**Business Report on**

***Predictive Modelling***

***Submitted to***



**Great Learning Olympus**

***Prepared by***

**Mrinmoy Majumdar**

**June-D Batch**

**Post Graduate Program in Data Science & Business Analytics**

**From**

****

UT Austin

**November, 2021**

TABLE OF CONTENT

**Problem 1-Linear Regression ………………………………………4**

**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

**2.**Impute null values if present, also check for the values which are equal to zero. Do they have any meaning or do we need to change them or drop them? Check for the possibility of combining the sub levels of an ordinal variables and take actions accordingly. Explain why you are combining these sub levels with appropriate reasoning…………………………………………………………14

**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 stats model. 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…………………………………………………....17

###### **4.**Inference: Basis on these predictions, what are the business insights and recommendations……………………………..................................................21

**Problem 2-Logistic Regression and LDA………………………….23**

**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..............................................23

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).............................................................31

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………………………………..32

4. Inference: Basis on these predictions, what are the insights and recommendations................................................................................................36

**LIST OF FIGURES**

|  |  |
| --- | --- |
| **Images** | **Page No** |
| **Info and null value checking** | **6** |
| **Histogram plot** | **7-8** |
| **Bar plot** | **8-12** |
| **Correlation plot** | **12** |
| **Pair plot** | **13** |
| **Box plot** | **14** |
| **Unique sorted object values** | **16** |
| **Coefficients of independent variables** | **18** |
| **Model intercept** | **18** |
| **Parameter values** | **19** |
| **Inferential statistics** | **19** |
| **Scatter plot for OLS model** | **20** |
| **Coefficient of determination** | **20** |
| **VIF values** | **22** |
| **Final Linear regression equation** | **22** |
| **Duplicate data check** | **24** |
| **Categorical univariate count plot** | **25** |
| **Pie chart** | **26** |
| **Joint plot** | **27** |
| **Strip plot** | **28** |
| **ROC curve** | **33** |
| **Confusion matrix** | **34** |
| **Classification report** | **35** |
| **Feature importance of model** | **37** |

**LIST OF TABLES**

|  |  |
| --- | --- |
| Table 1 | 5 |
| Table 2 | 5 |
| Table 3 | 7 |
| Table 4 | 11 |
| Table 5 | 12 |
| Table 6 | 14 |
| Table 7 | 16 |
| Table 8 | 17 |
| Table 9 | 17 |
| Table 10 | 19 |
| Table 11 | 23 |
| Table 12 | 24 |

**Problem 1-Linear Regression**

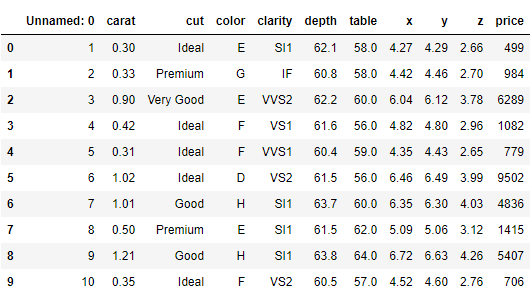
**Problem Statement:**

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

Data Dictionary:

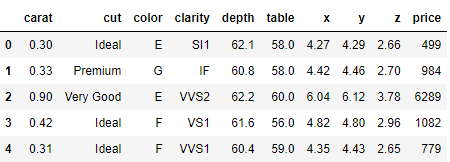
|  |  |
| --- | --- |
| Variable Name | Description |
| Carat | Carat weight of the cubic zirconia. |
| Cut | Describe the cut quality of the cubic zirconia. Quality is increasing order Fair, Good, Very Good, Premium, Ideal. |
| Color | Colour of the cubic zirconia with D being the worst and J the best. |
| Clarity | Clarity refers to the absence of the Inclusions and Blemishes. (In order from Worst to Best in terms of average 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. |

**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.**

Displaying cubic zirconia data:

**Table 1: Top 10 rows of cubic zirconia data Frame**

Unnamed: 0 column is not required as part of the linear regression model building, hence we will drop the column from the dataset.

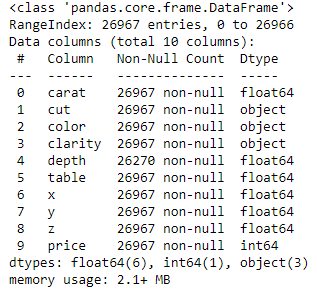


**Table 2: Data Frame without Unnamed: 0 variable**

## Basic EDA:

* Checking shape and information of data Frame

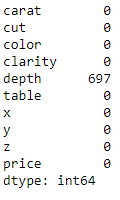
(26967, 10) – The data set contains 26967 observations of data and 10 variables. Previously there were 11 variables which included dropped column Unnamed: 0.



**Image 1: Information on cubic dataset**

### Majority of the data has 26967 instances with 10 attributes – 6 variables of float type, 1 variable of integer type and 3 variables of object type. Depth has 26270 instances which indicates presence of null values that will be seen and dealt with later.

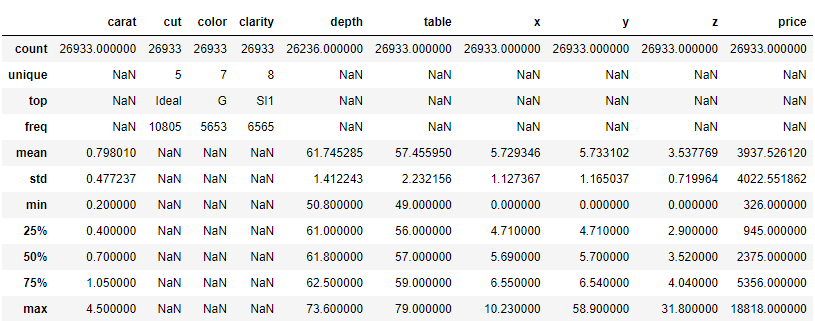
* Check the presence of missing values



**Image 2: Checking null values in data**

### There are 697 null values in depth column. There are 34 duplicate values in the dataset which was treated by dropping duplicates from the dataset. The shape of the dataset post dropping duplicates became (26933, 10).

* Checking summary of data Frame



**Table 3: Description of cubic dataset**

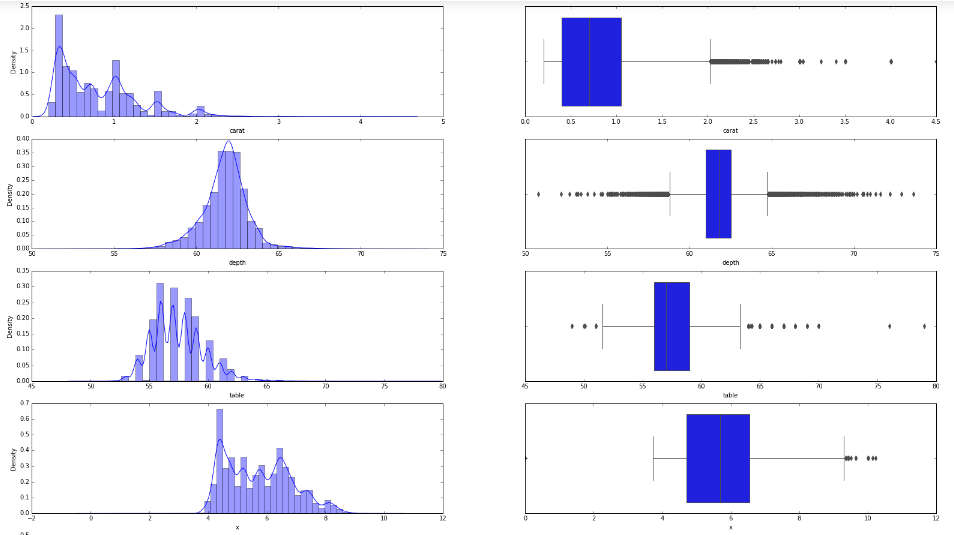
Looking at the 5 point summary, we can probably conclude that data is normally distributed as the mean and median of the columns are almost identical. The claims can be further solidified with the help of univariate, bivariate and multivariate analysis of feature columns along with its associated skewness.

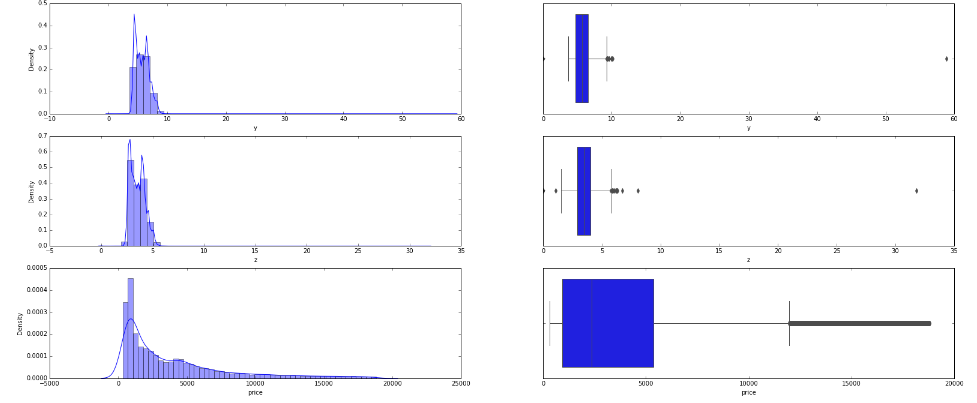
X, y and z which indicate length, width and height have minimum value as 0 which is not feasible and need to be treated.

