|  |
| --- |
| July 2021 | PG-DSBA-Online |



**PROJECT REPORT**

**PREDICTIVE MODELLING**

**SUBMITTED BY**

**DEV TRIPATHI**

**PROJECT REPORT ON**

**ADVANCE STATISTICS**

Submitted by

Dev Tripathi

**Contents**

[Problem 1 1](#_Toc78715873)

[Cubic Zirconia Dataset Analysis 1](#_Toc78715874)

[Information about dataset 1](#_Toc78715875)

[Data Dictionary for Cubic Zirconia Dataset 1](#_Toc78715876)

[EDA 2](#_Toc78715877)

[Linear Regression Model 9](#_Toc78715878)

[Inferences 11](#_Toc78715879)

[Problem 2 12](#_Toc78715880)

[Holiday Package Dataset Analysis 12](#_Toc78715881)

[Information about the Dataset 12](#_Toc78715882)

[Data Dictionary 12](#_Toc78715883)

[EDA 12](#_Toc78715884)

[Logit model & LDA 17](#_Toc78715885)

[Performance Metrics 17](#_Toc78715886)

[Inferences 19](#_Toc78715887)

**List of Figures**

[Figure 1: Sample of Cubic Zirconia Dataset 2](#_Toc78715606)

[Figure 2: Feature Info and null values count of Cubic Zirconia Dataset 2](#_Toc78715607)

[Figure 3: Univariate Curves for different features in the dataset 5](#_Toc78715608)

[Figure 4: Pairplot for Cubic Zirconia Dataset 6](#_Toc78715609)

[Figure 5: Correlation plot for different features 7](#_Toc78715610)

[Figure 6: Description of the dataset 8](#_Toc78715611)

[Figure 7: Labels of different categories present in features 9](#_Toc78715612)

[Figure 8: Coefficient values for features present in the dataset 10](#_Toc78715613)

[Figure 9: VIF values for features present in the dataset 10](#_Toc78715614)

[Figure 10: Sample of Holiday package dataset 13](#_Toc78715615)

[Figure 11: Feature info and null count of the dataset 13](#_Toc78715616)

[Figure 12: Value counts for each attribute in the dataset 14](#_Toc78715617)

[Figure 13: statistical Description of all the features 14](#_Toc78715618)

[Figure 14: Univariate Analysis for continuous variables 15](#_Toc78715619)

[Figure 15: Count plots for object/categorical type attributes 16](#_Toc78715620)

[Figure 17: Performance metrics using Logit model for train set (left) and test set (Right) 18](#_Toc78715621)

[Figure 18: Performance Metrics for LDA on Train set (left) and Test set (right) 18](#_Toc78715622)

[Figure 19: Coefficient for the variable present in the dataset for Logit model(left) and LDA (right) 19](#_Toc78715623)

**List of Tables**

[Table 1: Data Dictionary (Cubic Zirconia Dataset) 1](#_Toc78715888)

[Table 2: Performance Metrics (Linear Regression model) 9](#_Toc78715889)

[Table 3: Performance metrics for unscaled data 10](#_Toc78715890)

[Table 4: Data Dictionary for Holiday Package Dataset 12](#_Toc78715891)

## Problem 1

### **Cubic Zirconia Dataset Analysis**

### Information about dataset

The dataset is provided by a company named Gem Stones co ltd, which is a cubic zirconia manufacturer. This dataset contains 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).

### Data Dictionary for Cubic Zirconia Dataset

The dataset contains the following features and the information about these features is mentioned below:

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| Carat | Carat weight of the cubic zirconia. |
| Cut | Describes the cut quality of the cubic zirconia. Quality is increasing order Fair, Good, Very Good, Premium, Ideal. |
| Color | Color 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, FL = flawless, l1= level 1 inclusion) IF, VVS1, VVS2, VS1, VS2, Sl1, Sl2, l1 |
| Depth | The Height of cubic zirconia, measured from the Culet to the table, is divided by its average Girdle Diameter. |
| Table | The Width of the cubic zirconia's Table  expressed as a Percentage of its Average  Diameter. |
| Price | Price of the zirconia |
| X | Length of cubic Zirconia in mm. |
| Y | Width of cubic Zirconia in mm. |
| Z | Height of cubic Zirconia in mm. |

Table : Data Dictionary (Cubic Zirconia Dataset)

