.
* for all independent variables in our dataset the p-value is zero which means that we reject the null hypothesis and accept the alternate hypothesis i.e. for all the independent variables in our dataset there is a relation between x and y.
* the coefficients in the summary table indicate array of weights estimated by linear regression. This is the slope (m1, m2, m3,..). Higher slope will result in higher weightage.
* The coefficients tell us how one unit change in X can affect y. The sign of the coefficient indicates if the relationship is positive or negative.
* In our dataset for example an increase of carat i.e. the weight of the diamond by one unit will increase the price of diamond by 1.1808.
* If the level of significance is set to 5% (0.05), the p-values greater than 0.05 would indicate that the corresponding predictor variables are not significant.
* However, due to the presence of multicollinearity in our data, the p-values will also change.
* We need to ensure that there is no multicollinearity in order to interpret the p-values.

Check Multicollinearity

* There are different ways to indicate (or test) multicollinearity. One such way is Variation Inflation Factor.
* VIF measures the inflation in the variances of the regression coefficients estimates due to collinearities that exist among the predictors. It is a measure of how much the variance of the estimated regression is inflated by the existence of correlation among the predictable variables in the model.



Figure Initial VIF Values

* The VIF values indicate that the features carat, x, y and z are correlated with one or more independent features.
* Multicollinearity affects only the specific independent variables that are correlated. Therefore, in this case, we can trust the p-values of depth, table, cut, color and clarity variables.
* To treat multicollinearity, we will have to drop one or more of the correlated features (carat, x, y and z).
* We will drop the variable that has the least impact on the adjusted R-squared of the model.
* If VIF is 1, then there is no correlation among the 𝑘th predictor and the remaining predictor variables, and hence, the variance of 𝛽𝑘 is not inflated at all.
* If VIF exceeds 5, we say there is moderate VIF, and if it is 10 or exceeding 10, it shows signs of high multi-collinearity.
* We will create new data frames by dropping one column at a time from X\_train data frame. We will be dropping x, y, z and carat from X\_train and assign it to new variables that we create for the same. After dropping each independent variables that are having high multicollinearity, we will check R-Square and Adj. R-Square values.
* On dropping 'x', 'y' and 'z' adj. R-squared decreased by 0.001. Since adj. R-squared is not affected by dropping 'x', 'y' or 'z' we can remove them from the dataset one at a time.
* On dropping 'carat', adj. R-squared decreased by 0.042.This sharp decline indicates that 'carat' is an important predictor and shouldn't be removed.
* Since there is no effect on adj. R-squared after dropping the 'x', 'y' and 'z' column, we can remove it from the training set. We will remove it one at a time because if we drop all variables at once, we will throw away the information that is not captured by any of the three variables i.e. x, y or z.
* After removing a single variable at a time we will check the OLS Regression model results from the statsmodel and the VIF value for each variable.
* After removing all the required variables that were having high multicollinearity, we get the following statsmodel OLS Regression results:

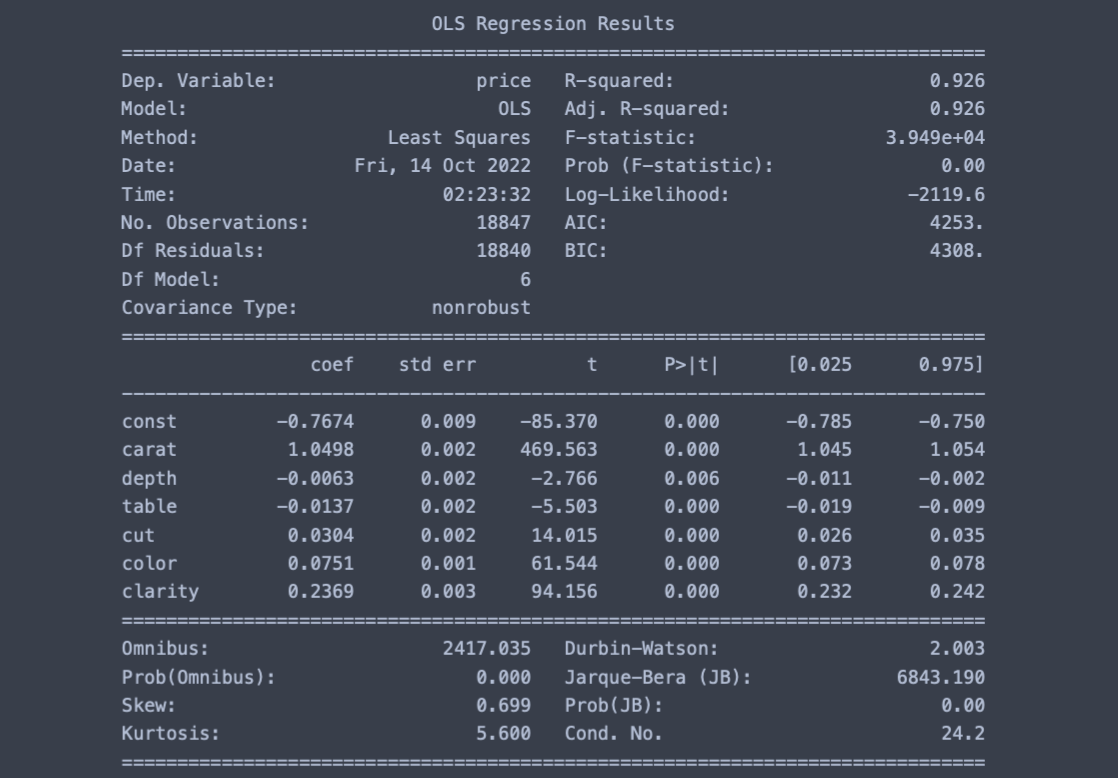


Figure OLS Regression Result Final

Following is the final VIF value for all the independent variables:

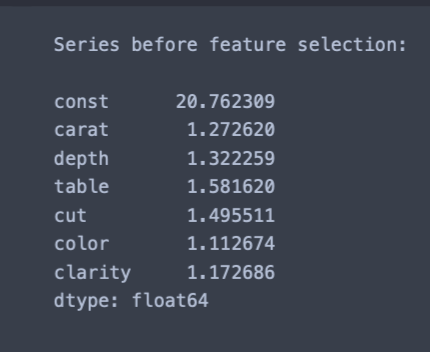


Figure VIF Final

**NOTE: Now we can observe that VIF for all the features is <2**

* Now that we do not have multicollinearity in our data, the p-values of the coefficients have become reliable and we can remove the non-significant predictor variable.
* We have created a data frame that will contain the column Actual Value which is nothing but the y\_train values, Predicted Value which is the Fitted Value from our model of statsmodel library that we had after removing multicollinearity and the Residuals which is nothing but Actual – Predicted.

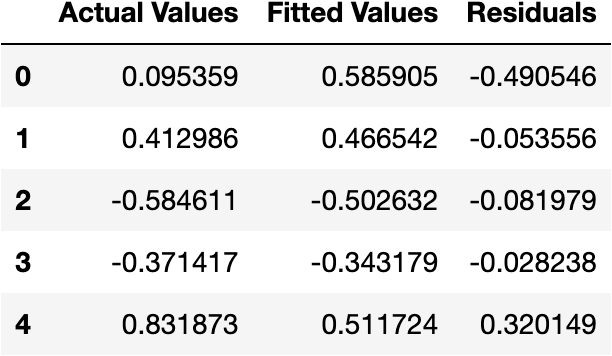


Figure Actual vs Predicted Cubic Zirconia

* The above Data Frame will also be used to calculate the Root mean Square Error and Mean Absolute Error.
  + RMSE on trained data - 0.2705122510413191
  + RMSE on test data - 0.27184897137847697
  + MAE on trained data - 0.1986602864058425
  + MAE on test data - 0.19541057067238712
* We can see that RMSE on the train and test sets are comparable. So, our model is not suffering from overfitting.
* MAE indicates that our current model is able to predict price within a mean error of 0.19 units on the test data.
* Hence, we can conclude the OLS Regression model result is good for prediction as well as inference purposes.