Price is the target column that will be used to build the model.

Cut, color and clarity are categorical columns while the rest are continuous columns.

**Univariate Analysis**

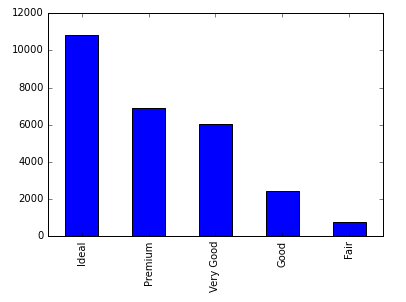




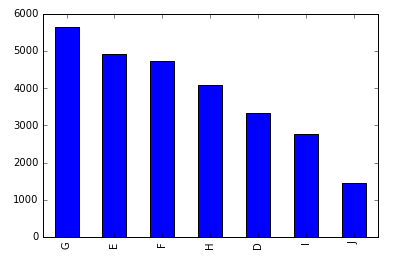
**Image 3: Histogram and boxplot of numerical columns**

Insights

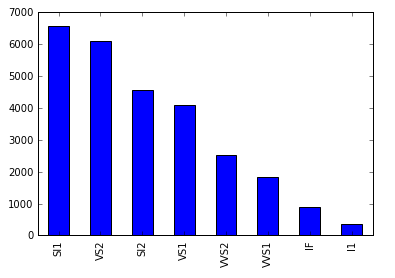
* There are outliers in the dataset as evident from the boxplot which need to be treated.
* Carat variable has multiple peaks which signifies multiple modes and is left skewed.
* Depth seems to have normal distribution.
* Table, x, z and price have positive skewness.
* Y has a distribution that is left skewed.



**Image 4: Bar plot showing cut column**



**Image 5: Color variable bar plot**

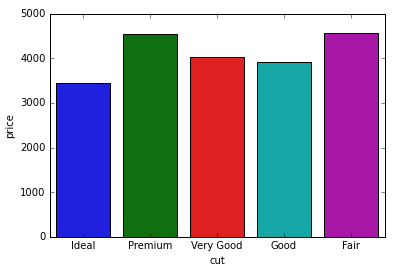


**Image 6: Bar plot of clarity feature**

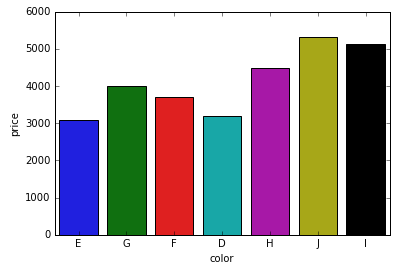
Observations**-**

* Ideal is the most preferred cut quality by manufacturers which is on expected lines.
* J is the best color but is the least selected by manufacturer which needs to be looked at.
* I1 has the best clarity and least prioritised option for manufacturers that requires investigation.

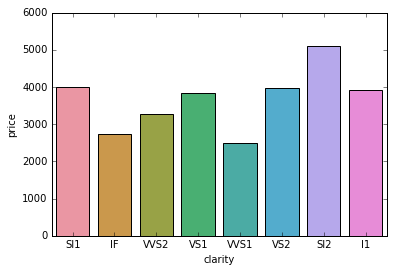
**Bivariate Analysis**



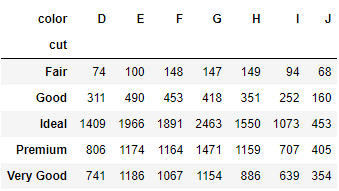
**Image 7: Cut vs price bar plot**



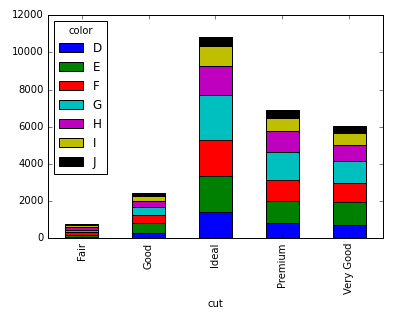
**Image 8: Bar plot for color vs price**



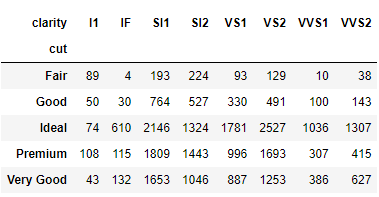
**Image 9: Clarity vs price bivariate plot**



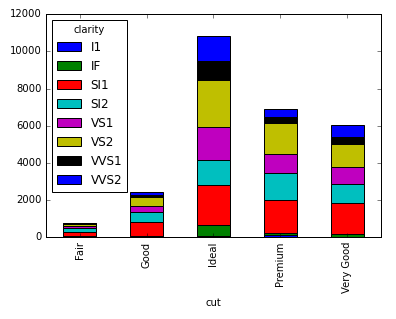
**Table 4: Crosstab of cut and color**



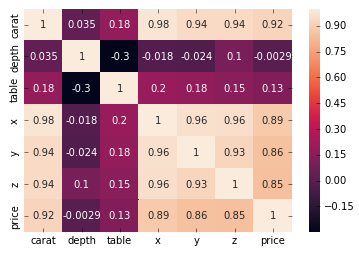
**Image 10: Stacked bar plot of categorical features**



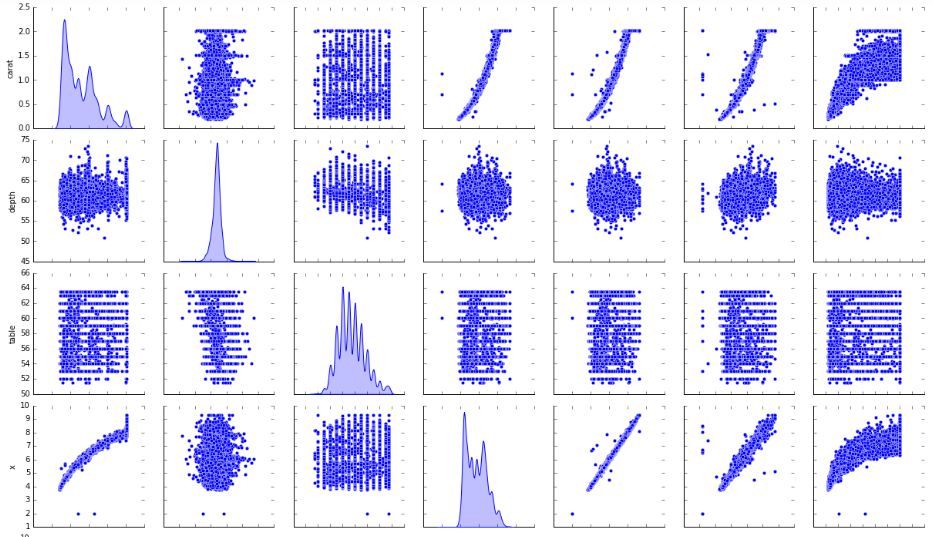
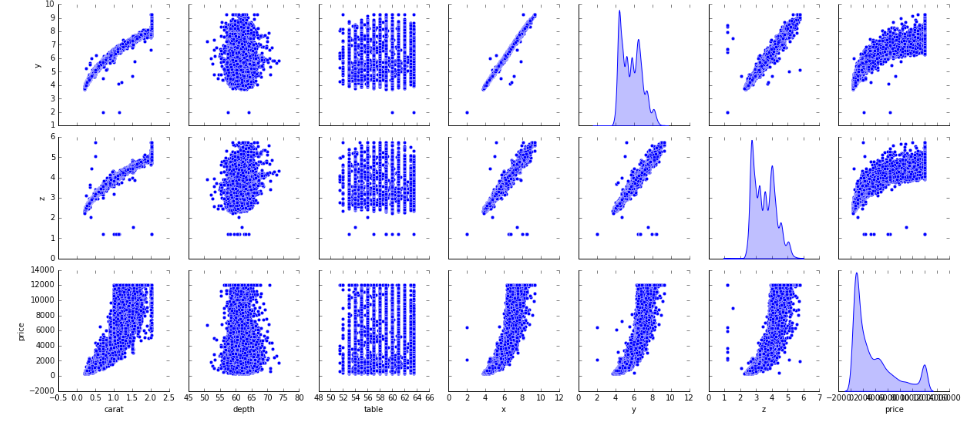
**Table 5: Categorical columns crosstab**



**Image 11: Stacked bar plot of cut and clarity**



**Image 12: Correlation plot**

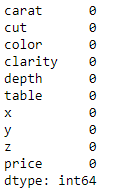
 

**Image 13: Pair plot of numerical features**

Observations**-**

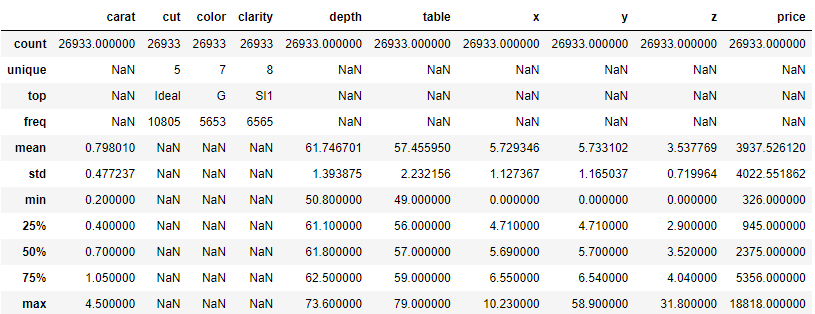
* Premium and fair cut quality has the highest price and ideal cut quality has the lowest price which explains the reason behind the manufacturer’s selection.
* J color has the highest price that is why most manufacturers least prefer this color.
* SI2 and WS1 has the highest and lowest prices respectively.
* Ideal cut and G color combinations is the most in-demand as indicated from crosstab.
* Ideal and VS2 cut-clarity combination is frequently used feature.
* Carat, x, y, z and price have a very high positive correlation with each other.
* The data distribution across various dimensions except depth do not look normal.
* Close observation between price and other attributes indicate the relationship is not really linear.