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

### EDA

#### Sample of Dataset-

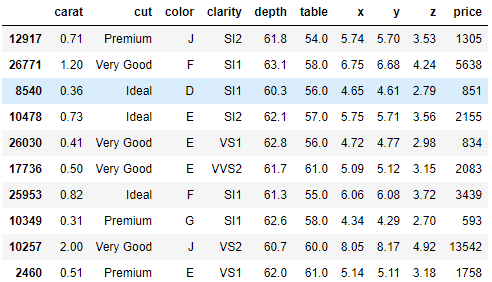


Figure 1: Sample of Cubic Zirconia Dataset

#### Variable Information-

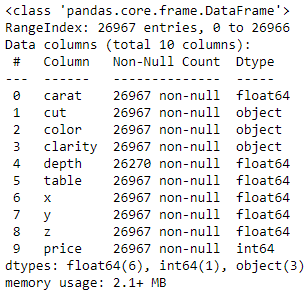
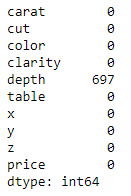
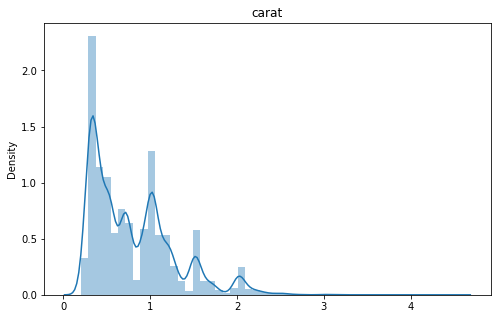
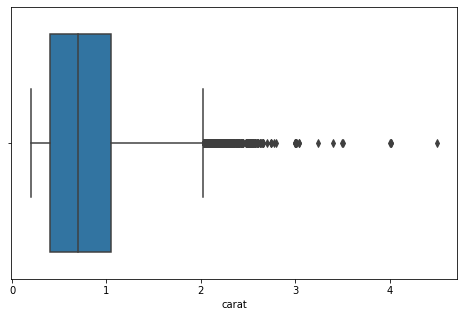
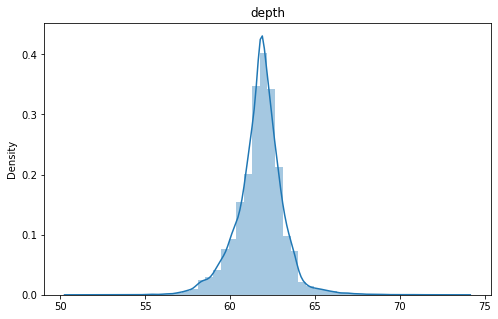
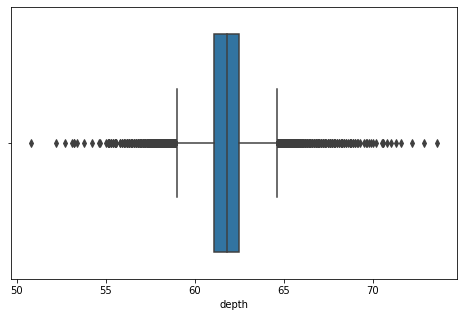
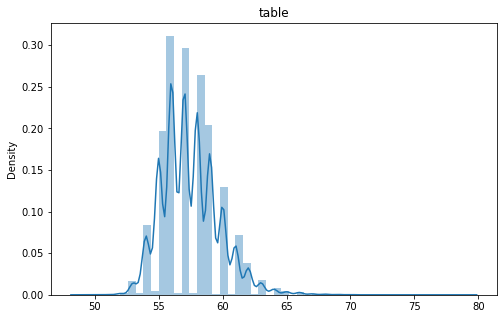
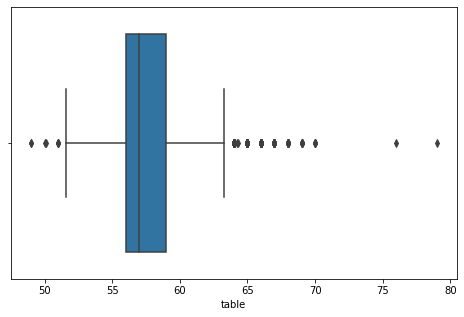
 