Linear Regression Final Equation:

***price = -0.7673715445099523 + 1.0497822414930829 \* ( carat ) + -0.006269671621101757 \* ( depth ) + -0.013654101923302173 \* ( table ) + 0.030390674652724495 \* ( cut ) + 0.07514752728304533 \* ( color ) + 0.236933010356209 \* ( clarity )***

* 1. Inference: Basis on these predictions, what are the business insights and recommendations.
* When carat increases by 1 unit, diamond price increases by 1.04978 units, keeping all other predictors constant.
* When cut increases by 1 unit, diamond price increases by 0.03039 units, keeping all other predictors constant.
* When clarity increases by 1 unit, diamond price increases by 0.236933 units, keeping all other predictors constant.
* when table increases by 1 unit, diamond price decreases by 0.013654 units, keeping all other predictors constant.
* when depth increases by 1 unit, diamond price decreases by 0.006269 units, keeping all other predictors constant.
* As per model these six attributes are the most important attributes for predicting target variable 'price' - 'Carat', 'clarity', 'Cut', 'color', 'table' and 'depth'.

Chart, scatter chart

Description automatically generated

Figure y\_pred vs y\_actual Cubic Zirconia

* we can see that the from the scatter plot between y predicted and y actual that it is a very strong corelation between the predicted y and actual y but there are lots of spread. That indicates some kind noise present on the data set i.e. Unexplained variances on the output.
* Linear regression Performance Metrics
  + intercept for the model: -0.7673715445099523
  + R square on training data: 0.9275358910663005
  + R square on testing data: 0.928081803037983
  + RMSE on Training data: 0.26856363030921065
  + RMSE on Testing data: 0.26962717962644167
* As the training data & testing data score are almost inline, we can conclude this model is a Right-Fit Model.
* The Gem Stones company should consider the features 'Carat', 'Cut', 'color', 'clarity', 'depth' and 'table' as most important for predicting the price. This will help us to distinguish between higher profitable stones and lower profitable stones so as to have better profit share.
* Stones having higher carat should be considered as the most profit making stones among all others.
* The 'Premium Cut' and 'ideal' Diamonds are the most Expensive, followed by 'Very Good' Cut, these should consider in higher profitable stones.
* We have found that Cut, Depth and Table attributes importance in predicting the price of cubic zirconia is the least as per the given data. So Business can check and analyse these features using some more dataset to understand the impact.
* The Diamonds clarity with 'VS1' &'VS2' are the most Expensive. So these two category also consider in higher profitable stones.
* Stones having higher weight and lower depth will earn huge profits.

# **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.

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.

# Data Description

|  |  |
| --- | --- |
| **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 |

# Sample of the dataset:

Table

Description automatically generated

Figure Sample Dataset Holiday Package

* Shape of the data is (872, 7) i.e. 7 columns and 872 rows.
* The sample of the dataset i.e. the first five rows appear to be perfect.
* After checking and gathering more information about the data, we find that there are total 872 entries in the dataset and columns are of data type object and int64.
* The column ‘Unnamed: 0’ can be removed from the data set as it’s an index row and will contribute nothing to our model.

# Exploratory Data Analysis

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

# Column Non-Null Count Dtype

------------------ --------------------- ---------

0 Holiday\_Package 872 non-null object

1 Salary 872 non-null int64

2 age 872 non-null int64

3 educ 872 non-null int64

4 no\_young\_children 872 non-null int64

5 no\_older\_children 872 non-null int64

6 foreign 872 non-null object

* There are total 872 rows and 7 columns in the dataset. All the columns are of object or int64 type.
* We have both Categorical and Continuous data.
* There are no null values in the data set as observed from the above table.

## Describing the Data:

Table

Description automatically generated

Figure Sample Data Description Holidayp

* The variable ‘Holiday\_Package’ tells us whether the employees opted for Holiday Package or not. It is a categorical variable. From the above table we can conclude that most of the employees did not opt for holiday package which is suggested by column ‘top’ in the above table with a count of 471 employees. The same is suggested from the column ‘freq’. The unique number of values for this variable is 2 i.e. yes or no.
* The variable ‘Salary’ tell us the Employee Salary. From the above table we can conclude that the average salary of the employee is 47729.1720 with a standard deviation of 23418.668. The median for the same is 41903.5 which is less than the mean. Therefore we can say that the variable ‘Salary’ is slightly skewed to the right. The maximum and minimum salary of Employee is 236961.0 and 1322.0 respectively.
* The variable ‘age’ tell us age of Employees in years. From the above table we can conclude that the average age of employees in years is around 40 years with a standard deviation of 10.55 years. The median for the same is 39.0 years which is slightly less than the mean. Therefore we can say that the variable ‘age’ is slightly skewed to the right. The maximum and minimum age of employees is 62.0 years and 20.0 years respectively.
* The variable ‘educ’ tell us years of formal education an employee has had. From the above table we can conclude that the average years of formal education is around 9.3 with a standard deviation of around 3 years. The median for the same is 9.0 years which is very slightly less than the mean. Therefore we can say that the variable ‘educ’ is very slightly skewed to the right. The maximum and minimum years of formal education that employees has had is 21.0 years and 1.0 years respectively.
* The variable ‘no\_young\_children’ tell us the number of young children i.e. children younger than 7 years that employee has. From the above table we can conclude that the maximum number of young children that an employee has is 3 and minimum number of young children that an employee has is 0. We can also observe that 75% of the employee has no young children.
* The variable ‘no\_older\_children’ tell us the number of older children. From the above table we can conclude that the maximum number of older children that an employee has is 6 and minimum number of older children that an employee has is 0. We can also observe that 75% of the employees has 2 older children.
* The variable ‘foreign’ tells us whether the employees are foreigner or not. It is a categorical variable. From the above table we can conclude that most of the employees are not foreigner which is suggested by column ‘top’ in the above table with a count of 656 employees. The same is suggested from the column ‘freq’. The unique number of values for this variable is 2 i.e. yes or no.

## Check for duplicate values in the dataset:

* We have no duplicate values in the dataset.

## Check for missing values in the dataset:

* We have no missing values in the dataset.

