Predictive Modelling Project

PGP – DSBA

May ‘22

Date: 30/10/2022

Swaathi Ramakrishnan

**Table of Contents**

**Contents**

Problem 1…………………………………………………………………………………………………………………………………..4

1. Problem Statement………………………………………………………………………………………………………………..4

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

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…………………………………………………………………………...10

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……………………………………………………………………………………………….11

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

Please explain and summarise the various steps performed in this project. There should be proper business interpretation and actionable insights present…………………………………………………………..15

Problem 2………………………………………………………………………………………………………………………………..16

2. Problem Statement………………………………………………………………………………………………………………16

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…………………………………………………………………………………………………………………………………...17

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).

………………………………………………………………………………………………………………………………………………..21

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……………………….22

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

Please explain and summarise the various steps performed in this project. There should be proper business interpretation and actionable insights present………………………………………………………….26

List of figures

Fig 1. Data description…………………………………………………………………………………………………..5

Fig 2. Univariate Analysis – Count Plots…………………………………………………………………………6

Fig 3. Box and Distribution plots…………………………………………………………………………………...7

Fig 4. Pair Plot……………………………………………………………………………………………………………….8

Fig 5. Correlation Heatmap…………………………………………………………………………………………...9

Fig 6. Imputing null values and dropping 0’s………………………………………………………………..10

Fig 7. Encoding the string variables……………………………………………………………………………..11

Fig 8. Linear Regression before outlier treatment………………………………………………………..11

Fig 9. Linear Regression after outlier treatment…………………………………………………………..11

Fig 10. Box Plot after outlier treatment………………………………………………………………………..12

Fig 11. Model 1 Summary…………………………………………………………………………………………….13

Fig 12. Model 2 Summary…………………………………………………………………………………………….13

Fig 13. Model 10 Summary…………………………………………………………………………………………..13

Fig 14. Model 11 Summary…………………………………………………………………………………………..14

Fig 15. Complete summary…………………………………………………………………………………………..14

Fig 16. VIF Values………………………………………………………………………………………………………….14

Fig 17. Descriptive statistics Dataset 2…………………………………………………………………………..17

Fig 18. Univariate Analysis Box Plot and Histogram……………………………………………………….18

Fig 19. PairPlot Holiday Package……………………………………………………………………………………19

Fig 20. Correlation Heatmap Holiday Package……………………………………………………………….20

Fig 21. ROC Curve on Train Data…………………………………………………………………………………...22

Fig 22. ROC Curve on Test Data……………………………………………………………………………………..22

Fig 23. Classification Report and Confusion Matrix on Train Data…………………………………..22

Fig 24. Classification Report and Confusion Matrix on Test Data…………………………………...22

Fig 25. ROC Curve on Train Data lbfgs solver………………………………………………………………….23

Fig 26. ROC Curve on Test Data lbfgs solver…………………………………………………………………...23

Fig 27. Classification Report and Confusion Matrix on Train Data lbfgs solver………………..23

Fig 28. Classification Report and Confusion Matrix on Test Data lbfgs solver………………….23

Fig 29. LDA Confusion Matrix………………………………………………………………………………………...24

Fig 30. LDA Classification Report…………………………………………………………………………………….24

Fig 31. ROC Curve for LDA……………………………………………………………………………………………...24

**Problem 1**

1. 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.