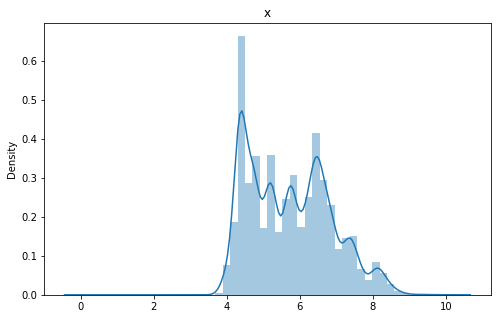
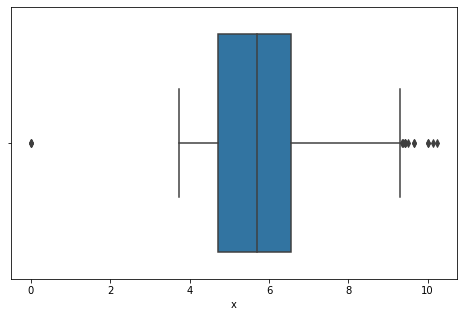
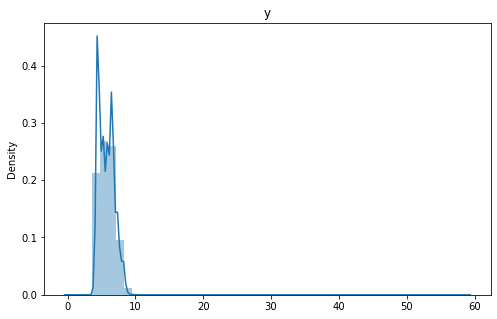
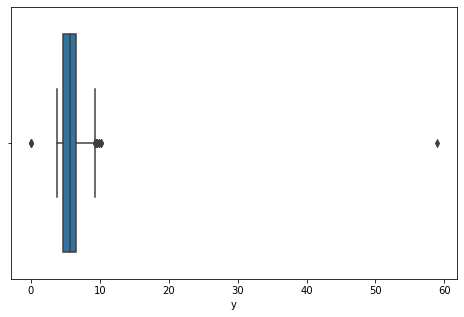
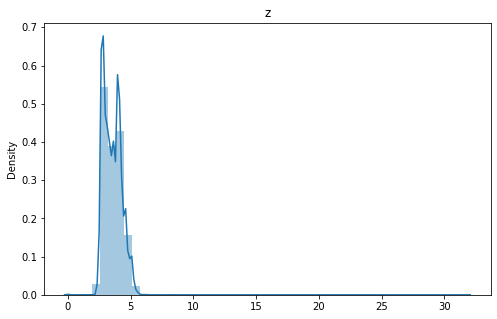
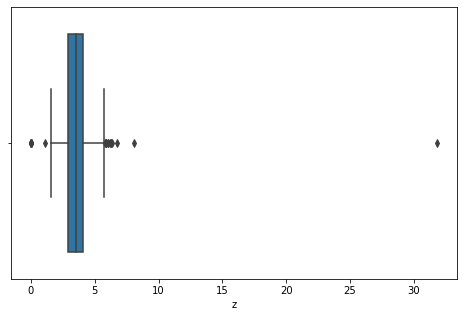
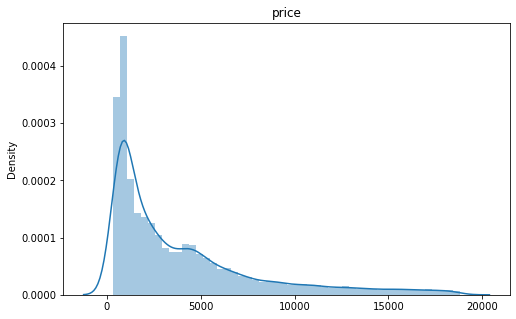
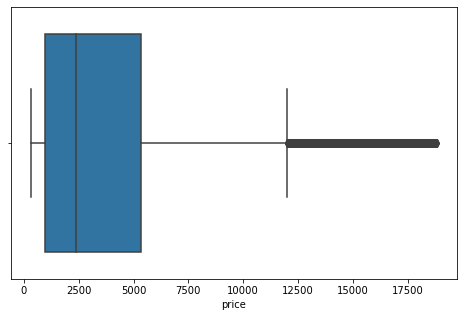
Figure 2: Feature Info and null values count of Cubic Zirconia Dataset

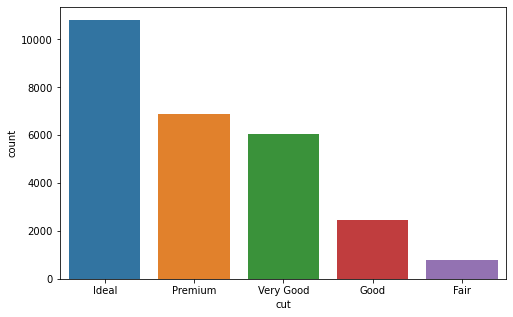
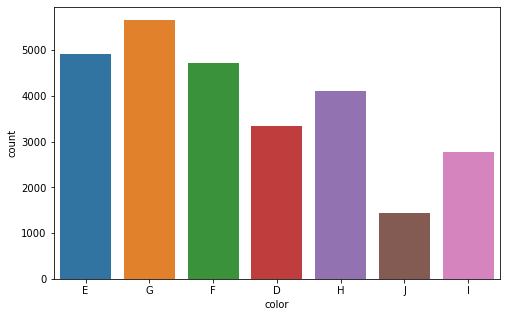
The dataset contains a total of 26967 entries and 10 features including our target variable which is **‘price’**. The **‘depth’** feature contains a total of 697 null values. None of the variables contains null values other than the **‘depth’** variable. There are three columns i.e., ’cut’, ’color’, and ‘clarity’ which are object data types. All other variables are of numeric data type.

#### Univariate Analysis-

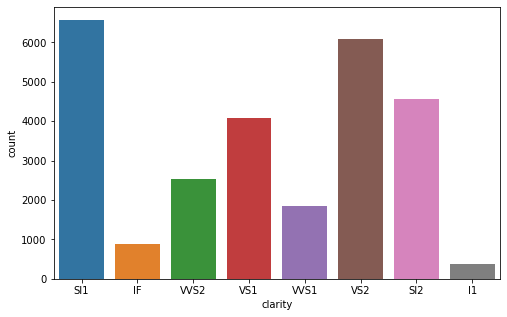


Figure 3: Univariate Curves for different features in the dataset

Now, conclusions drawn from this analysis can be briefly explained in the following points:

1. From the boxplots, it can be observed that most of the variables are having outliers
2. Though the X, Y, and Z variables have some very extreme outliers which can be problematic for further analysis.
3. All the variables are having Right skewed distributions except the ‘depth’ variable which seems to be having very little skewness.
4. Right skewed distributions show that zirconia having good values for attributes (e.g., Carat, length, width, etc.) are rare.
5. The count plot of cut shows that most of the product has Ideal or premium cut which is good.
6. The clarity of a very few zirconias lies in Fl(flawless) category. The most frequent clarity observed is the Sl1 type.

#### Bi-Variate Analysis-

Now for bivariate analysis, we have created a pair plot that simply shows a scatterplot for all the pairs of features present in the dataset.