Univariate Analysis

Box Plot

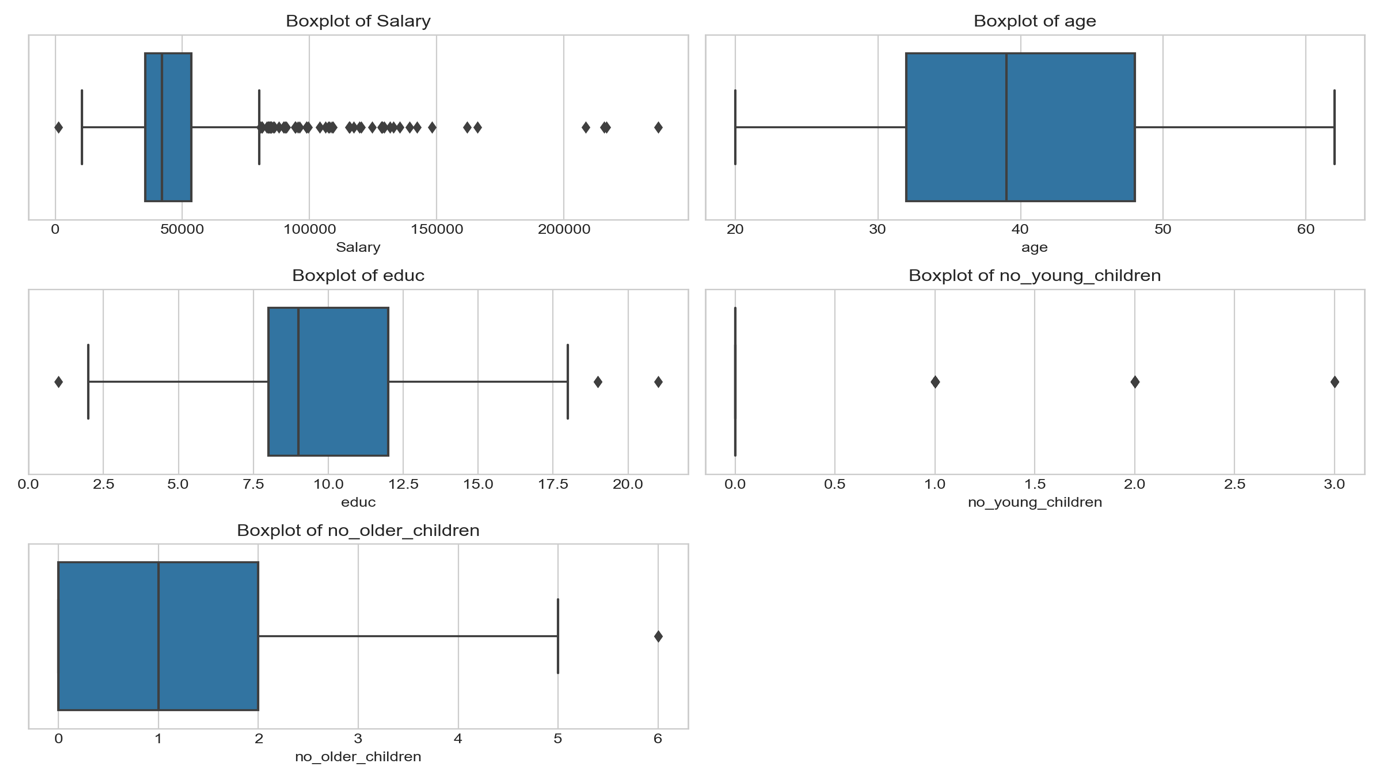


Figure Box Plot Holiday Package

* The variable ‘Salary’ is slightly skewed to the right. The median number of ‘Salary’ is approximately 41000. From the above graph we can see that Q1 = 35000 and Q3 = 53000 approximately. This means that Inter-Quartile range is approximately 18000 which means about 50% of ‘Salary’ is approximately between 53000 and 35000. The variable ‘Salary’ has a lot of outliers.
* The variable ‘age’ is slightly skewed to the right. The median number of ‘age’ is approximately 39 years. From the above graph we can see that Q1 = 32 and Q3 = 48 approximately. This means that Inter-Quartile range is approximately 16 years which means about 50% of ‘age’ is approximately between 33 and 48. The variable ‘age’ has no outliers.
* The variable ‘educ’ is slightly skewed to the right. The median number of ‘educ’ is approximately 8.5. From the above graph we can see that Q1 = 8 and Q3 = 12 approximately. This means that Inter-Quartile range is approximately 4 which means about 50% of ‘educ’ is approximately between 8 and 12. The variable ‘educ’ has few outliers.
* The variable ‘no\_young\_children’ is slightly skewed to the right. The median number of ‘no\_young\_children’ is 0. From the above graph we can see that Q1 = 0 and Q3 = 0. This means that Inter-Quartile range is also 0. The variable ‘no\_young\_children’ has no outliers. From the above graph we can conclude that the maximum number of young children that an employee has is 3 and minimum number of young children that an employee has is 0.
* The variable ‘no\_older\_children’ is almost normally distributed. The median number of ‘no\_older\_children’ is approximately 1.0. From the above graph we can see that Q1 = 0 and Q3 = 2. This means that Inter-Quartile range is 2 which means about 50% of ‘no\_older\_children’ is between 0 and 2. The variable ‘no\_older\_children’ has one outlier.

Distribution Plot

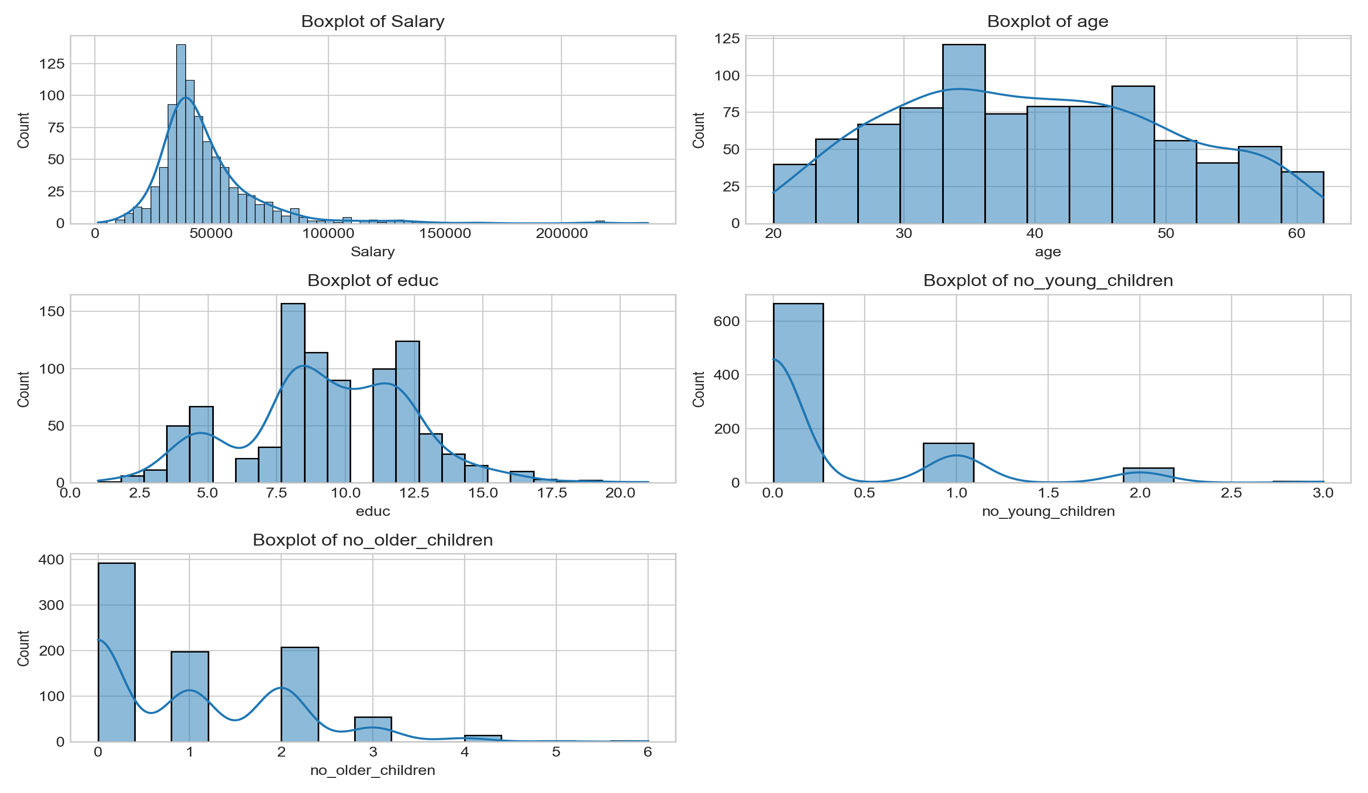


Figure Distribution Plot Holiday Package

* For the variable ‘Salary’ we can see that there is single mode in the dataset. The dist. plot shows the distribution of data from 0 to 200000 approximately. We can clearly see from the distribution plot that the variable is positively skewed. We can see that most of the ‘Salary’ for the employee is in the range 25000 to 55000.
* For the variable ‘age’ we can see that there could be chance of multi modes in the dataset. The dist. plot shows the distribution of data from 20 to 60 approximately. We can clearly see from the distribution plot that the variable is positively skewed. We can see that most of the ‘age’ for employee is in the range 33 to 45 years.
* For the variable ‘educ’ we can see that there could be chance of multi modes in the dataset. The dist. plot shows the distribution of data from 2.5 to 17.5 approximately. We can clearly see from the distribution plot that the variable is positively skewed. We can see that most of the ‘educ’ for employee is in the range 7.5 to 12.5 years.
* For the variable ‘no\_young\_children’ minimum number of children is 0 and maximum number of children is 3. Out of the total 872 employees we have 665 employees that do not have children younger than 7 years. 147 employees have one young child while 55 employees have 2 young children and only 3 employees have 3 young children.
* For the variable ‘no\_older\_children’ minimum number of children is 0 and maximum number of children is 6. Out of the total 872 employees we have 393 employees that do not have older children. 198 employees have one elder child while 208 employees have 2 elder children, 55 employees have 3 elder children, 14 employees have 4 elder children and 2 employees have 5 elder children and another 2 employees have 6 elder children.