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

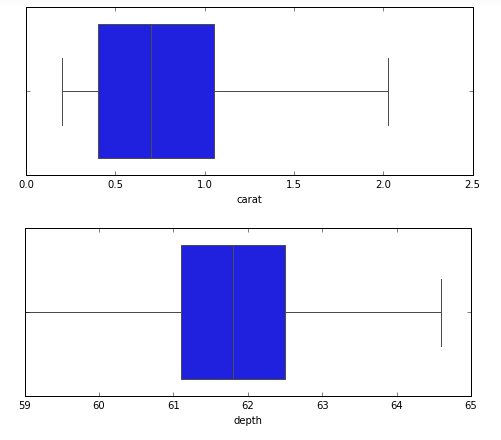


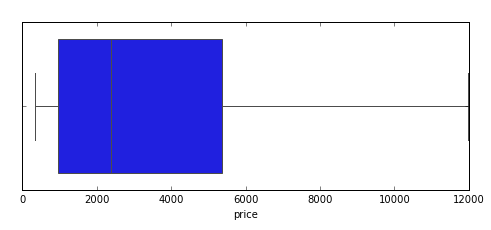
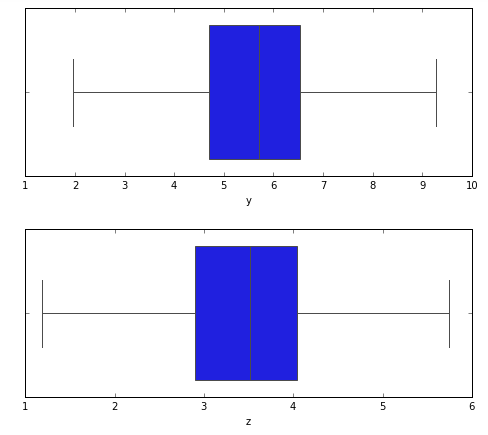
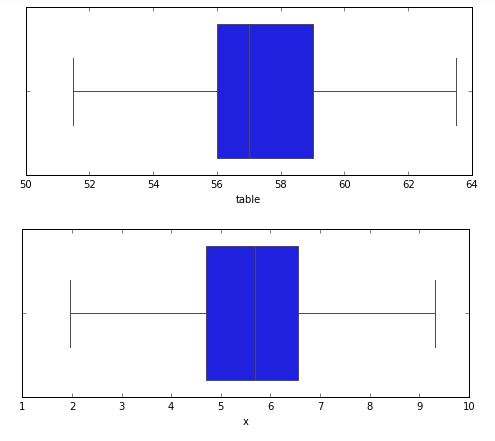
**Image 14: Null value check post imputation**

As we have seen in previous question, there were 697 null values in depth feature. There are various ways to treat missing values like dropping rows, replacing missing values with median values etc. Dropping the objectionable rows is not a good idea under all situations. Instead of dropping the rows, we replace the missing values with median values and thus the depth feature has no missing values as shown above.

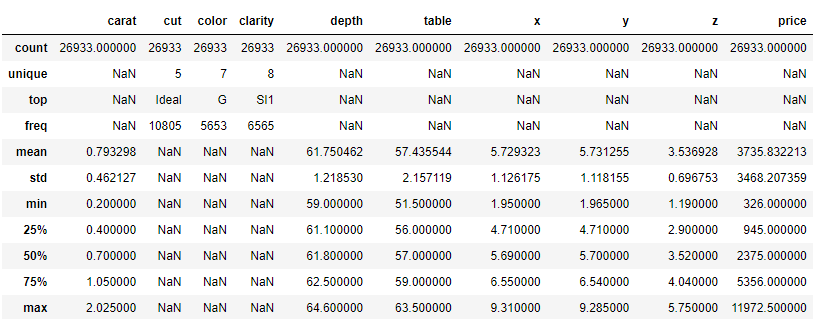


**Table 6: Dataset summary to show zero values**

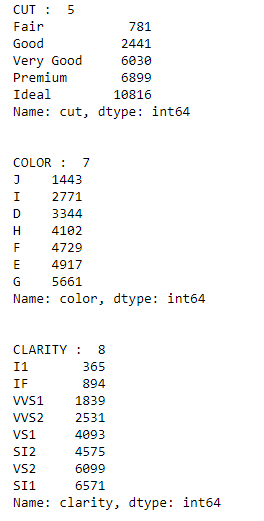
As seen from the summary above, there is presence of zeroes in x, y and z columns. They indicate presence of outliers in the data in previous question that needs to be removed using outlier treatment. 



**Image 15: Box plot post outlier removal**



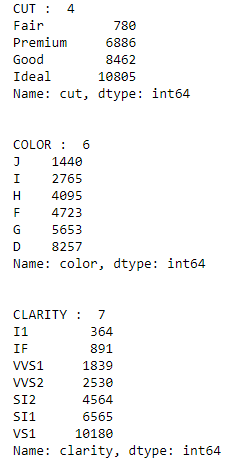
**Table 7: Zeroes have been removed**



**Image 16: Unique sorted values of categorical columns**

There are three ordinal variables in the dataset – cut, color and clarity. There are 5 categories of cut quality – Ideal, Premium, Very Good, Good and Fair. Color has 7 levels – G, E, F, H, D, I and J. Clarity has 8 different divisions – SI1, VS2, SI2, VS1, VVS2, VVS1, IF and I1.

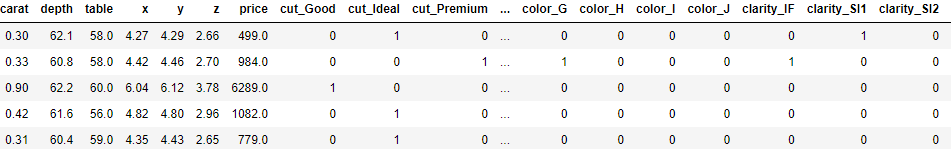
As seen previously in bivariate plots, cut quality ‘Good’ and ‘Very Good’ can be combined since their prices are identical. Similarly, colors ‘E’ and ‘D’ and clarity ‘VS2’ and ‘VS1’ can also be combined due to same price.



**Image 17: Categorical unique values post sub-level combination**

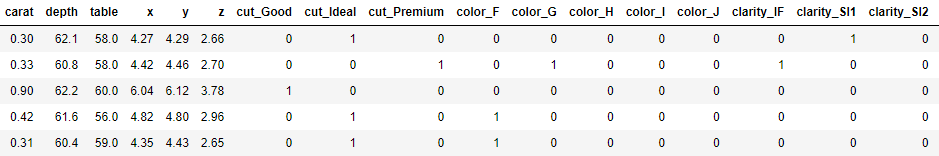
**Encode the data (having string values) for Modelling. Split the data into train and test (70:30). Apply Linear regression using scikit learn. Perform checks for significant variables using appropriate method from statsmodel. Create multiple models and check the performance of Predictions on Train and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

Categorical variables can be converted into dummy/indicator variables. Many columns will be created with distinct values which is known one-hot encoding. Cut, color and clarity are having string values which have been encoded before applying linear regression model.



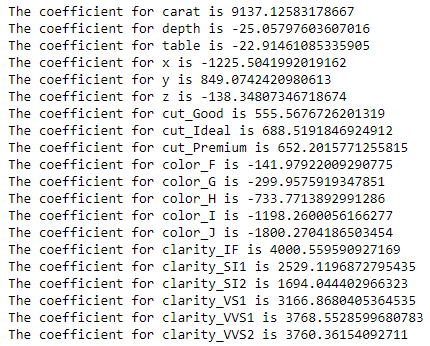
**Table 8: Dataset post encoding**

Train-test split is applied on the dataset with 70% training data and 30% testing data. All the predictor variables are copied into X data frame. Since price is dependent variable, so it is dropped. The price column is copied into y data frame.



**Table 9: Head of training set**

The linear regression function is invoked on the training data to find the best fit model. We can explore the coefficients of each of the independent attributes.



**Image 18: Coefficients of independent variables**

Let us check the intercept for the model.



**Image 19: Model intercept**

R-square for the training data = 0.939

Approximately, 94% of the variation in the price is explained by the predictors in the model for train set.

R-square for the testing data = 0.941

R-square values of both training and testing data are almost identical.

RMSE on training data = 854.20

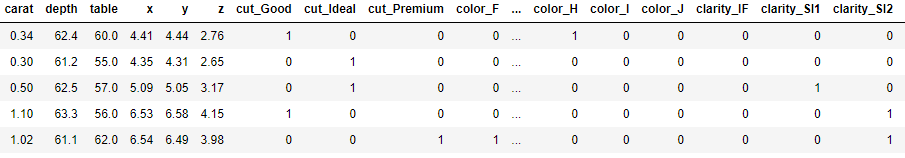
RMSE on testing data = 843.40

RMSE values of training and testing data are also close to each other.