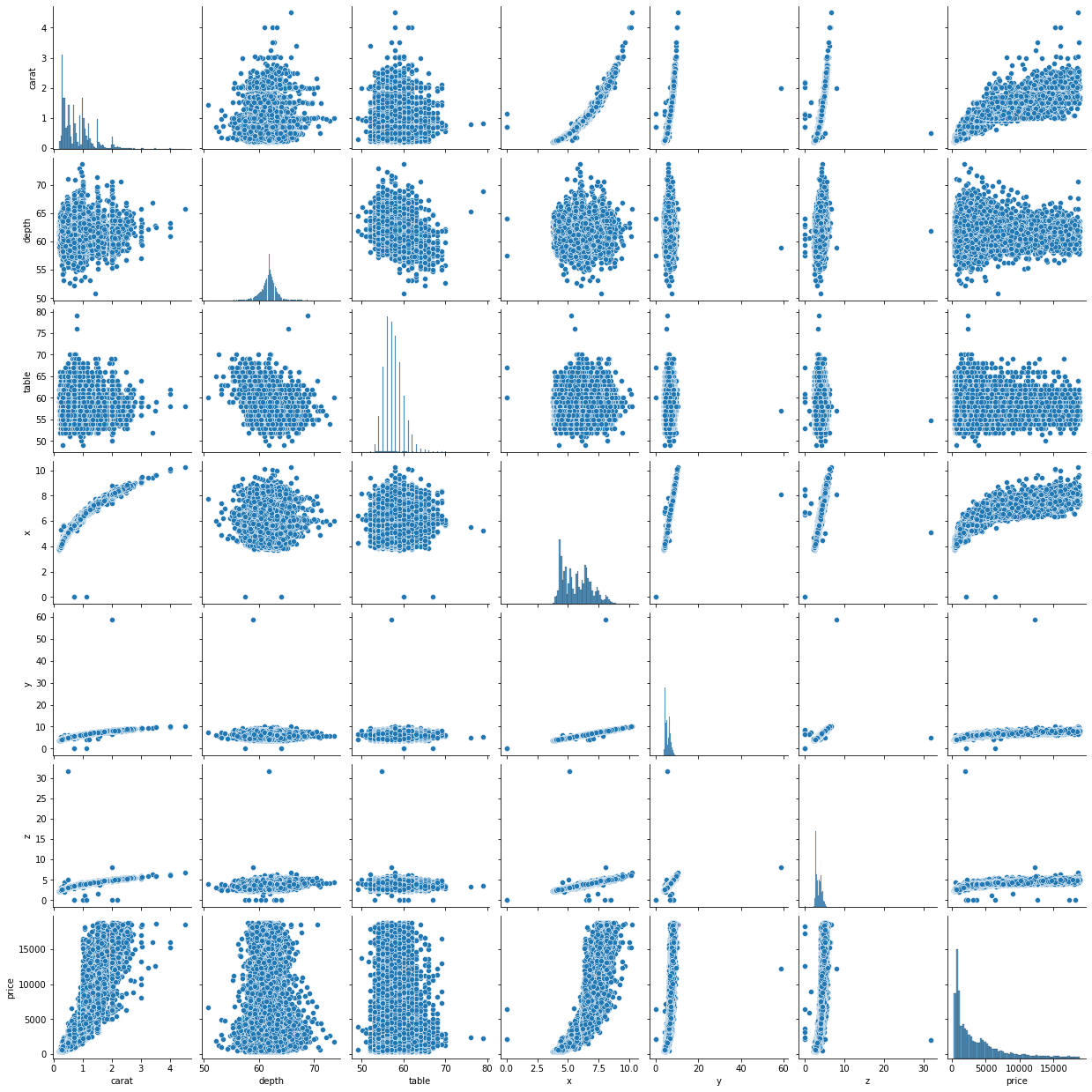


Figure 4: Pairplot for Cubic Zirconia Dataset



Figure 5: Correlation plot for different features

For the correlation plot encoding of categorical features is necessary. After encoding the categories in numeric form (please refer to the notebook for the labels/steps involved in encoding), the correlation plot was created. In the encoding process higher number was chosen to represent better attributes for example Fl(flawless) clarity was given 7 as the label.

Now, one can observe that the ‘carat’, ‘x’, ‘y’, ‘z’ and ‘price’ features have a high correlation between them (Which could be problematic for the regression model). The ‘clarity’ feature shows a negative correlation with x, y, z, and carat features and the ‘cut’ variable shows a negative correlation with the ‘table’ feature. The rest of the features have low to minimal correlation with any other variable. These correlations can be observed in pairplot also.

#### 1.2 Impute null values if present, also checks for the values which are equal to zero. Do they have any meaning or do we need to change them or drop them? Do you think scaling is necessary in this case?

**Sol:** As mentioned earlier, only the ‘depth’ variable has null values (null count = 697). Now, the imputation of these null values can be done in several ways like mean, median, mode, or model-based imputation. Here, **median imputation** has been done. Though the value of mean and median for the depth feature is almost the same (mean= 61.745147 and median = 61.8).