Univariate and Bivariate Analysis for Categorical Variables

Holiday\_Package and foreign

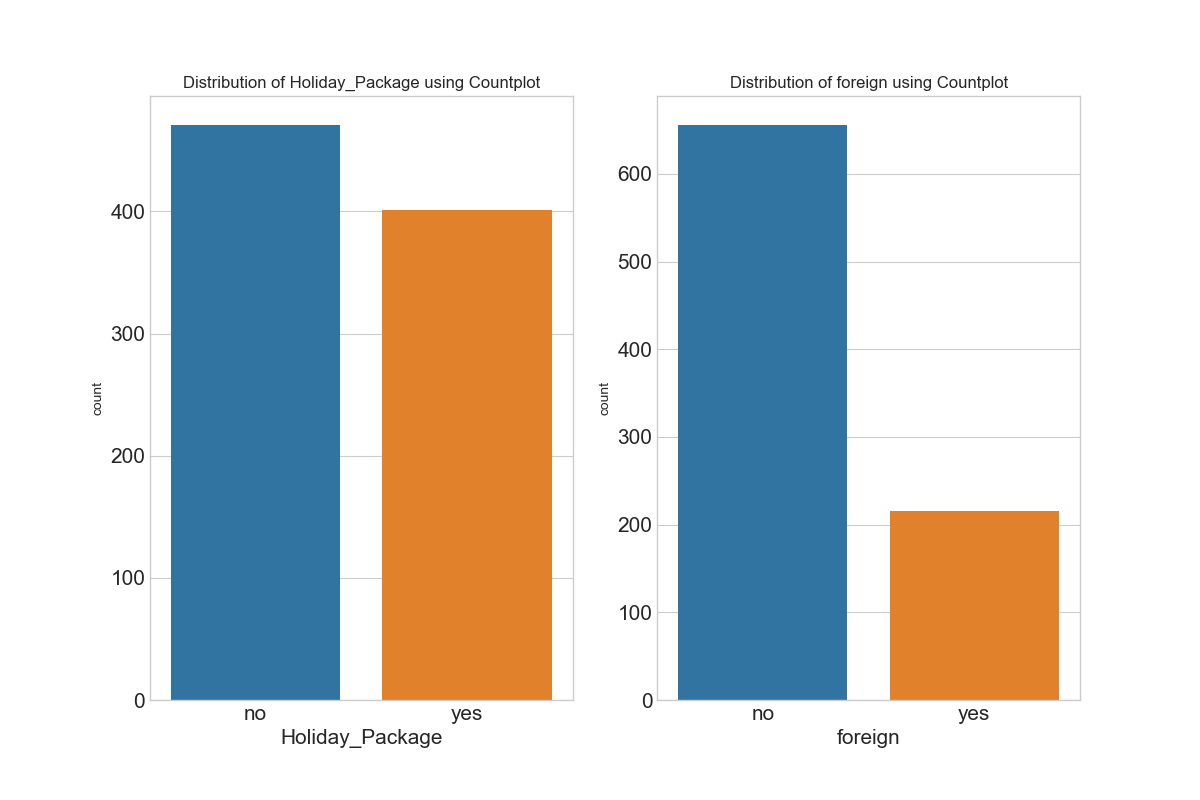


Figure Holiday Package and foreign Countplot

* From the above countplot of ‘Holiday\_Package’ we can see that less employees have opted for holiday package as compared to employees who did not opt for a holiday package. Around 401 employees have opted for the holiday package while 471 employees did not opt for the holiday package.
* From the above countplot of ‘foreign’ we can see that less employees are foreigners as compared to non-foreigners/Native. Around 656 employees are native employees while 216 employees are foreigners.

Holiday\_Package vs Salary

Chart, box and whisker chart

Description automatically generated

Figure Holiday\_Package vs Salary

* From the above Bar Plot of ‘Holiday\_Package’ and ‘Salary’ we can see that average salary of employee who did not opt for holiday package is higher as compared to salary of employee who did opt for the same.
* Across the four different boxplots, the median salary of the non-Foreigners/native employee rejecting the holiday package is highest while the median salary of the foreign employee accepting the holiday package is lowest. The employees rejecting the holiday package has more outliers.

Foreign vs Salary

Chart, box and whisker chart

Description automatically generated

Figure Foreign vs Salary

* From the above Bar Plot of ‘foreigner’ and ‘Salary’ we can see that average salary of employee who are non-foreigners/native is higher as compared to salary of employee who are foreigners.
* Across the two different boxplots, we can see that the median salary of non- foreigners/native is higher as compared to median salary of foreigners. Also, non-foreigners/native employees are having higher number of outliers as compared to foreigner employees in terms of their salary.

Holiday\_Package vs Age

Chart, box and whisker chart

Description automatically generated

Figure Holiday Package vs Age

* From the above Bar Plot of ‘Holiday-Package’ and ‘age’ we can see that employees that have not opted for holiday package are having higher age on an average as compared to employees who have opted for holiday package.
* Across the four different boxplots, we can see that the median age of non-foreigner/native employees who did not opt for holiday package is highest. The median age of foreigner employees who did opt for holiday package is lowest.

Holiday\_Package vs Educ

Chart, box and whisker chart

Description automatically generated

Figure Holiday Package vs educ

* From the above Bar Plot we can see that employees who did not opt for holiday package are having higher years of formal education on an average as compared to employees who did opt for the same.
* Across the four different boxplots, we can see that the median of years of formal education for non-foreigners employees who did opt for holiday package is equal to non-foreigner employees who did not opt for the same.

Holiday\_Package vs Young Children

Chart, box and whisker chart

Description automatically generated

* From the above box plot it is clear that for foreigner and native employees who are not having young children will opt for holiday package and the ones having young children will not opt for the same.

## Bivariate Analysis

## Pair Plot

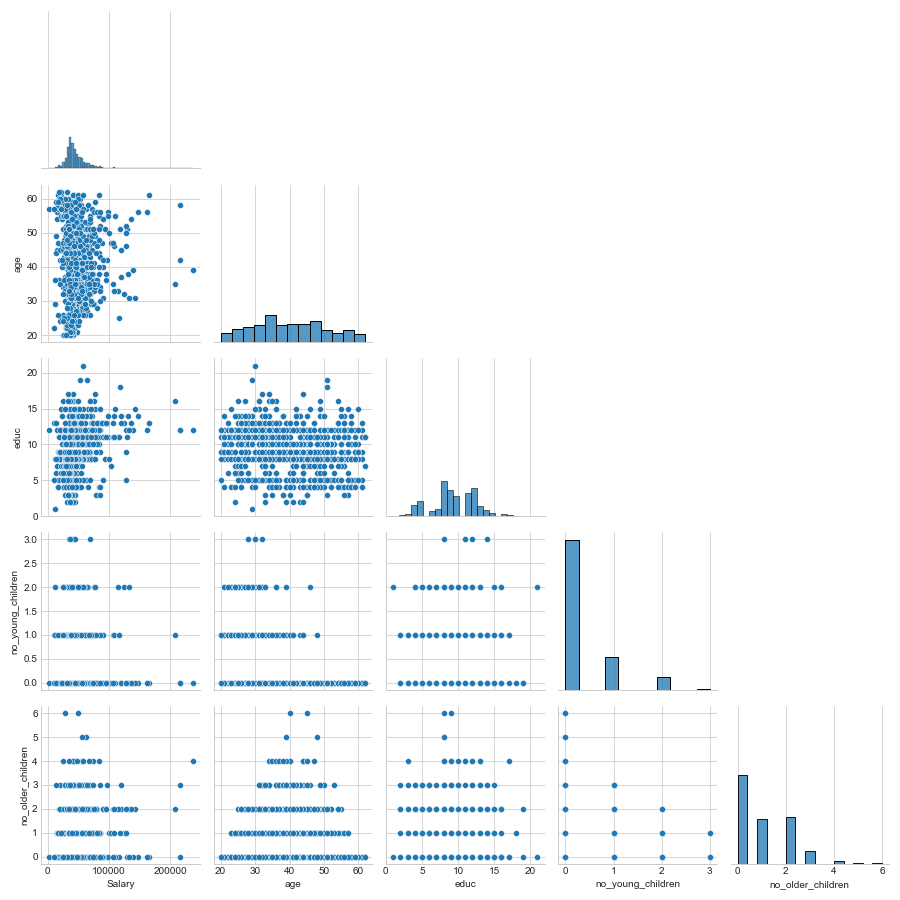


Figure Pair Plot Holiday Package

* The above Pairplot shows the relationship between the variables present in the Data Frame

using scatter plot and distribution of the variable using Histogram.

* There is a positive correlation between ‘salary’ and ‘educ’. This means that higher the years of formal education, higher the salary.
* There is a positive correlation between ‘no\_young\_children’ and ‘educ’. It’s not a weak positive correlation. However it indicated to some extent that higher the number of years of formal education means higher the number of children younger than 7 years.
* There is a positive correlation between ‘salary’ and ‘age’ but it is a weak correlation. This means that to some extent higher the age of employee means higher the salary. However there are many cases where we can see that even for the higher age, the salary is relatively low and for a lower age, the salary is relatively high.
* There is a positive correlation between ‘salary’ and ‘no\_older\_children’. It’s not a very strong positive correlation. However it indicated to some extent that higher the salary of the employee means more the number of older children.
* There is a somewhat strong negative correlation between ‘educ’ and ‘age’. This means that higher the age of the employee, lower the years of formal education. However there are a lot of cases in which at a very low age employees are having higher years of formal education and vice-versa.
* There is a very strong negative correlation between ‘age’ and ‘no\_young\_children’. This means higher the age of the employee lower the number of children younger than 7 years.