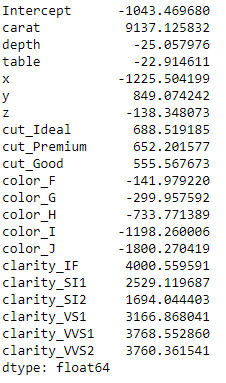
Linear regression using scikit learn seems to be a very good model based on the values as shown above.

Now, we will look at linear regression model using statsmodel.

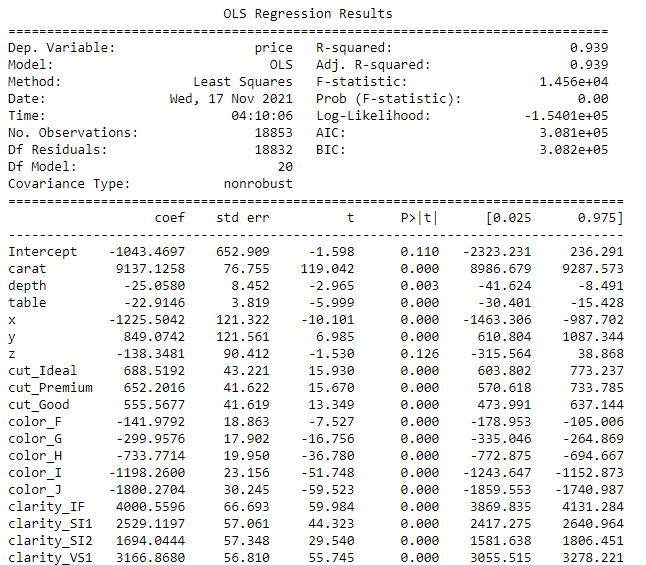
R^2 is not a reliable metric as it always increases with addition of more attributes even if the attributes have no influence on the predicted variable. Instead we use adjusted R^2 which removes the statistical chance that improves R^2.Scikit does not provide a facility for adjusted R^2.so we use statsmodel, a library that gives results similar to what you obtain in R language. This library expects the X and Y to be given in one single data frame.

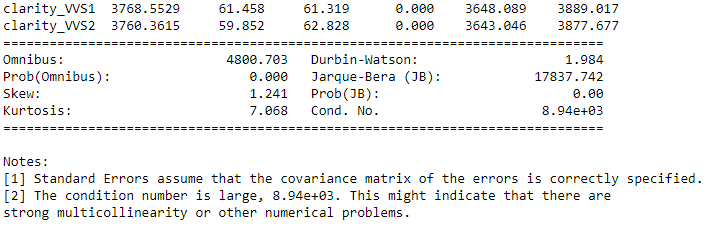


**Table 10: Training data head**



**Image 20: Parameter values of OLS model**





**Image 21: Inferential statistics**

The overall P value is less than alpha, so rejecting H0 and accepting Ha that at least 1 regression co-efficient is not 0. Here all regression coefficients are not 0.

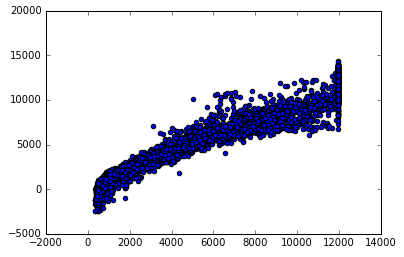
Let us check the sum of squared errors by predicting value of y for test cases and subtracting from the actual y for the test cases. We will do the same for train cases. Under root of mean squared error is standard deviation i.e. average variance between predicted and actual.

MSE for training data = 854.20

MSE for testing data = 843.40

The above two values from OLS model are same as found for linear regression model. So there is average of 843 (round off) price difference from real price on an average for both the models investigated till now.

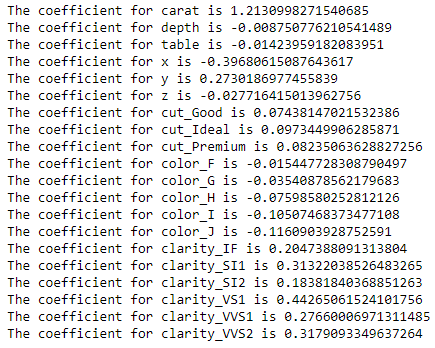
Since this is regression, plot the predicted y value vs actual y values for the test data. A good model's prediction will be close to actual leading to high R and R^2 values.



**Image 22: Scatter plot for OLS model**

We can improve the model by improving existing R^2 value. The independent attributes have different units and scales of measurement .It is always a good practice to scale all the dimensions using z scores or some other method to address the problem of different scales.

We applied z-score for this problem to check if the model performance improves further.



**Image 23: Coefficient of determination for scaled linear regression model**

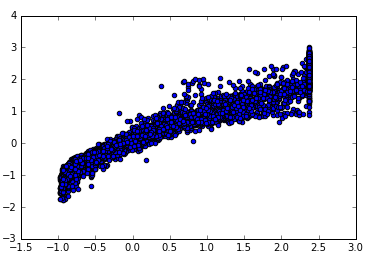
The coefficient of determination for most of the attributes reduces as compared to original linear regression model.



**Image 24: Intercept of scaled linear regression model**

The intercept of the scaled linear regression is decreased to a great extent as compared to simple linear regression model.

Scaled regression model score for testing data remains same as the simple linear regression model. MSE of scaled regression model for testing set is 0.243 which drastically dropped in comparison to simple linear model.



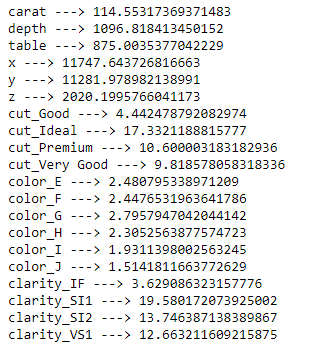
**Image 23: Scatter plot for Scaled linear model**

There is not much difference in scatter plots for both scaled linear and simple linear models.

We prefer to choose either simple linear regression model or OLS model but not the scaled linear regression model based on the different parameters like coefficient of determination, intercept, scatter plot, accuracy score for training and testing data, R^2 , adjusted R^2, MSE and RMSE values.

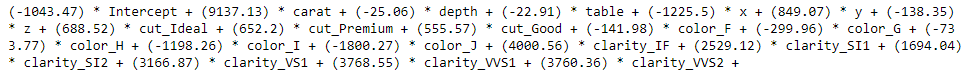
**Inference: Basis on these predictions, what are the business insights and recommendations.**

The variance inflation factor is used to check for multicollinearity of dataset. The below snapshot values clearly indicates presence of multicollinearity of data and to reduce it we can skip few columns while performing OLS or simple linear regression model. This further leads to validation of model performance based on other related model performance measures like accuracy and root mean squared error .We will look at business insights and recommendations based on predictions made simple by linear regression model and OLS model. In this case we have decided not to remove any feature while looking for ways to improve the model as more or less important parameters remain the same.



**Image 24: VIF values of each feature**

The final linear regression model equation is as follows –



**Image 25: Final linear regression equation**

When carat increases by 1 unit, price increases by 9137 units, keeping all other predictors constant. Similarly, when cut\_Ideal increases by 1 unit, price increases by 688 units, keeping all other predictors constant.

There are also some negative co-efficient values, for instance, depth has its corresponding co-efficient as -25.06. This implies, when depth increases, the price decreases by 25 units, keeping all other predictors constant.

Business insights and recommendations

* The five most important attributes for determining price of diamond are carat, clarity\_IF, clarity\_VVS1, clarity\_VVS2 and clarity\_VS1.
* Presence of inclusions and blemishes do not lead to less profits for the company rather it contributes to higher price range of diamonds.
* The manufacturer should focus most on carat weight of the diamond to bring in more sales to the company.
* The selected model can be termed as right fit due to exceptional training and testing accuracy results.
* There were presence of outliers, null values and duplicates in the dataset which were removed otherwise it could have brought the model right to its knees.
* Price is inversely proportional to color of diamond, hence neglecting this attribute won’t affect sales as such.
* Width of cubic zirconia needs more focus from manufacturer while designing rather than length and height of the same.
* Depth or table features could be removed during model evaluation but it may lead to increase in accuracy but decrease in intercept further.

**Problem 2: Logistic Regression and LDA**

**Problem Statement:**

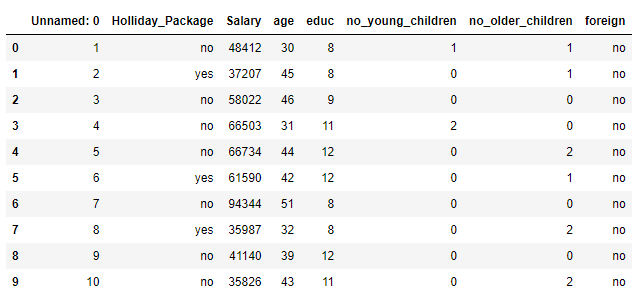
You are hired by a tour and travel agency which deals in selling holiday packages. You are provided details of 872 employees of a company. Among these employees, some opted for the package and some didn't. You have to help the company in predicting whether an employee will opt for the package or not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

Data Dictionary:

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

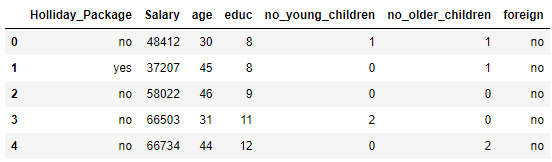
**Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, write an inference on it. Perform Univariate and Bivariate Analysis. Do exploratory data analysis.**

**Displaying Holiday Package data –**

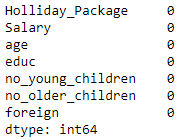


**Table 11: Holiday package dataset**

Unnamed: 0 is not required as part of logistic regression model and linear discriminant analysis, so it is dropped.