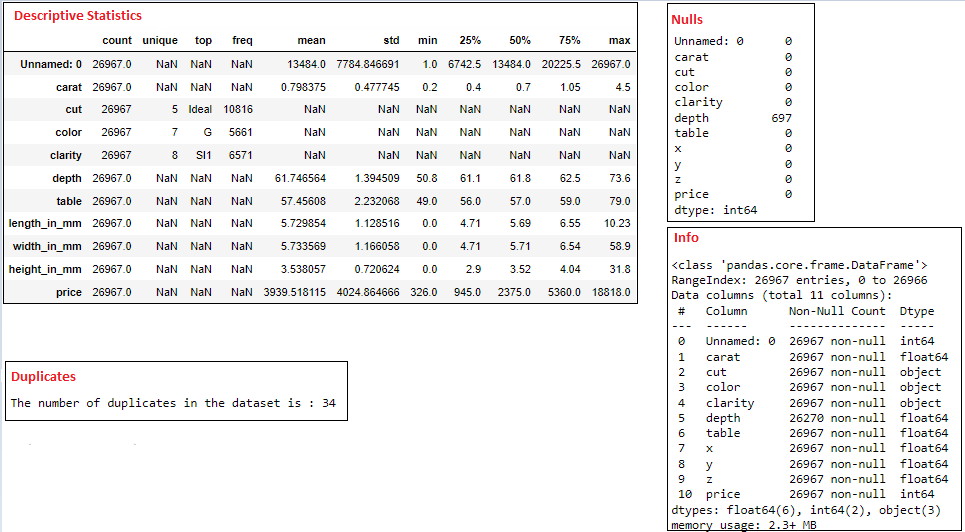
Data Dictionary

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

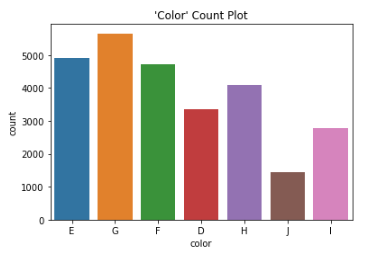
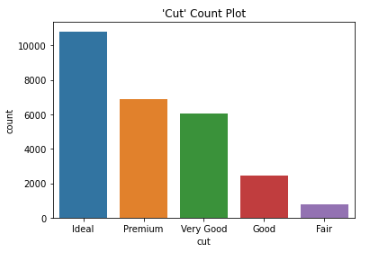
1.1. Read the data and do exploratory data analysis. Describe the data briefly. (Check the null values, Data types, shape, EDA, duplicate values). Perform Univariate and Bivariate Analysis.

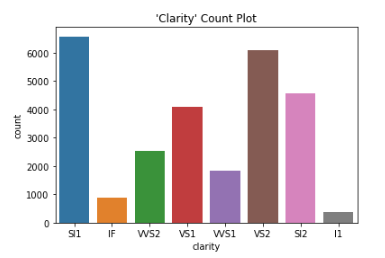
From the below figure *Fig1.* we can arrive at the following observations:

* There are 26967 rows and 11 columns
* Cut, color and clarity are object datatypes, and the rest numeric
* Depth column has 697 null values (this will be imputed with median)
* Renaming the columns as below
  + x = length\_in\_mm
  + y = width\_in\_mm
  + z = height\_in\_mm
* Most of the cut quality is Ideal, color is G and clarity is SI1
* Unnamed: 0 column can be removed from the dataset as it does not add any value
* The dependant variable is Price and its value ranges from minimum 326.00 to maximum 18818.00. The mean is 3940.00 and its median is 2375.00
* There are 0’s in length, width and height columns
* There are 34 duplicate values in the dataset. These will be dropped