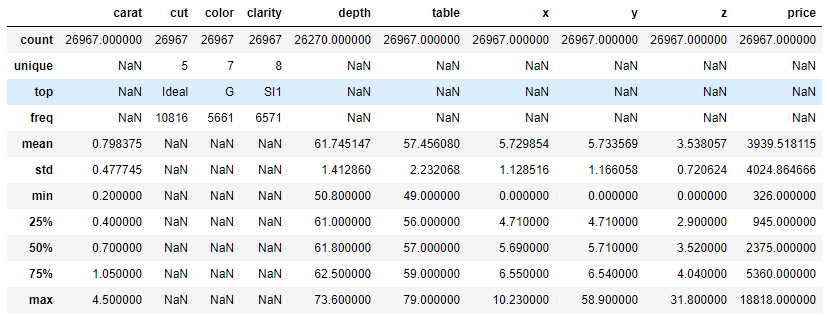


Figure 6: Description of the dataset

Now, from the description of the data, one can observe that only ‘x’, ‘y’, and ‘z’ columns have minimum values 0 (which was earlier shown in the respective boxplots). Since these are physical dimensions of zirconia the 0 value seems absurd. Though, only 11 observations were found to have the x y or z value as 0. Hence, considering the number of observations available to us which is around 27k, these samples were **dropped for further analysis**.

The scaling of the dataset becomes necessary for any model if the range of values for different features is very different. For regression models scaling is necessary for interpreting the feature importance and also preventing the model from being biased. Now, in our case, the ‘price’ ranges from around 4k to 18k whereas the values for ‘carat’ ranges from 0.2 to 4.5. Hence, due to this huge difference in ranges, scaling of the features was done before proceeding to model and prediction steps.

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

Sol: As mentioned earlier the encoding was done such that the better attributes are labeled with higher values. The codes for the same are given below:

Figure 7: Labels of different categories present in features

Now, after encoding and scaling, the dataset was split into train and test datasets for evaluation of the model. As per the instructions, the 70,30 split was performed (70 percent samples for train and 30 percent samples for test set).

### Linear Regression Model

After the train and split, a Linear regression model was fitted to the training data The actual model obtained from the analysis can be summarized by the following equation-

(-7598.6) \* Intercept + (10875.33) \* carat + (140.45) \* cut + (333.46) \* color + (501.04) \* clarity + (96.36) \* depth + (-19.08) \* table + (-1754.0) \* x + (2594.81) \* y + (-2794.93) \* z = Price of zirconia

What it represents is that, if the carat value is to be increased by a single unit (1 unit) the resulting increase due to this price change is around 11k units (keeping other variables constant). Performance metrics were calculated for the train and test set which are following:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **R^2** | **MAE** | **MSE** | **RMSE** |
| **Train set** | 0.907 | 0.203 | 0.092 | 0.304 |
| **Test set** | 0.913 | 0.201 | 0.088 | 0.296 |

Table : Performance Metrics (Linear Regression model)

From the metrics obtained from the model, one can say that the model is working fine for both the train and test set. Also, the model performance is comparable for both sets considering the above-mentioned metrics. Though one must keep in mind that these values are for scaled data, not the actual data. For unscaled data the metrics are given below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **R^2** | **MAE** | **MSE** | **RMSE** |
| Train set | 0.907 | 814.905 | 1493407 | 1222.051 |
| Test set | 0.913 | 809.19 | 1417539 | 1190.604 |

Table : Performance metrics for unscaled data

Hence, from the above RMSE values we can say that, if this model is used for predicting the price of zirconia, the actual and predicted value can differ by around 1200 units.

Now, if we look at the coefficients (for scaled dataset) of different features, we can say that ‘carat’, ‘x’, ‘y’, ‘z’ (means the dimensions of zirconia) and ‘clarity’ are better predictors for the price compared to other variables. But one must keep in mind that multicollinearity hasn’t been considered in the analysis.

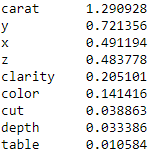


Figure : Coefficient values for features present in the dataset

Though in the analysis here, we have ignored the VIF (variance inflation factor) calculation for addressing the multicollinearity problem. We can look at the VIF values to say which features are redundant for the prediction.

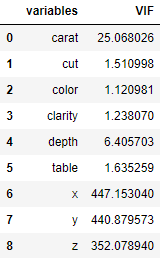


Figure : VIF values for features present in the dataset

As it is visible from the values that the x,y,z, and carat have very high VIF values, which means one can eliminate these features and keep only the best ones (predictors) for the model. Though in this analysis, the variables were simply assumed to be independent.

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

Sol:

### Inferences

The linear regression model is one of the simplest yet powerful models available. The inferences drawn from the above analysis can be discussed in the following points:

1. The price of cubic zirconia is highly affected by its dimension and carat value.
2. The dimensions correlate with each other, we can be interpreted as most of the crystals mentioned in the dataset are not flaky.
3. For further optimization, one can remove the features with high values of VIF (>10) and try to get better results.
4. Multicollinearity is present in the dataset though for this analysis it was not considered to be problematic.
5. Error distribution obtained from the analysis was found to be slightly skewed to left.
6. Overall performance of the model was found to be good in this analysis, though this model is not recommended to be used for real-world cases because multicollinearity is present which can make the model unstable.