## Heat Map

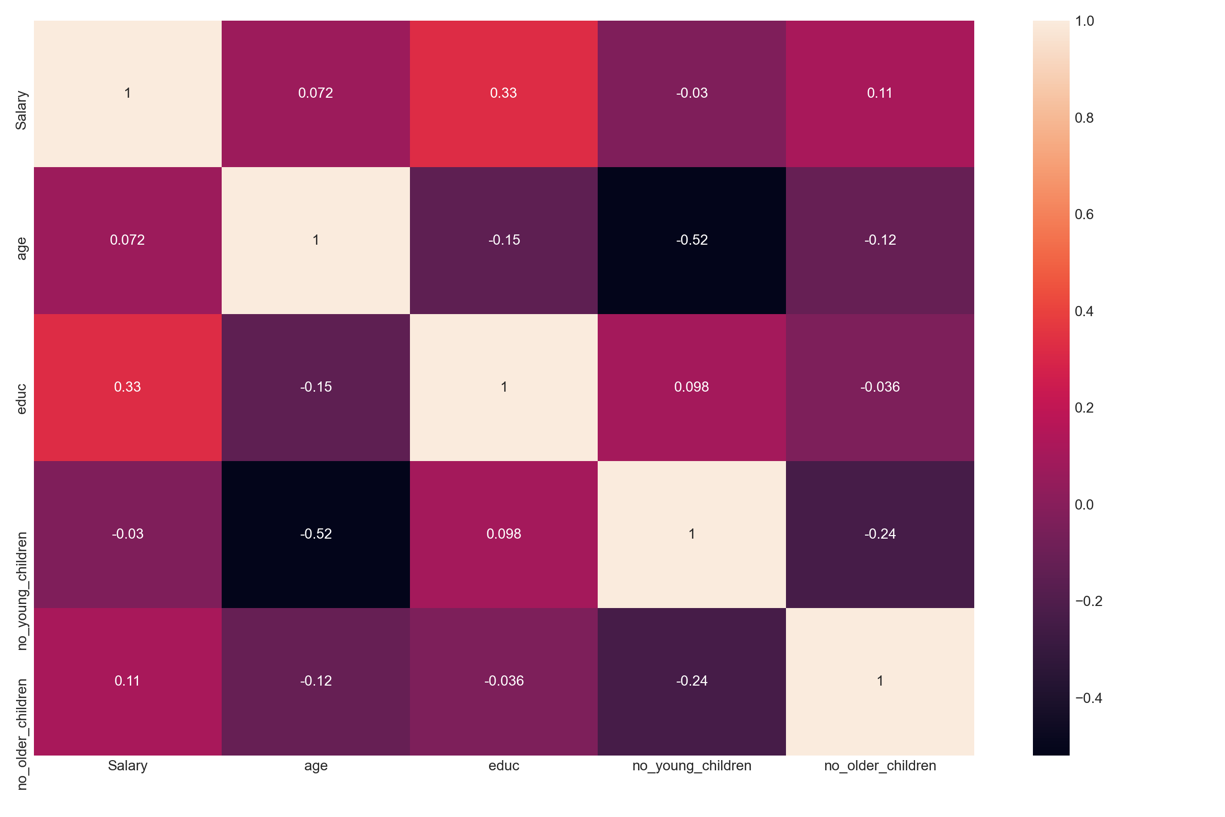


Figure Heat Map Holiday Package

From the above Heat Map we can observe the following points:

* ‘Salary’ and ‘educ’ are strongly positively correlated (0.33).
* ‘Salary’ and ‘no\_older\_children’ are positively correlated (0.11).
* ‘no\_young\_children’ and ‘educ’ are positively correlated (0.098). It is a very weak positive correlation.
* ‘educ’ and ‘age’ are negatively correlated (-0.15). It is somewhat strong negative correlation.
* ‘age’ and ‘no\_young\_children’ are strongly negatively correlated (-0.52).

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).

**NOTE: As we have seen from the Boxplots that we had outliers present across all the columns of the dataset, we will proceed by treating the same because treating outliers sometimes results in the models having better performance.**

Outlier Treatment

We have outliers in Salary, age, educ, no\_young\_children and no\_older\_children.

* An observation is considered to be an outlier if that particular observation has been mistakenly captured in the data set. Treating outliers sometimes results in the models having better performance but the models lose out on generalization.
* From the Initial diagnosis, we can say that the column age, no\_young\_children, educ and no\_older\_children are having valid outliers.
* But since there are too many outliers in the column Salary we can choose to remove it.

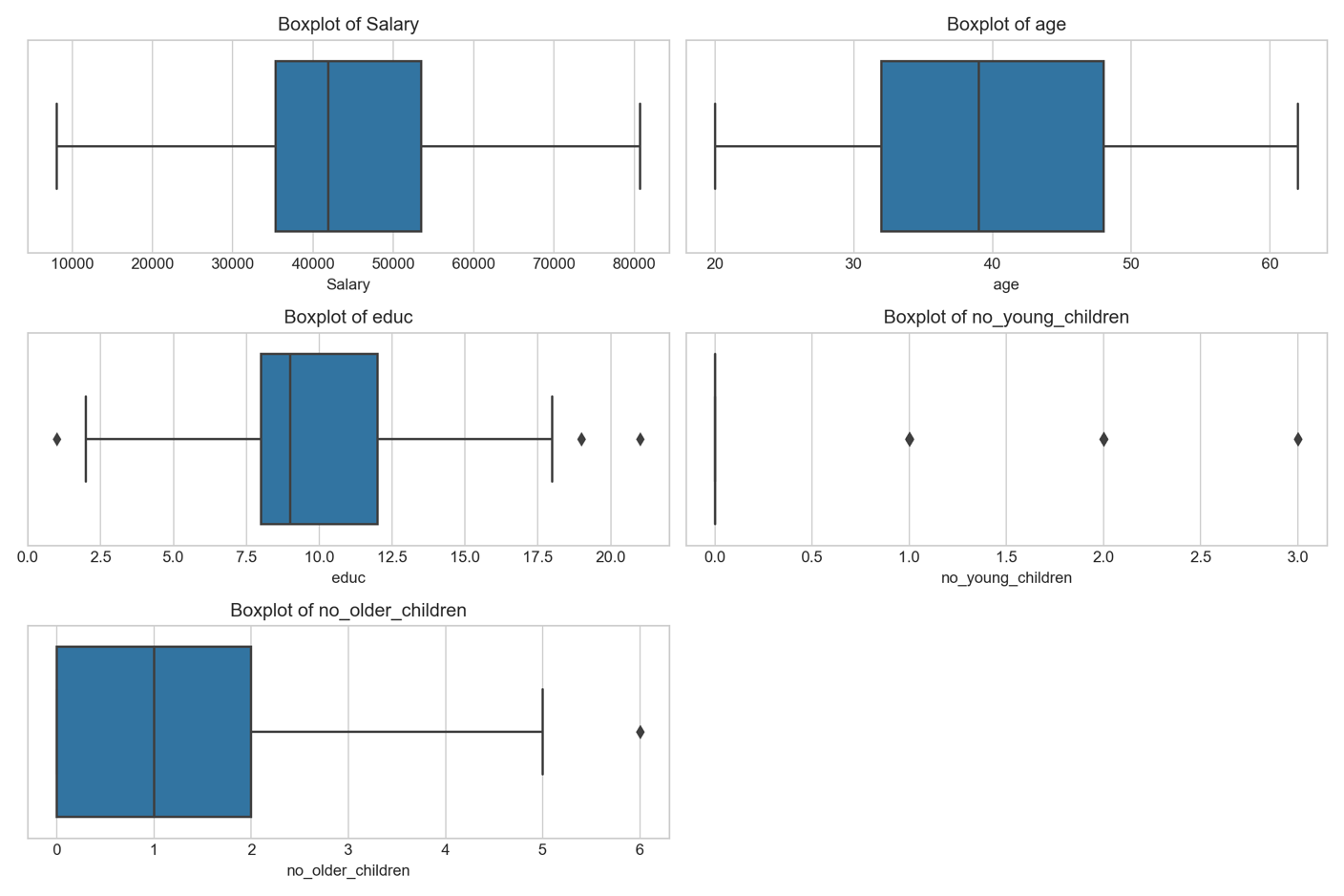


Figure Boxplot outlier treatment Holiday Package

Encode the Data

* The columns ‘Holiday\_Package’ and ’foreigner’ are of type object. We have to convert them into 0’s and 1’s for Modelling. We will be Encoding Categorical data into integer format so that data with converted categorical value can be provided to the model to give and improve predictions.
* ‘Holiday\_Package’ is the categorical target column whereas ‘foreigner’ is the independent categorical column.
* The columns ‘Holiday\_Package’ and ‘foreigner’ are Categorical data. Since there is no implied order within the categorical data type it is not of type Ordinal. Since these categorical variables are binary variables the possible values are not ordered hence we will proceed by replacing yes/no with 0 or 1.