*Fig 1. Data description*

Univariate Analysis

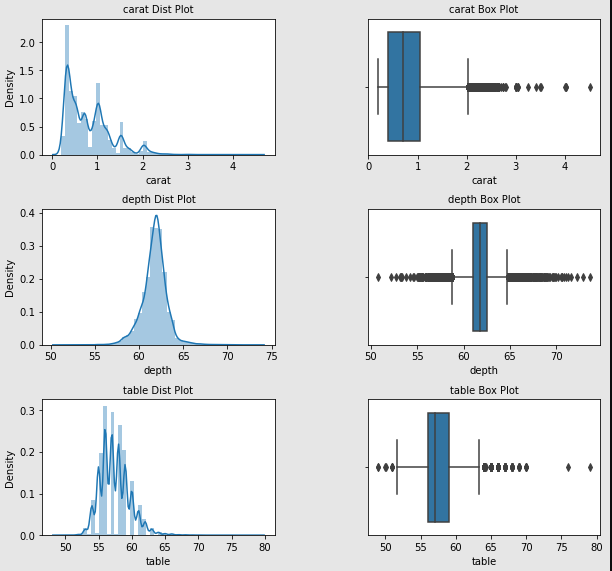


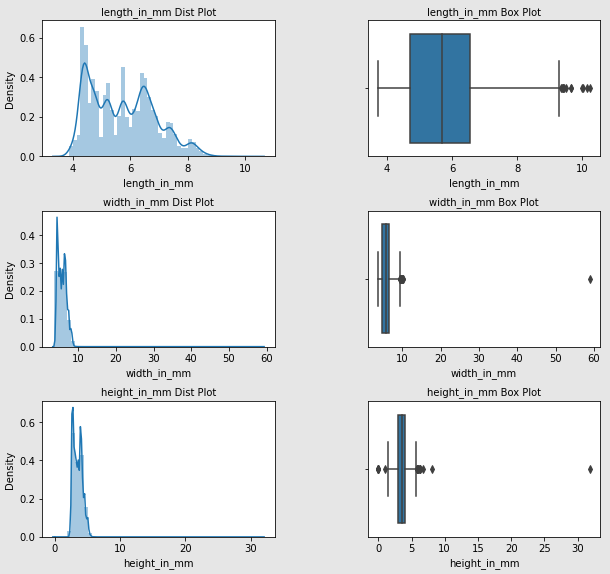


*Fig 2. Univariate Analysis – Count Plots*

From the above count plots we can observe that:

* Ideal is the most preferred type of diamond cutting
* G is the color that is used the most and J the least
* SI1 and VS2 have the most clarity





*Fig 3. Box and Distribution plots*

The above box plot shows that there are many outliers in the dataset

Distribution is right-skewed for carat, length\_in\_mm and it is normal from width\_in\_mm, height\_in\_mm and table variables.