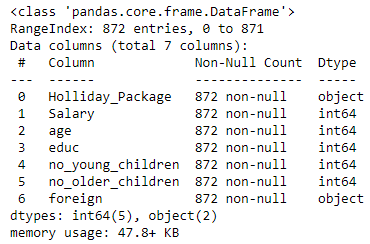


**Table 12: Dataset post dropping a feature**



**Image 26: Null value check**

There are no missing values in the dataset.



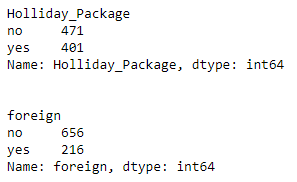
**Image 27: Information on dataset**

5 variables are numeric and remaining categorical. Categorical variables are not in encoded format.



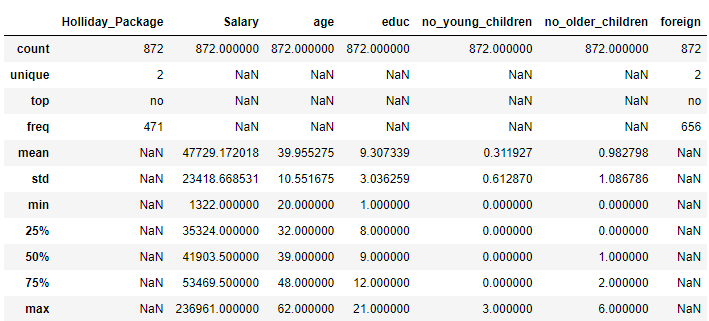
**Image 28: Duplicate data check**

There are no duplicate values in the dataset. The data set contains 872 observations of data and 7 variables.



**Image 29: Unique counts of all objects**

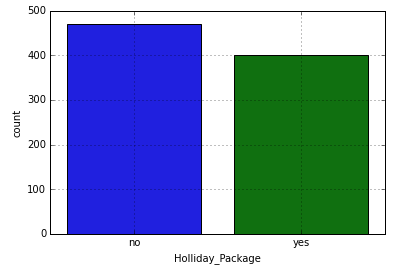
There are 216 foreigners out of 872 employees in a company. 401 employees have opted for holiday package which is less as compared to those who have not.



**Table 13: Descriptive statistics of data frame**

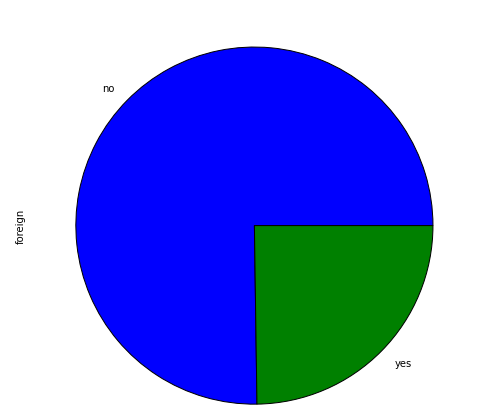
Age and education variables have mean and median values close to each other indicating normal distribution. Salary feature has mean much greater than median, hence it has a skewed distribution. 872 employees working in the company have age ranging from 20 to 62 with maximum 6 older children and 3 young children and earning a reasonable salary.

**Univariate Analysis**



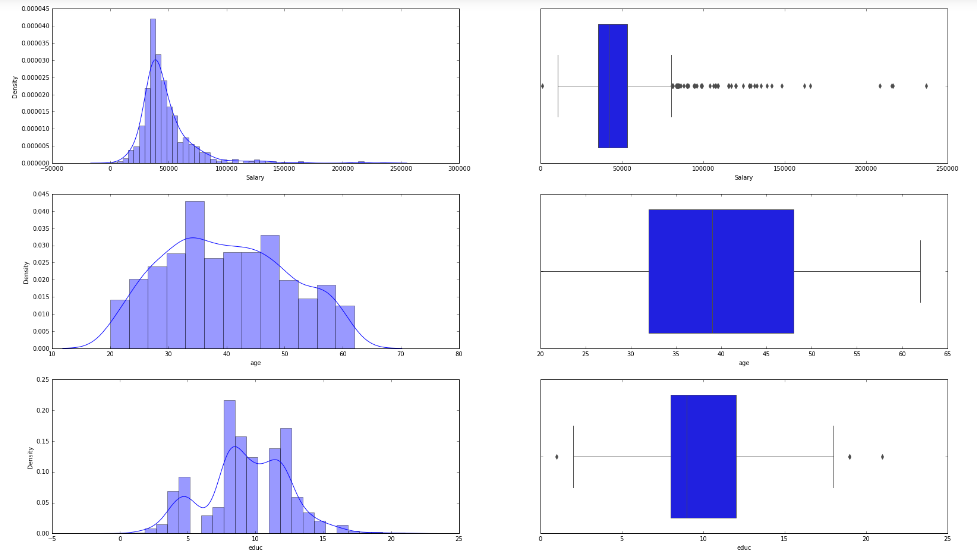
**Image 30: Count Plot of holiday package**

The count of employees with no holiday package is more compared to those having one.



**Image 31: Pie chart of foreign employees**

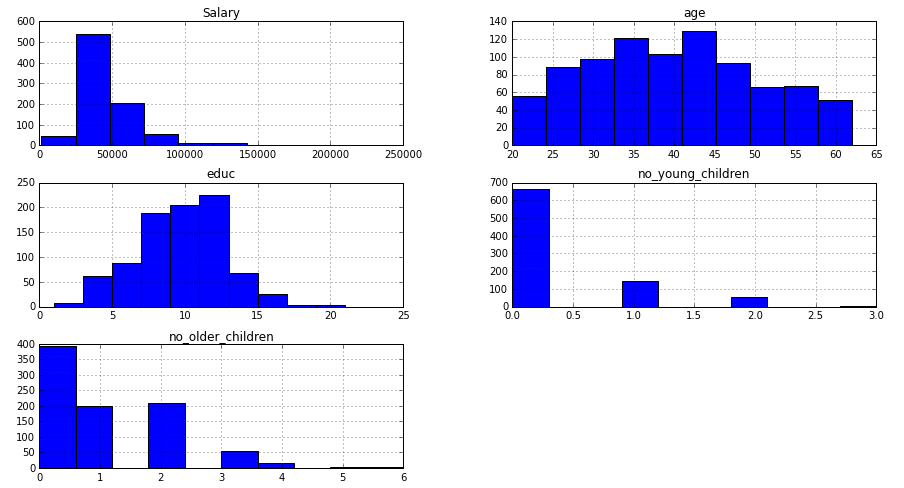
There are less foreign employees in the company who have choice of opting for holiday package.



**Image 32: Spread of data using plots for continuous columns**

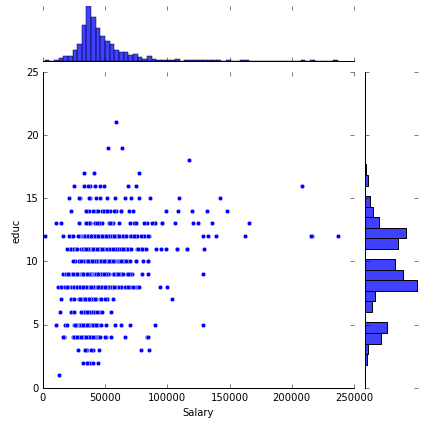
Salary and education are having outliers in the data. Age has no outliers in the data. Salary and age seem to have a normal distribution whereas education has multiple peaks in the histogram plot which indicates left skewness.

Although outliers exists as per the boxplot, by looking at the data distribution in descriptive statistics, the values are not too far away. Treating the outliers by converting them to min/max values will cause most variables to have values to be the same. So, outliers are not treated in this case.



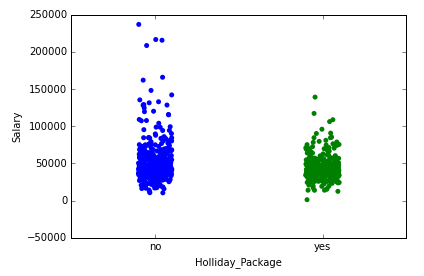
**Image 33: Distribution of numeric features**

**Bivariate Analysis**



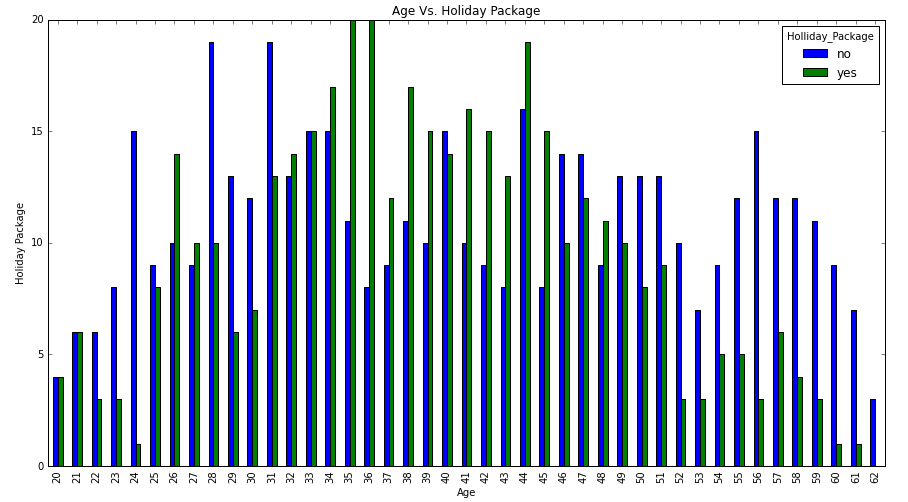
**Image 34: Salary vs education joint plot**

There are very few employees with higher years of education level earning a salary package from a lower to higher range. Most of the employees are having a low salary package independent of years of education which is explained by large concentration of points in the above plot.