## Problem 2

### **Holiday Package Dataset Analysis**

### Information about the Dataset

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

### Data Dictionary

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

Table : Data Dictionary for Holiday Package Dataset

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

Sol:

### EDA

#### Sample of dataset

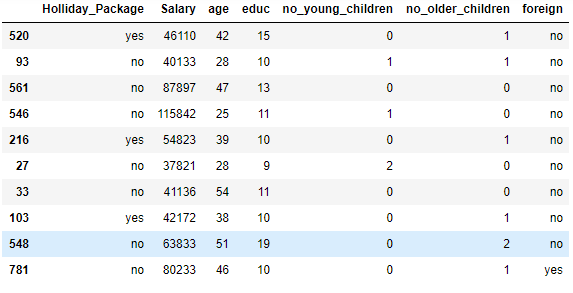


Figure 10: Sample of Holiday package dataset

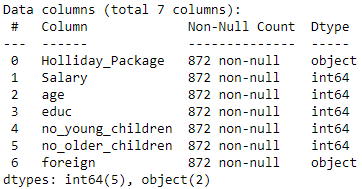
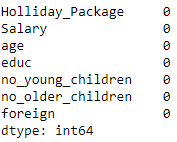
 

Figure 11: Feature info and null count of the dataset

As we can observe that the dataset has a total of 872 entries and no null values. There is a total of 6 variables including our target variable which is ‘Holiday\_package’. The individual value counts for these object type variables are following:

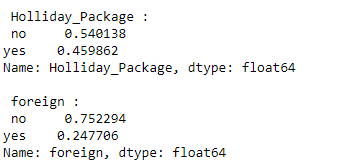


Figure 12: Value counts for each attribute in the dataset

As we can see that there are 54 percent ‘no’ values and 46 percent ‘yes’ values. So, we can say that the target variable is fairly balanced.

#### Descriptive Statistics

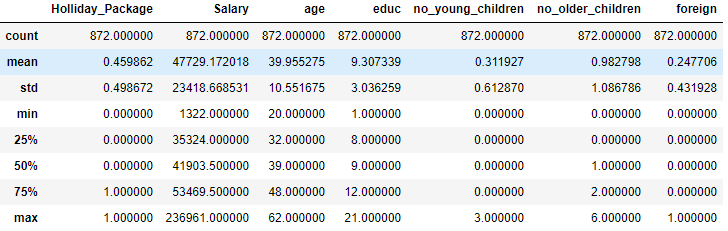
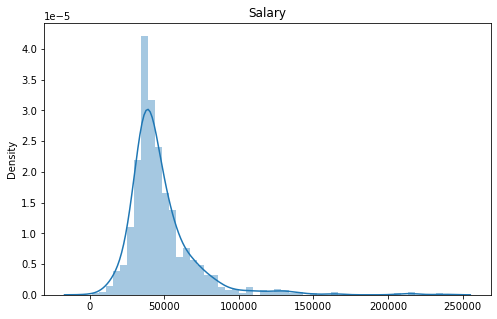
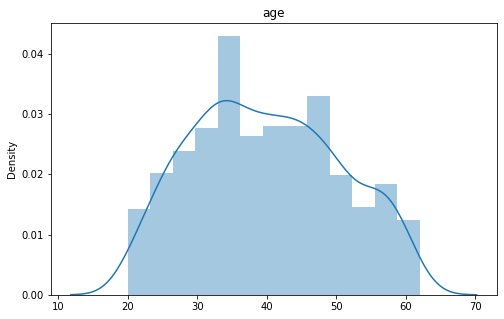
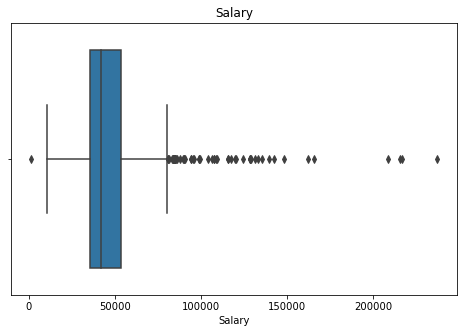
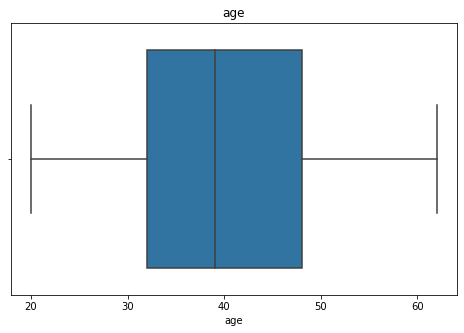


Figure 13: statistical Description of all the features

#### Univariate Analysis

For univariate analysis we can look at distribution plot and boxplots of the attributes present in the dataset:

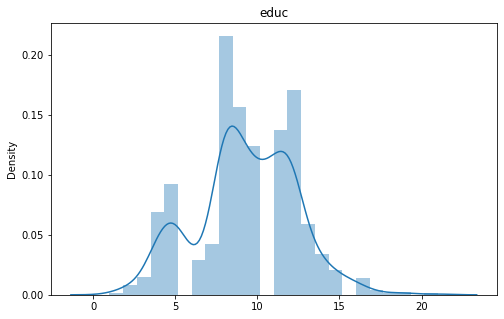
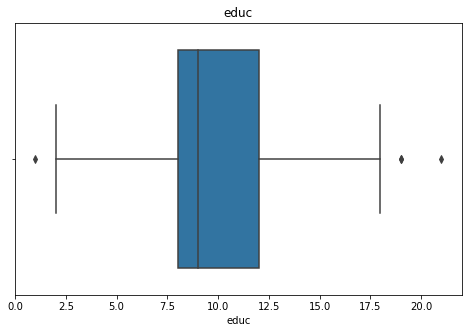
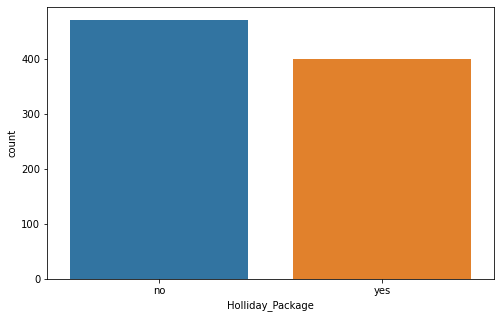
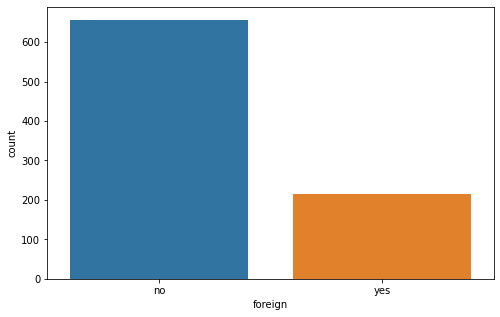
 