Bivariate Analysis

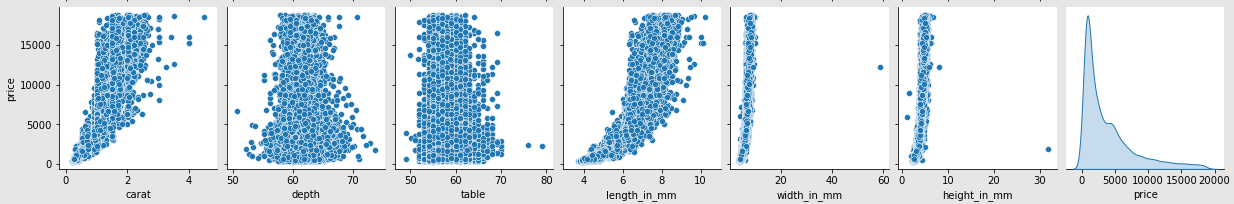
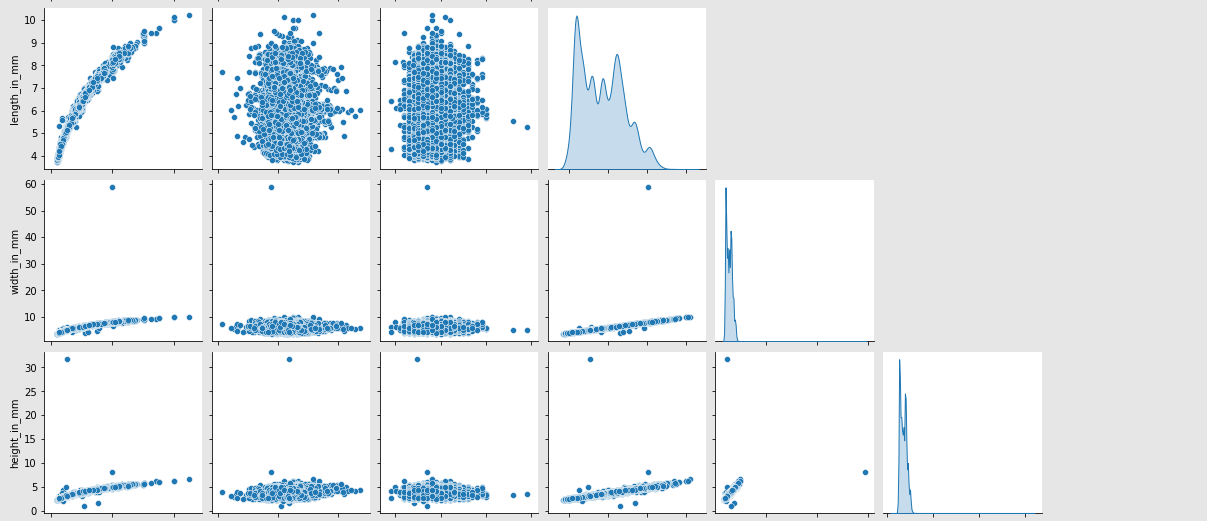
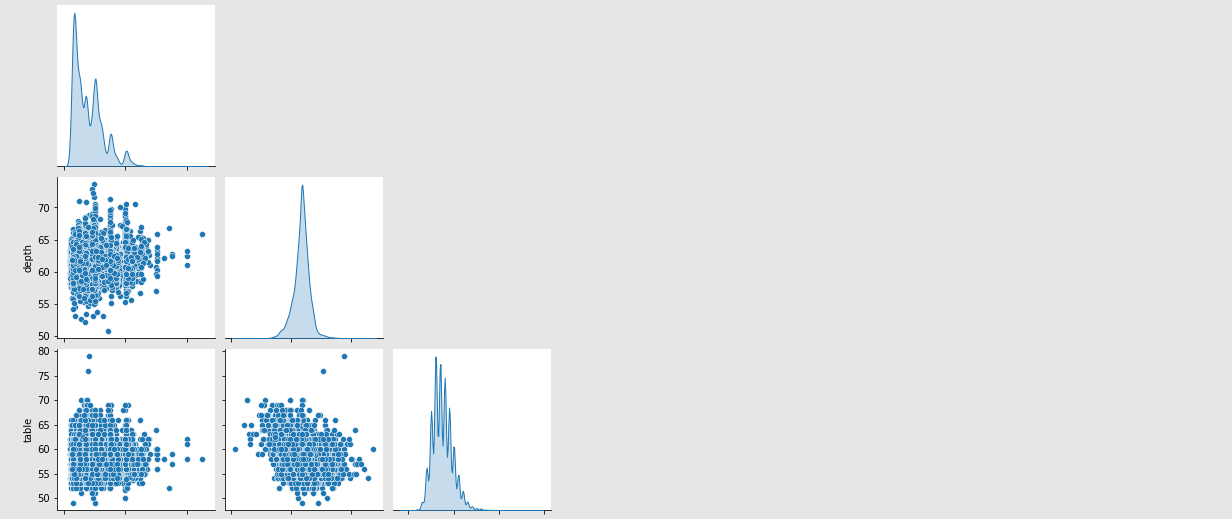
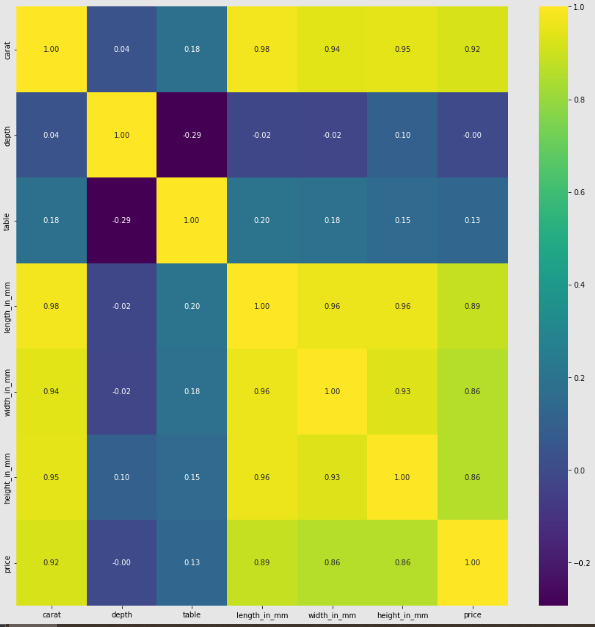


Fig 4. Pair Plot

It can be inferred from the above pair plot that price has a good relationship with carat, depth and length\_in\_mm. Similarly the measures are also well correlated with each other.