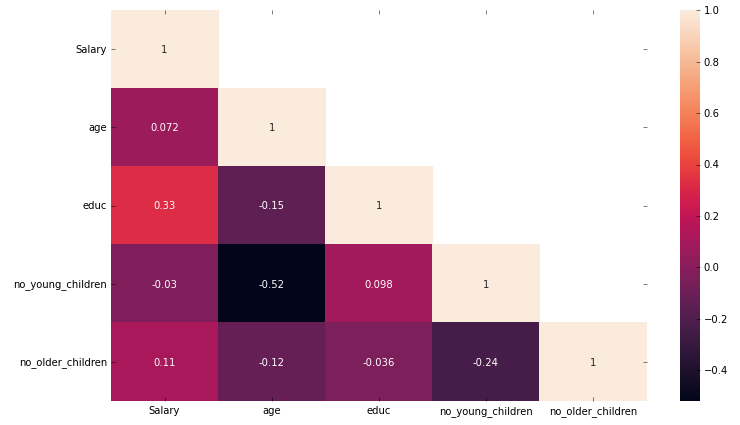
**Image 35: Strip plot of numeric and categorical combination**

We can see from the above plot that employees with no holiday package have a higher salary package which needs focus from the travel agency. Median salary package employees have both opted and not opted for the holiday package.



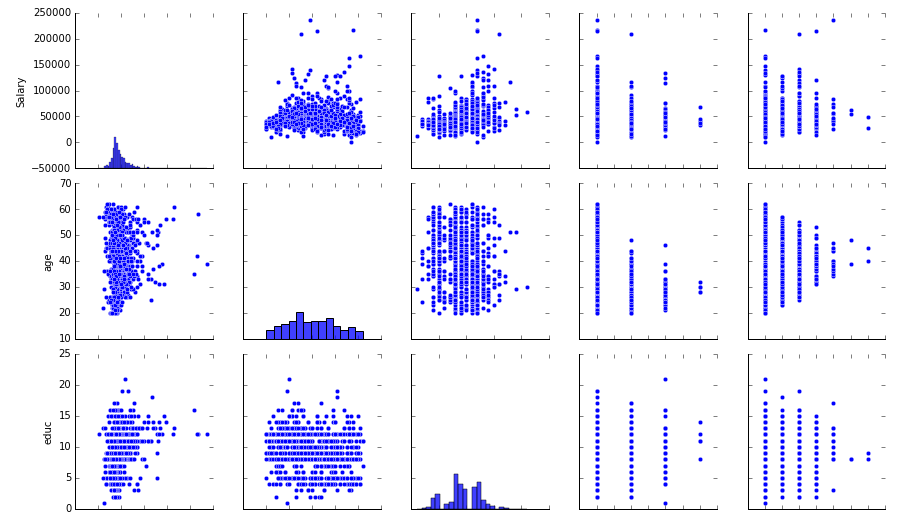
**Image 36: Bar plot of age vs holiday package**

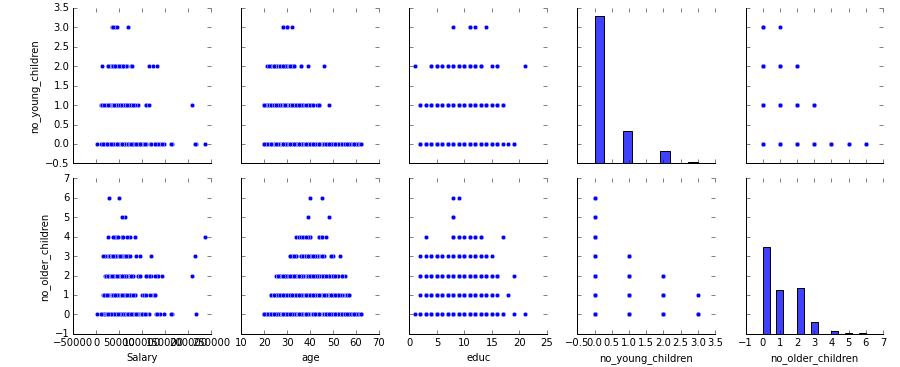
Employees with age 35-36 are highly probable to select the holiday package offered by the tour company. 28 and 31 aged employees are reluctant to opt for holiday package which should be the target customers for the travel firm. Employees at the tail end of their careers do not want to go on a tour whereas employees at the early stage of their careers are equally likely to select the package.



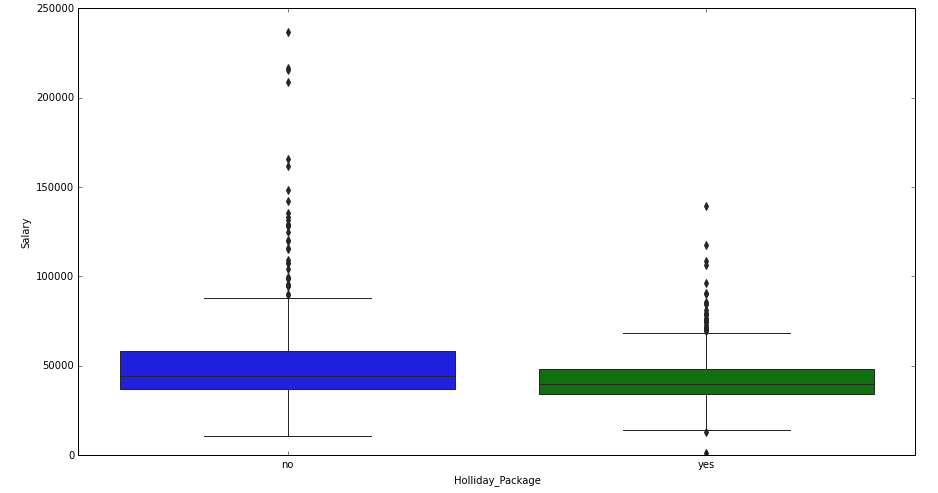
**Image 37: Heat map of numerical columns**

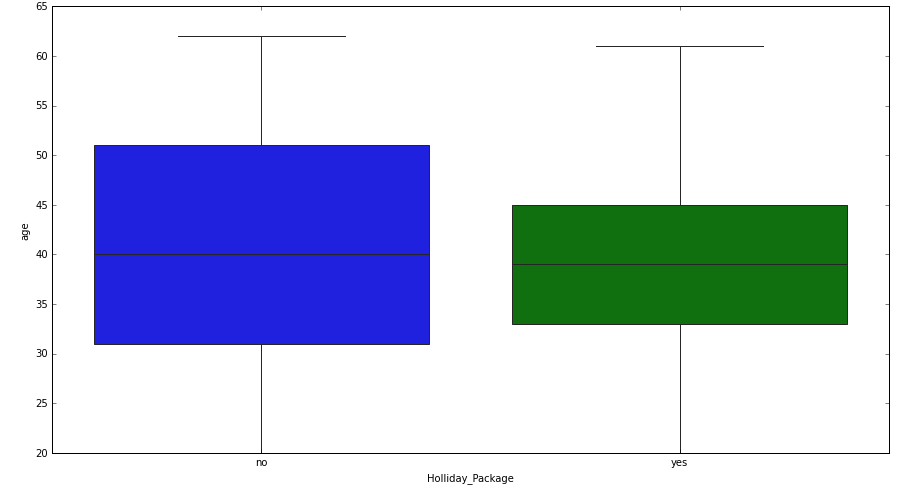
Salary has positive correlation with age, education and number of older children. Age has a negative correlation with education, number of younger and older children. Education is having positive correlation with only number of younger children.