After replacing yes/no with 0 or 1 we get the following dataframe:

Table

Description automatically generated

Figure Data Encoding Holiday Package

Train | Test Split

* We will first create ‘X’ and ‘y’ which will contain Features/predictable variables and Target. ‘X’ will contain Features of the dataset and ‘y’ will contain the Target variable.
* To fairly evaluate our model’s performance, we don’t want to evaluate it on the same data it was trained at. Therefore, we will separate out a training set and a test set. There are four components to this – X\_train, X\_test, y\_train, y\_test. We will use Sklearn library in which we have sub-package model selection.

From model selection we will import ‘train\_test\_split’. We will keep the size of train and test in the ration of 70:30.

Logistic Regression Model

* We now invoke the logistic regression function. Logistic Regression is a machine learning algorithm based on supervised learning. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.
* We create an instance of Logistic Regression model. We have various parameters for the same. These parameters are called the hyperparameters. After creating the model we will train the model on the training dataset using the fit method. We will use this model which has been fit using the training dataset to make predictions of the dataset.
* After creating Logistic Regression Model we are making some adjustments to the parameters in the Logistics Regression class to get a better accuracy.
  + Max\_iter: Maximum number of iterations taken for the solvers to converge.
  + n\_jobs: an integer or None (default) that defines the number of parallel processes to use.
  + penalty: is a string ('l2' by default) that decides whether there is regularization and which approach to use. Other options are 'l1', 'elasticnet', and 'none'.
  + solver: Algorithm to use in the optimization problem. Default is ‘lbfgs’.

**LogisticRegression(max\_iter=10000, n\_jobs=2, penalty='none', solver='newton-cg',**

**verbose=True)**

* After making predictions on the Training Data and Test Data, we will retrieve predicted classes and probabilities.
* Following is the table of predicted classes and probabilities for Test Data.

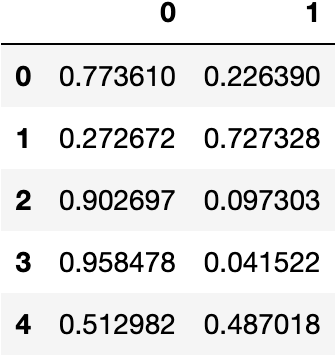


Figure Predicted Classes LR

Logistic Regression Model Using Grid Search

* GridSearchCV allows us to do Grid Search i.e. we can pass multiple input values for all the parameters and find out which combination of value will give the highest precision. After applying Grid Search validation, the best parameters were as below:
  + penalty':['l2','none'],
  + 'solver':['sag', 'lbfgs', 'newton-cg'],
  + 'tol':[0.0001,0.00001]

GridSearchCV(cv=3,

estimator=LogisticRegression(max\_iter=10000, n\_jobs=2, penalty='none', solver='newton-cg', verbose=True),

n\_jobs=-1,

param\_grid={'penalty': ['l2', 'none'],

'solver': ['sag', 'lbfgs', 'newton-cg'],

'tol': [0.0001, 1e-05]},

scoring='f1')

* After making predictions on the Training Data and Test Data, we will retrieve predicted classes and probabilities.
* Following is the table of predicted classes and probabilities for Test Data.

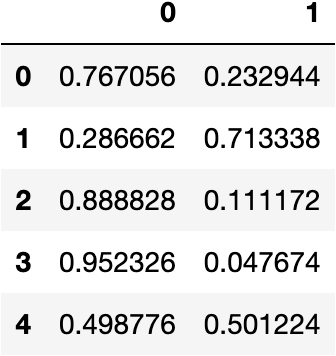


Figure Predicted Classes LR Grid

Linear Discriminant Analysis

* LDA uses linear combinations of independent variables to predict the class in the response variable of a given observation. LDA assumes that the independent variable are normally distributed and there is equal variance / covariance for the classes. LDA is popular, because it can be used for both classification and dimensionality reduction.
* Discriminant Analysis is used for classifying observations to a class or category based on predictor (independent) variables of the data.

Linear Discriminant Analysis model

* We will create a Linear Discriminant analysis function. We generate a model from the same.
* After creating a model, we will use it to predict for whatever data we might chose to put into it.
* After making predictions on the Training Data and Test Data, we will retrieve predicted classes and probabilities.
* Following is the table of predicted classes and probabilities for Test Data.

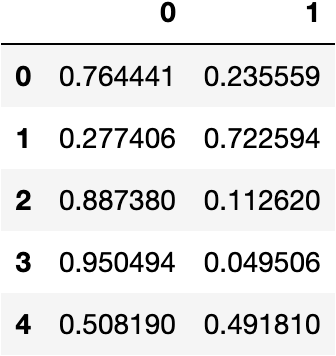


Figure Predicted Classes LDA

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.

Model Evaluation for Logistic Regression Model

**Confusion Matrix**

* A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. It is used to determine the performance of the classification models for a given set of test data.

**Confusion Matrix – Training Data**

**Chart

Description automatically generated**

Figure Confusion Matrix Training Data LR

**Confusion Matrix – Testing Data**

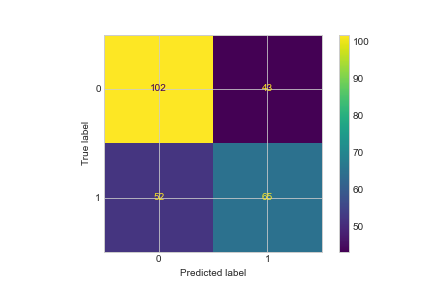


Figure Confusion Matrix Test Data LR

* In our dataset for the target variable 'Holiday\_Package', 0 means the employee has not opted and 1 means the employee has opted for a holiday package.
* if we study the confusion matrix, we will find that False Negative i.e. FN is the most important for us because it signifies the condition when the employee did not opt for a holiday package but the model predicted that the employee will opt for the same. It will be a more costlier mistake for the agency. This is the type 2 error.
* the second most important metric for us is True Positive i.e. TP because it signifies the condition when the Employee did opt for holiday package and the model also predicted the same.
* From the above table we can see that for training data FN = 124 and TP = 160 and testing data FN = 52 and TP = 65.
* The performance metric which considers FN and TP as the measuring parameter is Recall or Sensitivity. FN is the Type 2 error and for Type 2 error we consider Recall or Sensitivity as the optimised performance measure for our problem statement.

**Classification Report**

* A classification report is a performance evaluation metric in machine learning. It is used to show the precision, recall, F1 Score, and support of your trained classification model.

**Classification Report – Training Data**

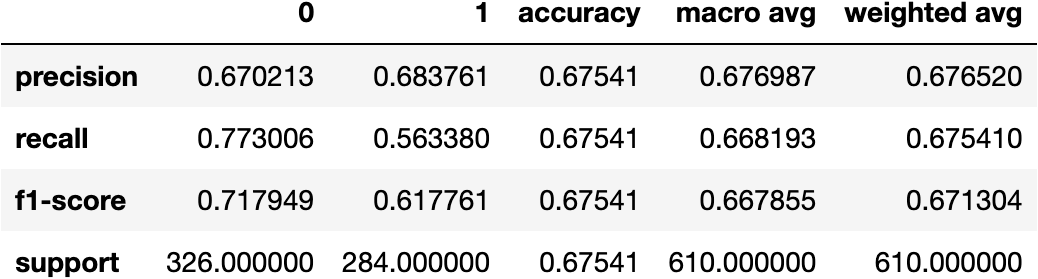
****

Figure Classification Report Training Data LR

**Classification Report – Testing Data**

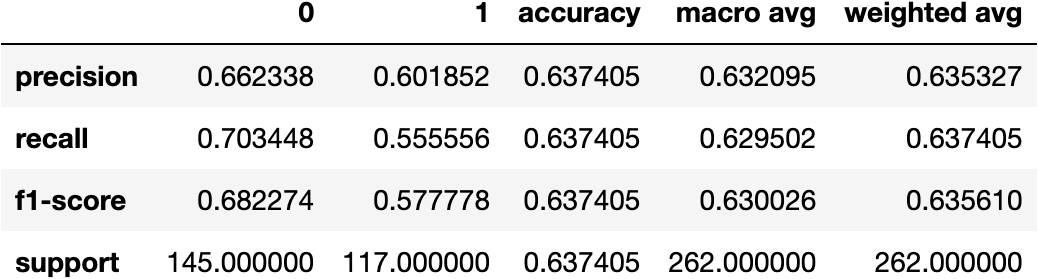


Figure Classification Report Test Data LR

* Accuracy of the model is the correct predictions divided by total number of observations. It suggests how accurately does the model classify the data points. Accuracy for training data is 67.5% and Accuracy for testing data is 63.7%. Higher the Accuracy of the model, stronger the prediction lower the accuracy weaker the prediction.
* The Recall of the model tells us how many of the true data points are identified as the true data points by the model. It describes the Type 2 error. Our variable of interest is 1 i.e. employees who will opt for holiday package. Recall for training data is 56.3% and Recall for testing data is 55.55%.
* The precision of the model tell us among the data points identified as Positive by the model, how many are really positive. It describes the Type 1 error. precision for training data is 68.3% and precision for testing data is 60.1%.