Multivariate Analysis



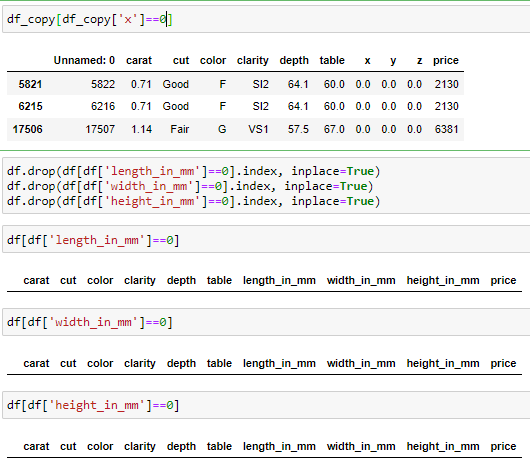
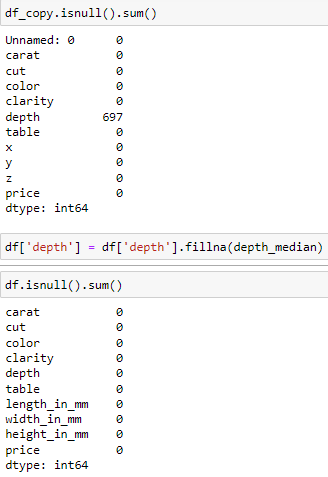
*Fig 5. Correlation Heatmap*

The multivariate analysis shows us that the height, width and length od diamonds are highly correlated to the carat variable and also a multicollinearity is present with them. The coefficient of correlation between price and carat is 0.92 which indicates that it is a strong positive one.

**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.**

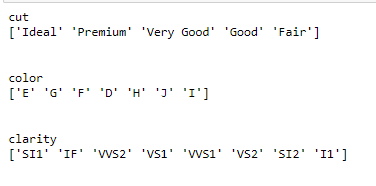
Null values were present in the depth variable, hence imputing them with the median values.

0’s were present in the length, width and height variables, which do not make any meaning to have 0 as a measure of diamond. This could be bad data. Hence dropping those values from the dataset.



*Fig 6. Imputing null values and dropping 0’s*

**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.**

 feature: cut

[1, 2, 3, 4, 5]

Categories (5, int64): [1, 2, 3, 4, 5]

[0 1 2 3 4]

feature: color

[2, 4, 3, 1, 5, 7, 6]

Categories (7, int64): [1, 2, 3, 4, 5, 6, 7]

[1 3 2 0 4 6 5]

feature: clarity

[1, 7, 6, 3, 5, 4, 2, 8]

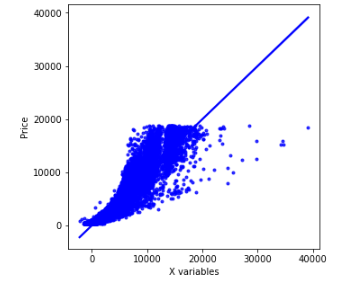
Categories (8, int64): [1, 2, 3, 4, 5, 6, 7, 8]

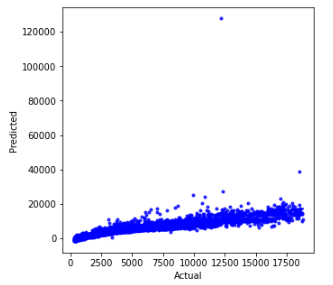
[0 6 5 2 4 3 1 7]

*Fig 7. Encoding the string variables*

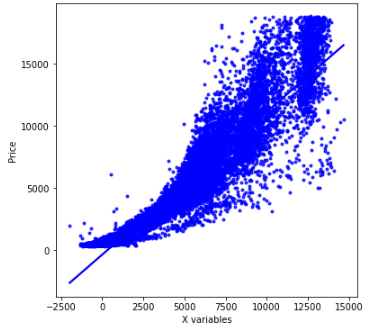
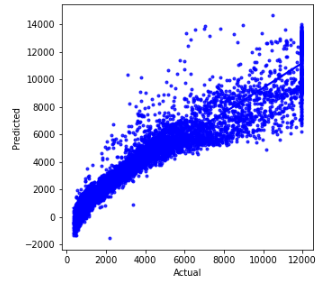
The string variables cut, color and clarity have been encoded in ascending order from worst to best because we cannot do linear regression with string variables.