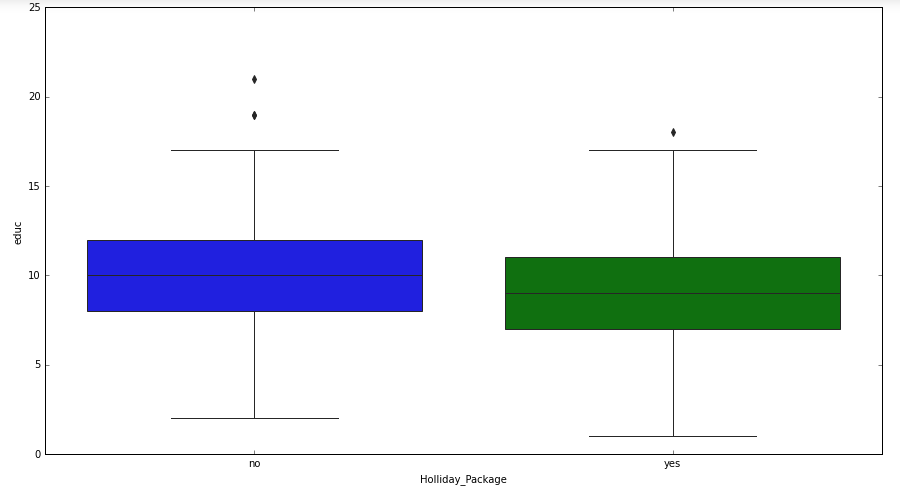


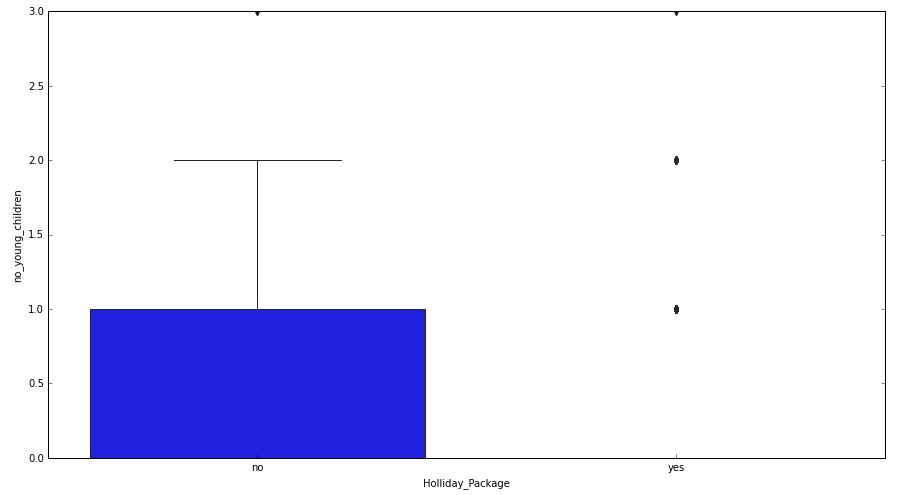


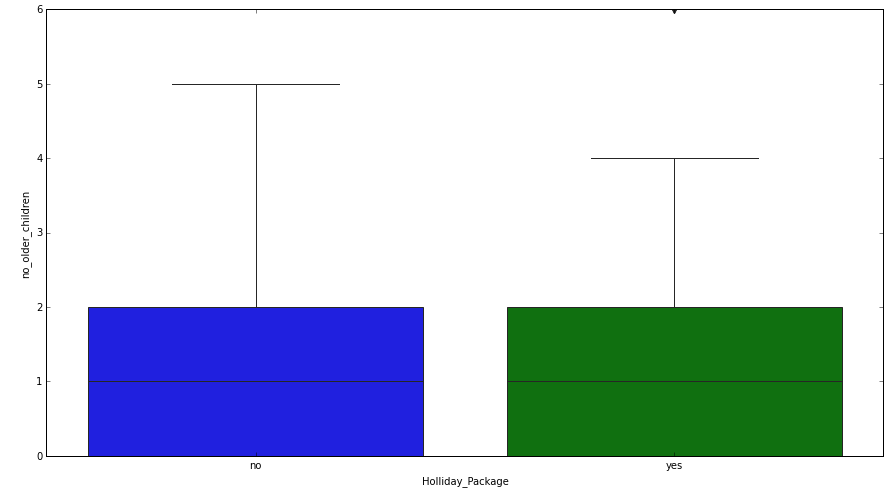
**Image 38: Pair plot of holiday package dataset**







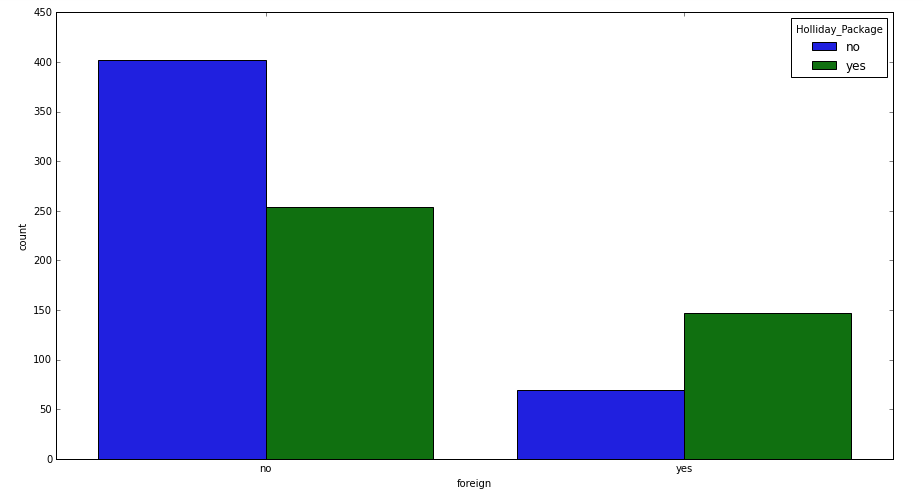




**Image 39: Bivariate analysis of numerical features with target variable**

Insights

* Number of older children shows similar distribution between holiday package and no holiday package, and is normally distributed.
* Salary shows clear distinction between package and no package. Employees who has no package shows a wider distribution indicating more salary with many outliers indicating few employees who have more salary still has no package whereas employees who have package has smaller distribution.
* Age and education have similar medians for employees with and without salary package.



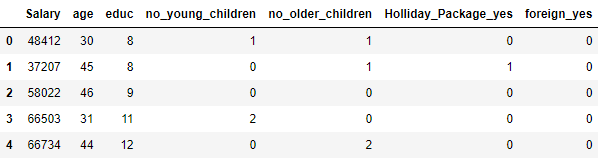
**Image 40: Bivariate analysis of categorical variable with target variable**

Employees who are not foreigners are willing to opt for holiday package from the travel firm but in that category most of them are reluctant whereas though the count is less, majority of foreigners are interested in taking up holiday package.

**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).**

There are two variables in the data with string values – foreign and Holliday\_Package.

We encode the data by converting object type variables into dummy variables which is also known as dummy encoding technique.



**Table 14: Dataset post dummy encoding**

The target variable is Holliday\_Package which has combined ‘yes’ and ‘no’ values into a single column Holliday\_Package\_yes post encoding.

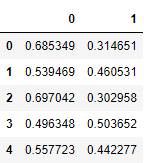
We split the data using train-test split technique in a ratio of 70:30. The predictor variables are copied into X data frame and target variables are copied into Y data frame.

**Image 41: Training and testing probabilities of target variable**

We will apply logistic regression model to the dataset. We are making some adjustments to the parameters in the Logistic Regression Class to get a better accuracy. The below hyper-parameters have been used to fit the logistic regression model as shown below –

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



**Table 15: Predicted Classes and Probabilities**

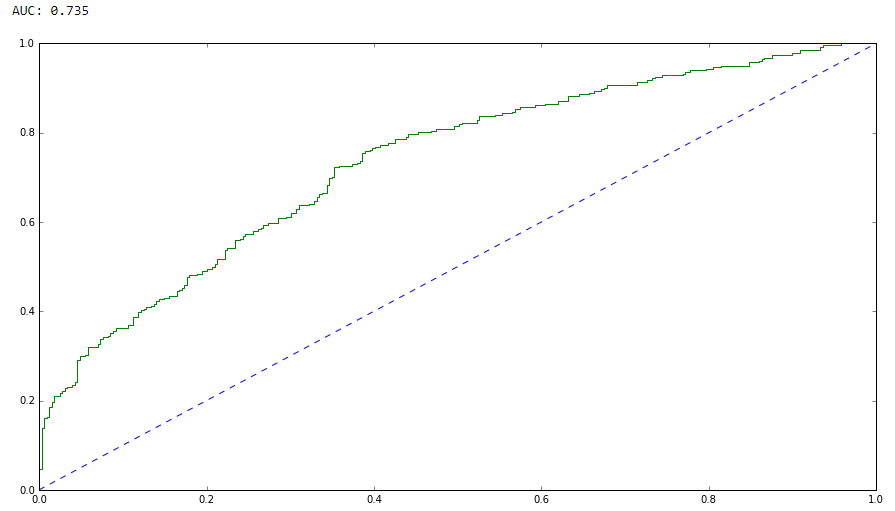
We will build Linear Discriminant Analysis model on the training and testing data post-split.

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

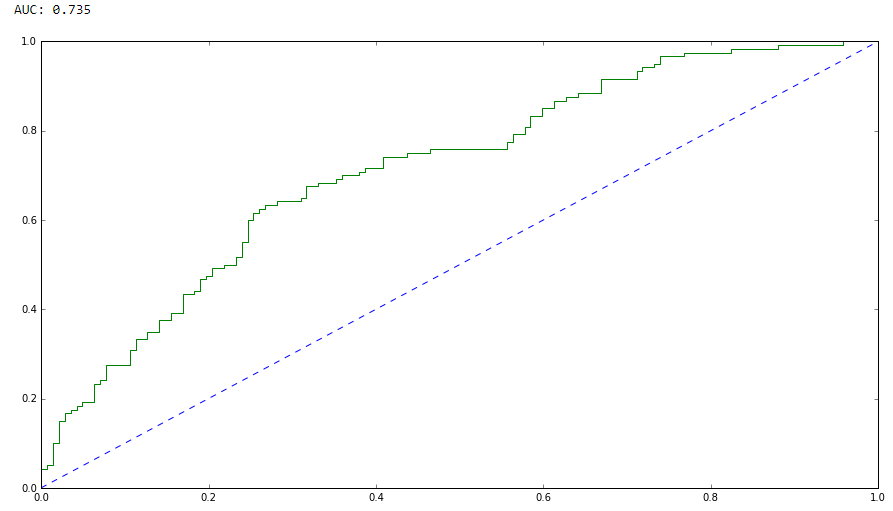
Accuracy of logistic regression model on training data = 0.6672

Accuracy of logistic regression model on testing data = 0.6526

The logistic regression model scores also known as accuracy for both training and testing samples are not very far away from each other. Hence, it seems to be a good model.