Figure 14: Univariate Analysis for continuous variables

**Observations-**

* The education variable has a few outliers. But these outliers are genuine outliers. Also, the frequency of the curve is highest between 8 to 12. This means a majority of the employees have 8 to 12 years of formal education.
* Age variable have no outliers. From the distribution curve, we can see that majority of the people who are an employee in the company are between the age group of 30 to 50.
* The salary variable has a lot of outliers. The distribution was found to be right-skewed.

Now, let’s look at the count plot for each object type variable present in the dataset.**** ****

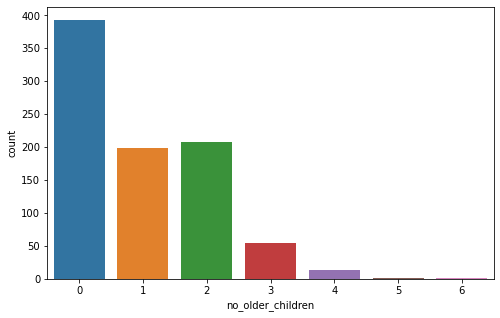
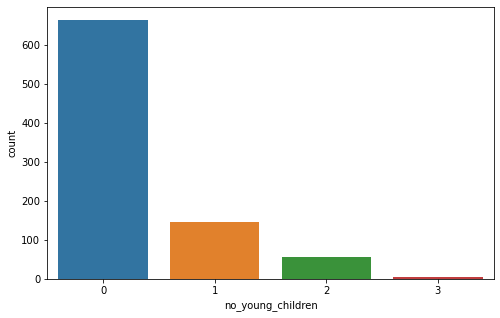
 

Figure 15: Count plots for object/categorical type attributes

**Observations:**

* The majority of the employees are natives. Only 24 percent of employees are foreigners.
* Less than 100 employees have 3 or more older children.
* Half of the number of employees have no children.
* The target variable i.e., Holiday\_package is fairly balanced as found earlier from value counts.

#### 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).

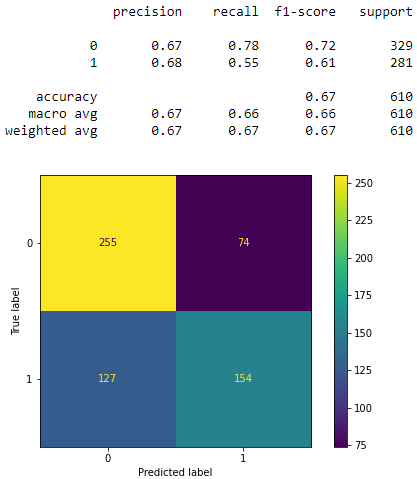
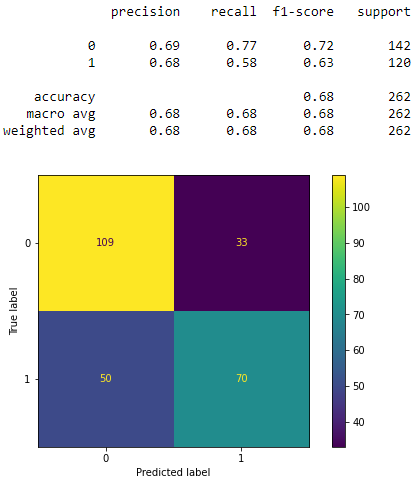
### Logit model & LDA

Now, before proceeding the further analysis we have to encode the string type data. Since the categories available in both the target variable (Holiday\_package) and the foreigner variable is ‘yes’ and ‘no’. Hence these values were simply be replaced by 0 and 1 as, 0 refers to ‘no’ and 1 refers to the ‘yes’ category.

After doing this, we can proceed to train and test split. As per the instructions, the 70,30 split was performed (70 percent samples for train and 30 percent samples for test set). Here, one must look out for the class balance ratio in train and test. Now, the Logistic regression model was fitted to train data, and predictions were made for train and test data. Similarly, Linear Discriminant Analysis (LDA) was performed on the dataset.

#### 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.

### Performance Metrics

Now, if we look at the performance metrics of models, these come out to be very similar. Though the logistic regression model needed hyperparameter tuning to get the performance comparable to the LDA. 