Linear regression using scikit learn Linear regression using statsmodel



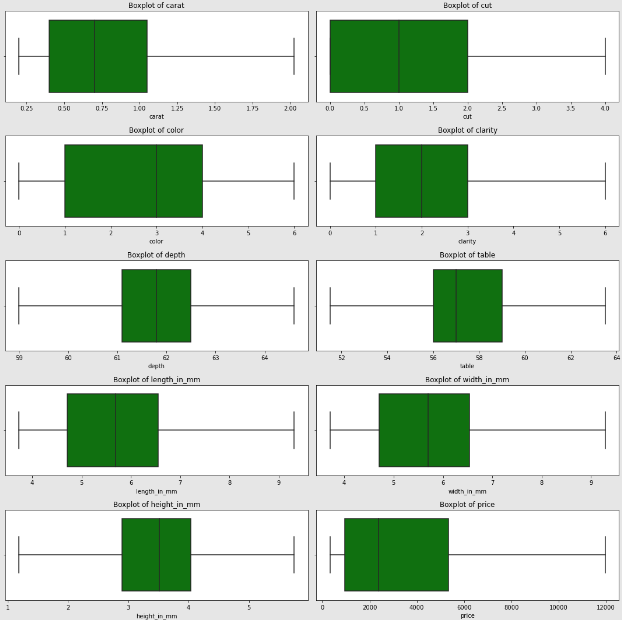


*Fig 8. Linear Regression before outlier treatment* Linear regression after outlier treatment (scikit learn) (statsmodel)



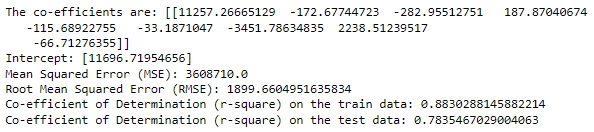
*Fig 9. Linear Regression before outlier treatment* Linear regression after outlier treatment

(scikit learn) (statsmodel)

 Fig 10. Box Plot after outlier treatment

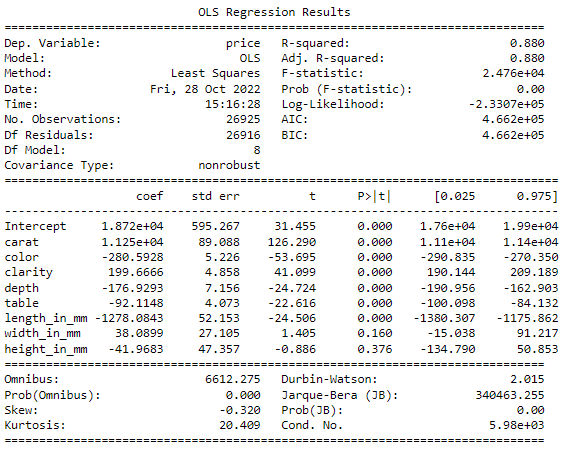
Summaries:

Model1

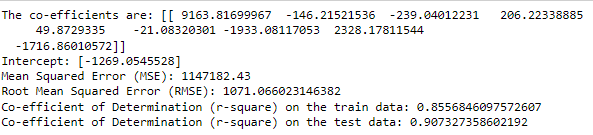


*Fig 11. Model 1 Summary*

Model2

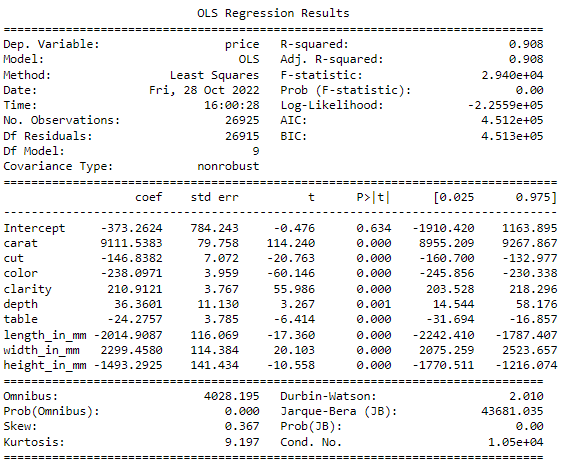


*Fig 12. Model 2 Summary*

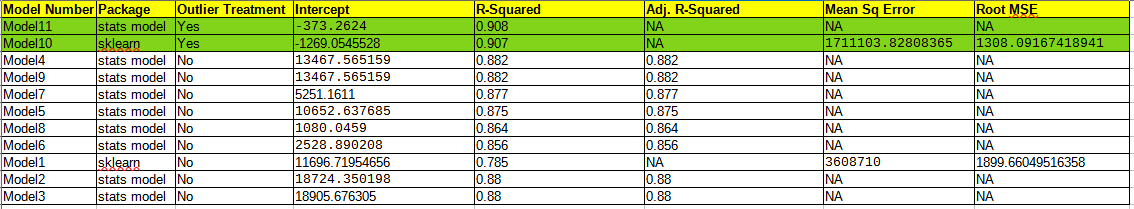


*Fig 13. Model 10 Summary*

Model11

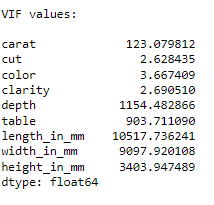


*Fig 14. Model 11 Summary*

*Fig 15. Complete summary*

From the above figures we can infer that the r-squared values have improved for the models after outlier treatment. After outlier treatment, Model 10 and Model 11 have seemed to perform better after testing various models, parameters and formulas.

Checking multi-collinearity with Variance Inflation Formula (VIF)



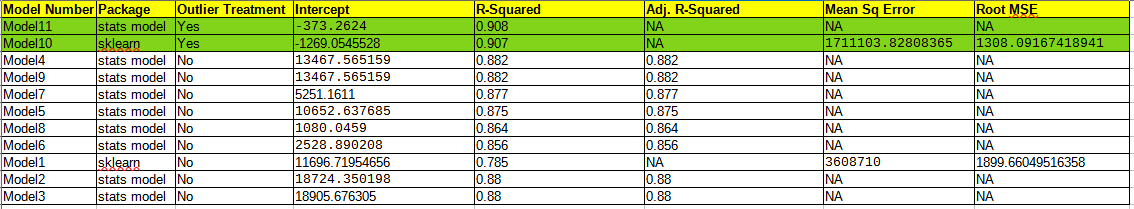
*Fig 16. VIF Values*

The above VIF values indicate that cut, color and clarity have high multi-collinearity.

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

Please explain and summarise the various steps performed in this project. There should be proper business interpretation and actionable insights present.

* EDA was performed by doing univariate analysis where the distribution of each variable was studied, bivariate analysis where the pairplot was reviewed and finally doing a multivariate analysis to study the correlation between the variables.
* The data was split into train and test and was analysed for various factors.
* This was further studied by using scikit learn and statsmodel methods to find the r-squared value, adjusted r-squared, SME and RSME values to find out which model performed the best.



* All model performances were comparable initially. After treating outliers the models have performed better. This indicates that outliers affect prediction of models.
* Fig 9 indicates that there is a high linear relationship between actual and predicted values
* R-squared value of 90% indicates that the accuracy is good and model 10 and 11 have outperformed the other models.
* There is a high-multi collinearity in the dataset which has been observed from the VIF values.
* Carat, color, cut, clarity, length\_in\_mm, width\_in\_mm, height\_in\_mm, depth and table influence the price variable.
* Increase in carat weight will increase the price of diamond.
* Width, height and length of diamonds also influence price in a great way.
* Better colour diamonds are priced better.
* Data collection could have been more appropriate because, there were some bad data in the x, y and z variables. These records were deleted because they do not add any value to the size of diamonds and hence it might not affect the price variable. The business should try and capture better data in the future as it does not create a good report while forecasting high level of prices and business in diamonds.