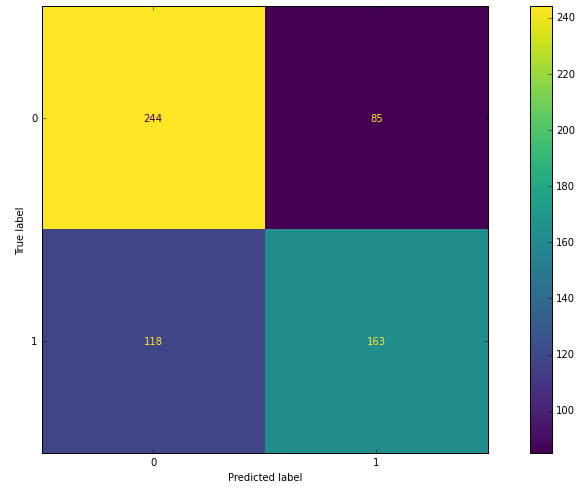


**Image 42: ROC curve for logistic regression training data**

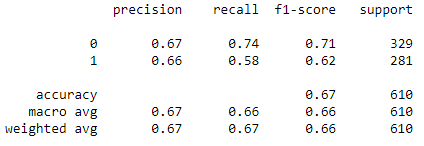


**Image 43: ROC curve for logistic regression testing data**

The ROC\_AUC score is 0.735 for both train and test data and the ROC plot for both train and test samples looks identical for logistic regression model.

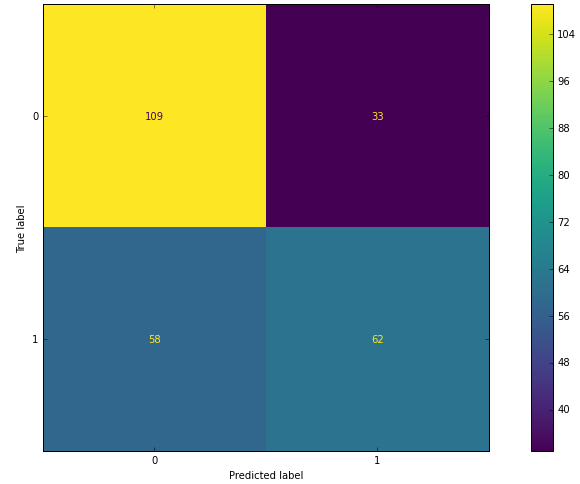


**Image 44: Confusion matrix of logistic regression trained data**

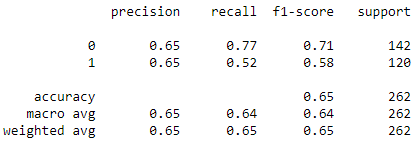


**Image 45: Classification report of logistic regression trained data**

The true positive and true negative values of trained data are higher than false positive and false negative values which signifies that predictions have been good on the trained part of the data and the precision, recall and f1-scores suggest the same.

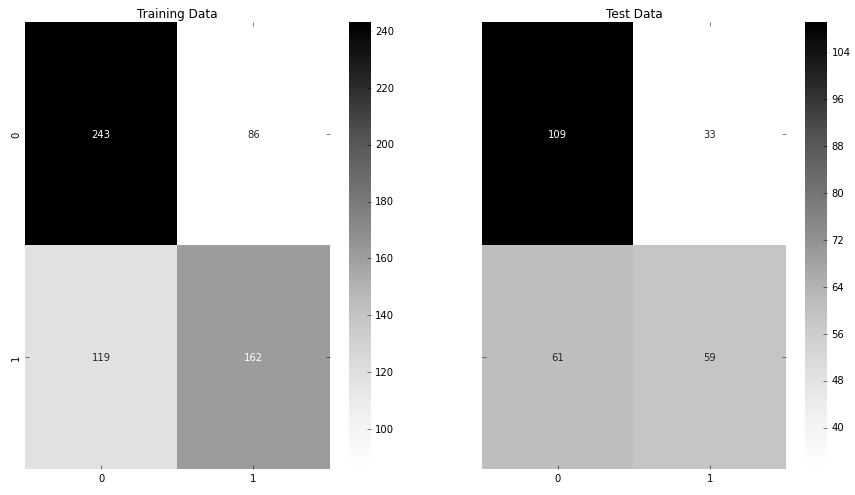


**Image 46: Confusion matrix of logistic regression tested data**

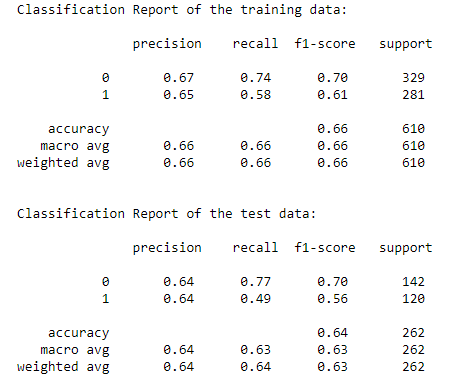


**Image 47: Classification report of logistic regression tested data**

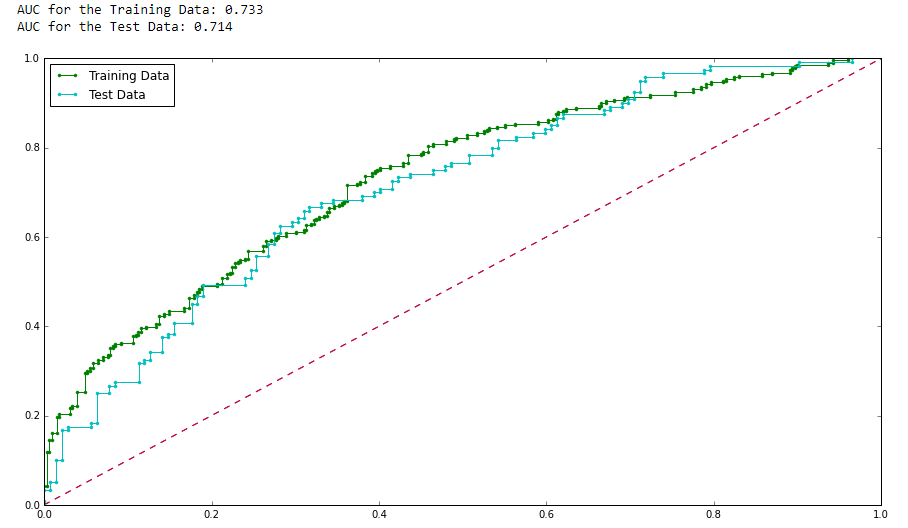
Testing data has room for improvement as false positives and false negatives outweigh their true counterparts but the precision, recall and f1-scores are quite close to the trained data which represents decent model selection.



**Image 48: Confusion matrix comparison for LDA**



**Image 49: Classification report comparison for LDA**



**Image 50: ROC plot and AUC score comparison for LDA**

LDA also looks to be quite a fine model considering all the factors as shown above.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Logistic regression train | Logistic regression test | LDA train | LDA test |
| Accuracy | 0.67 | 0.65 | 0.66 | 0.64 |
| AUC | 0.73 | 0.73 | 0.73 | 0.71 |
| Recall | 0.58 | 0.52 | 0.58 | 0.49 |
| Precision | 0.66 | 0.65 | 0.65 | 0.64 |
| F1-score | 0.62 | 0.58 | 0.61 | 0.56 |

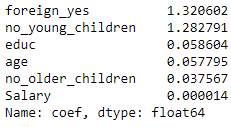
**Table 16: Logistic regression and LDA model comparison**

The logistic regression model is the better optimized model as compared to LDA model as the model performs better on both training and testing data without compromising on accuracy, AUC, recall, precision and f1-scores and the dependent variable has binary classification which is best suited for logistic regression. For more than two classes of dependent feature, LDA would have been selected.

**Inference: Based on these predictions, what are the business insights and recommendations?**

***Insights and Recommendations* –**

* **The model performance can be improved by working around the hyper-parameters of logistic regression and LDA.**
* **Employees above 50 are not interested in holiday packages. The tour and travel firm can offer holiday packages to places suitable to this category of people.**
* **Employees with high salary packages are not taking up the package, so in order to attract them the travel firm can provide perks and vacation leave in consultation with the company they work for.**
* **Special discounts can be offered to national employees to lure them into opting for a package as they are reluctant to select one.**
* **Feedback needs to be taken from employees not opting for packages and their suggestions on areas of improvement which will lead to further planning and action from the firm.**
* **Average salaried employees can be grouped together for a discounted package plan that they can opt without hesitation and second thoughts from the budget point of view.**
* **Holiday packages to places where younger children are most likely to enjoy themselves can be offered to their parents/employees to increase sales.**
* **Foreign tour packages is also a good option to bring in more revenue.**
* **Important factors for agency to consider while selling tour packages to their prospective customers in decreasing order are as follows –**



**Image 51: Feature importance of model**