**AUC and ROC Curve**

* ROC curve is the technique for visualising the output of classification models to find out how good the performance is. It is a graph showing the performance of a classification model at all classification thresholds.

**AUC and ROC Curve – Training Data**

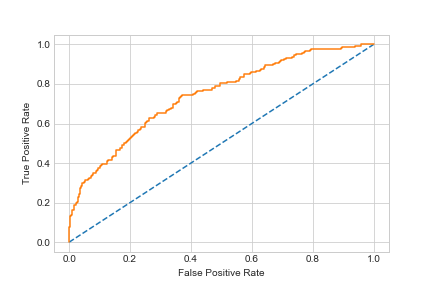
****

Figure auc\_roc Curve training data LR

**AUC and ROC Curve – Testing Data**

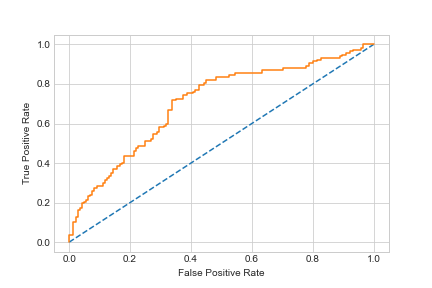


Figure auc\_roc curve test data LR

* It is the graph that is calculated between the true positive rate and false positive rate in the confusion matrix.
* True positive rate = TP/total positive - y axis
* False positive rate = FP/total negative - x axis
* ROC graph is drawn to show a trade-off between the benefits (TP) and costs (FP)
* The above graph is constructed between 0 and 1 in x axis and 0 and 1 in y axis. This represents the percentage or the probabilities.
* The x axis is the negative scenario and y axis is the positive scenario. We expect a hight true positive rate and low false positive rate.
* For the testing data ROC graph if we take 0.2 as the cut-off to separate or classify the positives and the negatives the corresponding true positive rate is around 0.45.
* For the training data ROC graph if we take 0.2 as the cut-off to separate or classify the positives and the negatives the corresponding true positive rate is around 0.55.
* We know that the steeper the ROC curve, the stronger the model and the flatter the ROC curve , the weaker the model.
* AUC means Area under the ROC Curve. Larger the AUC, better the model because the more steeper it will be. The AUC for training data is 74.2% and for testing data it is 70.5%.

Model Evaluation for Logistic Regression Model using Grid Search

**Confusion Matrix**

* A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. It is used to determine the performance of the classification models for a given set of test data.

**Confusion Matrix – Training Data**

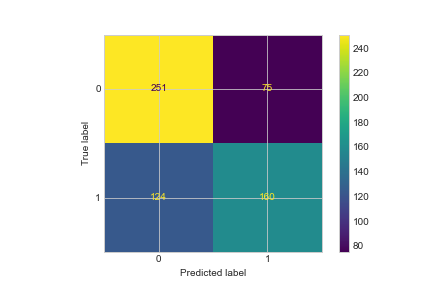
****

Figure Confusion Matrix Training Data LR Grid

**Confusion Matrix – Testing Data**

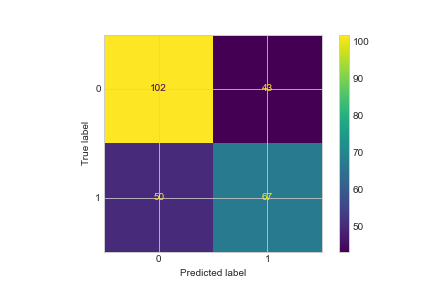
****

Figure Confusion Matrix Test Data LR Grid

* In our dataset for the target variable 'Holiday\_Package', 0 means the employee has not opted and 1 means the employee has opted for a holiday package.
* if we study the confusion matrix, we will find that False Negative i.e. FN is the most important for us because it signifies the condition when the employee did not opt for a holiday package but the model predicted that the employee will opt for the same. It will be a more costlier mistake for the agency. This is the type 2 error.
* the second most important metric for us is True Positive i.e. TP because it signifies the condition when the Employee did opt for holiday package and the model also predicted the same.
* From the above table we can see that for training data FN = 124 and TP = 160 and testing data FN= 50 and TP = 67.
* The performance metric which considers FN and TP as the measuring parameter is Recall or Sensitivity. FN is the Type 2 error and for Type 2 error we consider Recall or Sensitivity as the optimised performance measure for our problem statement.

**Classification Report**

* A classification report is a performance evaluation metric in machine learning. It is used to show the precision, recall, F1 Score, and support of your trained classification model.

**Classification Report – Training Data**

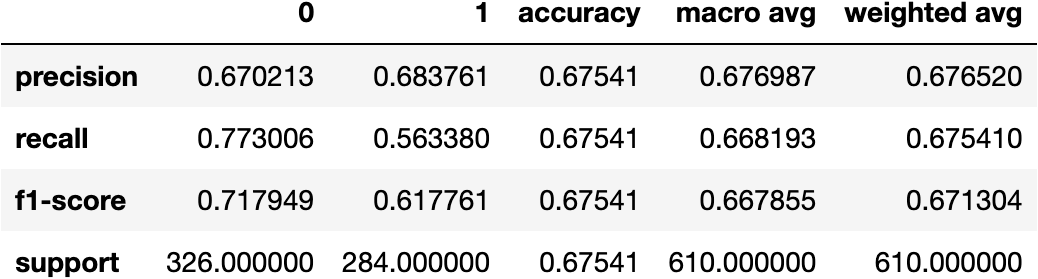
****

Figure Classification Report Train data LR Grid

**Classification Report – Testing Data**

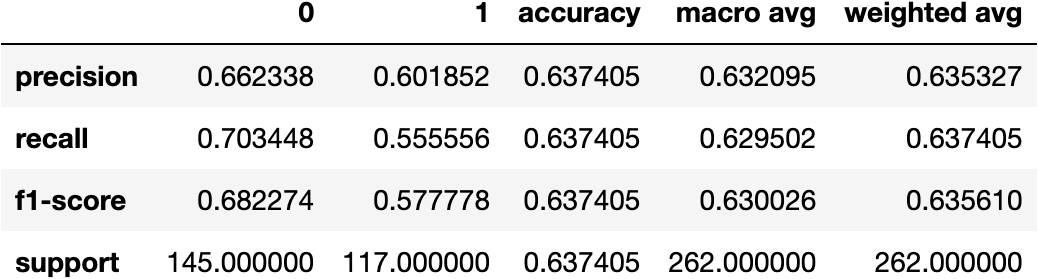


Figure Classification Report Test data LR Grid

* Accuracy of the model is the correct predictions divided by total number of observations. It suggests how accurately does the model classify the data points. Accuracy for training data is 67.5% and Accuracy for testing data is 63.7%. Higher the Accuracy of the model, stronger the prediction lower the accuracy weaker the prediction.
* The Recall of the model tells us how many of the true data points are identified as the true data points by the model. It describes the Type 2 error. Our variable of interest is 1 i.e. employees who will opt for holiday package. Recall for training data is 56.3% and Recall for testing data is 55.55%.
* The precision of the model tell us among the data points identified as Positive by the model, how many are really positive. It describes the Type 1 error. precision for training data is 68.3% and precision for testing data is 60.1%.