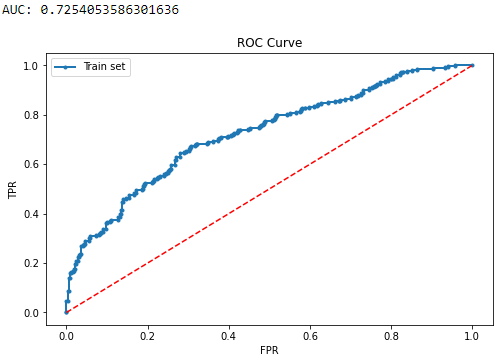
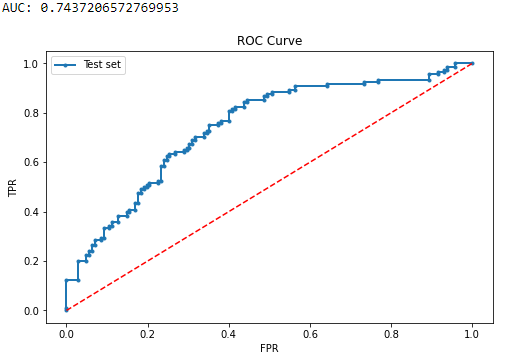
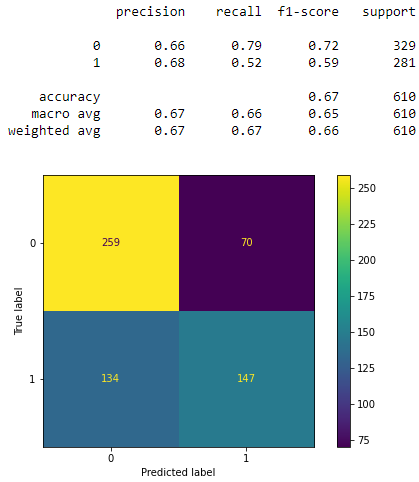
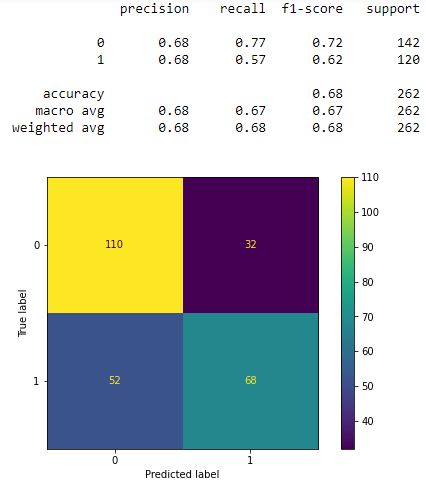
 

Figure 17: Performance metrics using Logit model for train set (left) and test set (Right)

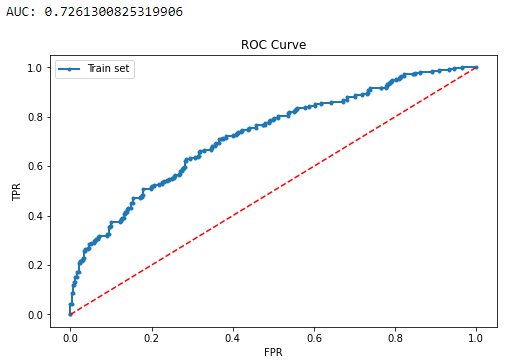
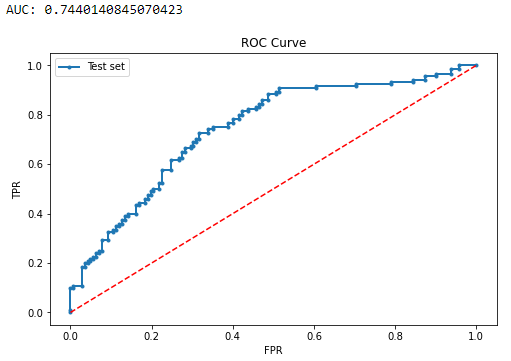
 

Figure 18: Performance Metrics for LDA on Train set (left) and Test set (right)

As it can be observed in the above figures, both the models are struggling to make good predictions considering the overall accuracy of the models. One of the reasons can be due to weak predictors available as we observed in Bi-variate analysis. LDA model is performing slightly better than the logit model considering the AUC scores. Hence, we can say that the LDA model is slightly more optimized for the predictions.

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

### Inferences

Now, if we look at the feature importance values calculated by the models, they come out to almost equal:

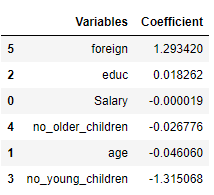
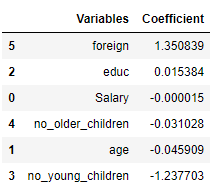
 

Figure 19: Coefficient for the variable present in the dataset for Logit model(left) and LDA (right)

Notice that the values for age, salary, educ, and no\_older\_children are almost zero which means that these are weak predictors for the target variables. On the other hand, the ‘foreign’ feature is more important for class 1 and ‘no\_young\_children’ is important for class 0 in both models.

Now, the following conclusions can be drawn from this analysis:

* Most of the features available in the dataset are weak predictors for the target feature.
* Model performance for both the models were coming out to be almost the same. Though for the LDA model the AUC score was slightly better on both train and test sets.
* More features and data of the employees are required for making the model perform better.
* Based on the analysis, the number of older children and foreigner status are important factors for predicting the holiday package status.