**AUC and ROC Curve – Training Data**

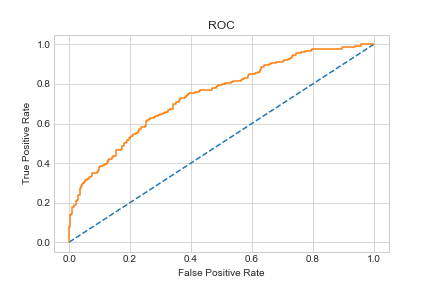


Figure auc\_roc curve Train Data LR Grid

**AUC and ROC Curve – Testing Data**

Chart, line chart

Description automatically generated

Figure auc\_roc curve Test Data LR Grid

* It is the graph that is calculated between the true positive rate and false positive rate in the confusion matrix.
* True positive rate = TP/total positive - y axis
* False positive rate = FP/total negative - x axis
* ROC graph is drawn to show a trade-off between the benefits (TP) and costs (FP)
* The above graph is constructed between 0 and 1 in x axis and 0 and 1 in y axis. This represents the percentage or the probabilities.
* The x axis is the negative scenario and y axis is the positive scenario. We expect a hight true positive rate and low false positive rate.
* For the testing data ROC graph if we take 0.2 as the cut-off to separate or classify the positives and the negatives the corresponding true positive rate is around 0.45.
* For the training data ROC graph if we take 0.2 as the cut-off to separate or classify the positives and the negatives the corresponding true positive rate is around 0.55.
* We know that the steeper the ROC curve, the stronger the model and the flatter the ROC curve , the weaker the model.
* AUC means Area under the ROC Curve. Larger the AUC, better the model because the more steeper it will be. The AUC for training data is 74.2% and for testing data it is 70.5%.

Model Evaluation for Linear Discriminant Analysis

**Confusion Matrix**

* A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. It is used to determine the performance of the classification models for a given set of test data.

**Confusion Matrix – Training Data**

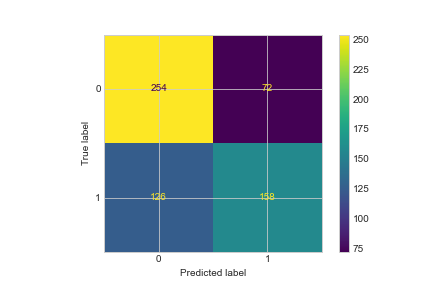


Figure Confusion Matrix Train data LDA

**Confusion Matrix – Testing Data**

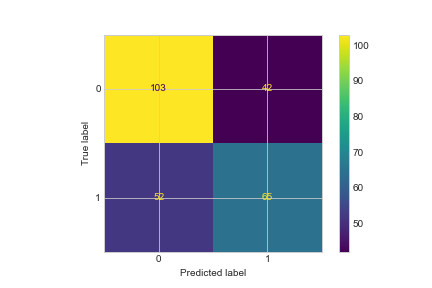


Figure Confusion Matrix Test Data LDA

* In our dataset for the target variable 'Holiday\_Package', 0 means the employee has not opted and 1 means the employee has opted for a holiday package.
* if we study the confusion matrix, we will find that False Negative i.e. FN is the most important for us because it signifies the condition when the employee did not opt for a holiday package but the model predicted that the employee will opt for holiday package. It will be a more costlier mistake for the agency as they will spend resources towards the Employee for whom the model predicted that employee will opt but after spending all the resources the employee did not opt and hence a costlier mistake. This is the type 2 error.
* the second most important metric for us is True Positive i.e. TP because it signifies the condition when the Employee did opt for holiday package and the model also predicted the same.
* From the above table we can see that for training data FN = 126 and TP = 158 and testing data FN= 52 and TP = 65.
* The performance metric which considers FN and TP as the measuring parameter is Recall. FN is the Type 2 error and for Type 2 error we consider Recall or Sensitivity as the optimised performance measure for our problem statement.

**Classification Report**

* A classification report is a performance evaluation metric in machine learning. It is used to show the precision, recall, F1 Score, and support of your trained classification model.

**Classification Report – Training Data**

Table

Description automatically generated

Figure Class Report Train Data LDA

**Classification Report – Testing Data**

Table

Description automatically generated

Figure Class Report Test Data LDA

* Accuracy of the model is the correct predictions divided by total number of observations. It suggests how accurately does the model classify the data points. Accuracy for training data is 67.5% and Accuracy for testing data is 64.1%. Higher the Accuracy of the model, stronger the prediction lower the accuracy weaker the prediction.
* The Recall of the model tells us how many of the true data points are identified as the true data points by the model. It describes the Type 2 error. Our variable of interest is 1 i.e. employees who will opt for holiday package. Recall for training data is 55.6% and Recall for testing data is 55.55%.
* The precision of the model tell us among the data points identified as Positive by the model, how many are really positive. It describes the Type 1 error. precision for training data is 68.69% and precision for testing data is 60.74%.

**AUC and ROC Curve – Training Data**

**Chart, line chart

Description automatically generated**

Figure auc\_roc curve Train Data LDA

**AUC and ROC Curve – Testing Data**

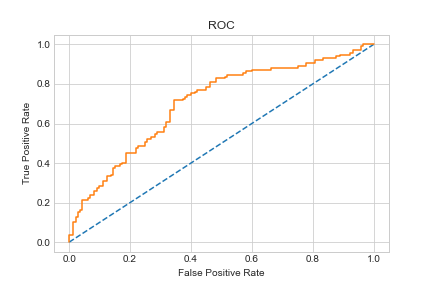


Figure auc\_roc curve Test Data LDA

* It is the graph that is calculated between the true positive rate and false positive rate in the confusion matrix.
* True positive rate = TP/total positive - y axis
* False positive rate = FP/total negative - x axis
* ROC graph is drawn to show a trade-off between the benefits (TP) and costs (FP)
* The above graph is constructed between 0 and 1 in x axis and 0 and 1 in y axis. This represents the percentage or the probabilities.
* The x axis is the negative scenario and y axis is the positive scenario. We expect a hight true positive rate and low false positive rate.
* For the testing data ROC graph if we take 0.2 as the cut-off to separate or classify the positives and the negatives the corresponding true positive rate is around 0.45.
* For the training data ROC graph if we take 0.2 as the cut-off to separate or classify the positives and the negatives the corresponding true positive rate is around 0.55.
* We know that the steeper the ROC curve, the stronger the model and the flatter the ROC curve , the weaker the model.
* AUC means Area under the ROC Curve. Larger the AUC, better the model because the more steeper it will be. The AUC for training data is 73.9% and for testing data it is 70.3%.

Comparison between the Models

* On comparing the models Logistic Regression and Linear Discriminant Analysis we find that the performance of Train and Test data is very similar.
* As we can see the results of Confusion Matrix, ROC\_AUC score, ROC\_AUC curve for both the models is equivalent with each other. Therefore it is very difficult to differentiate between the two on the basis of their performance. However, we also know that LDA comes to our rescue in situations where logistic regression is unstable when
  + classes are well separated- Logistic Regression lacks stability when the classes are well- separated. That’s when LDA comes in.
  + the data is small
  + we have more than two classes- LDA is a better choice whenever multi-class classification is required and in the case of binary classifications, both logistic regression and LDA are applied.
* In our dataset we know that classes are well separated and the data given to us is also small (7 columns and 872 rows)
* Therefore we can infer that LDA will be a better model for our dataset.

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

* From our Analysis we found out that foreigner and native employees aged above 45 are not interested in holiday packages. Also foreigner and native employees aged below 30 has also not shown much interest in opting for holiday packages. So this shows that age is one of the factors in determining whether an employee will opt for a holiday package or not. Employees age between 30 and 50 do buy holiday packages whether they are foreigners or natives. To get the attention of employees above 45 we can them their field of interest and try to include that in the holiday package or should organise their holiday around their interest. Same should be done for employees having age less than 30 years.
* Salary is also one of the important factors in determining whether an employee will opt for holiday package or not. Foreigner or native employees having salary in the range of 30000 – 50000 have opted for holiday packages. To attract employees having salary higher than 150000 we should include beautiful couple destinations as employees with mostly higher age will be having salary around the same. Higher age indicates that mostly employees will be married and therefore we can include couple destinations like Europe etc. We can also include luxurious stays for the same. We can have longer stays for the holiday package. Also, Employees having lower salary should be given discount so that they can also opt for holiday packages.
* For the employees which have children older or younger we should have family holiday packages for them.
* There is a significant difference in the number of foreigners who has opted for holiday packages as compared to those who did not opt for the same. Perhaps we should include holiday packages which is packed with cultural events and historical information about the place so that foreigners can learn as well as enjoy after opting for the same.
* If employee is foreigner or native and employee is not having young children, chances of opting for Holiday Package is good. Special offer can be designed to domestic employees to opt for Holiday Package. For employees having young children packages can be modified to make infant and young children friendly.