**Problem 2**

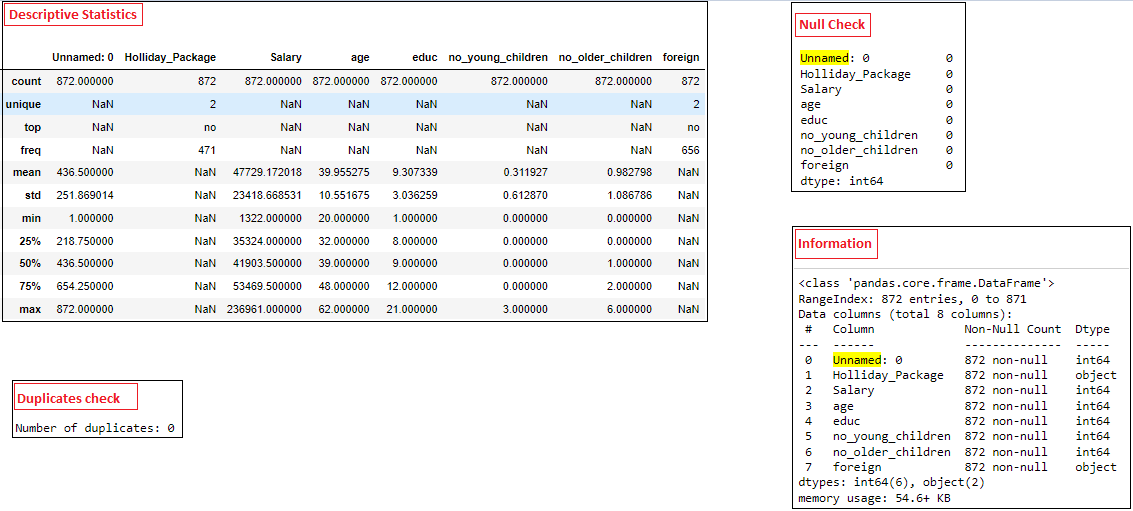
**2. Problem Statement**

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 |

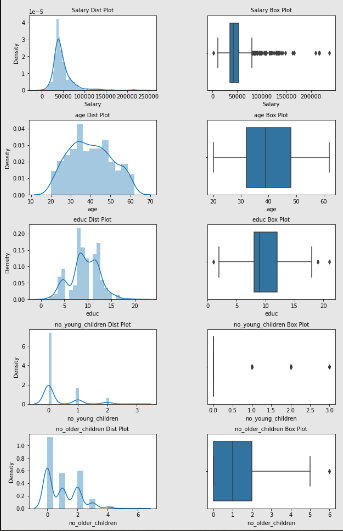
2.1 Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, write an inference on it. Perform Univariate and Bivariate Analysis. Do exploratory data analysis.



*Fig 17. Descriptive statistics Dataset 2*

* Unnamed:0 Column will be removed as it does not add any value for further analysis
* There are no null values in the dataset.
* There are no duplicates in the dataset.
* 2 variables are of object datatype and the rest are integer datatype.
* Out of 872 employees, 471 have not opted for holiday package
* Age of employees range from 20 to 62 years. The average age is 39 years. The median is also 39, hence this might be normally distributed.
* Out of 872 employees, 656 are not foreign nationals.

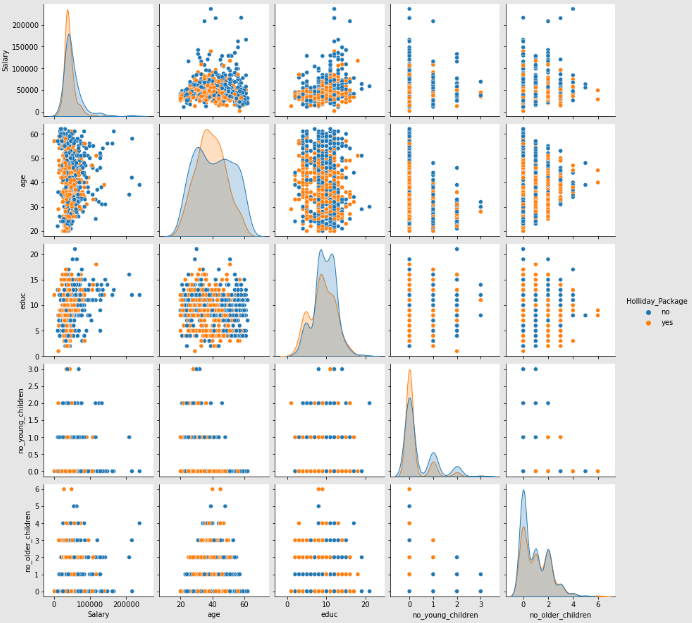
Univariate Analysis



*Fig 18. Univariate Analysis Box Plot and Histogram*

* Salary, no of older children and no of younger children attributes are right skewed.
* Education and Age follow a normal distribution.
* There are many outliers present.

Bivariate Analysis



*Fig 19. PairPlot Holiday Package*

The pairplot does not show any correlation between the variables.

Multivariate Analysis

Correlation Heatmap

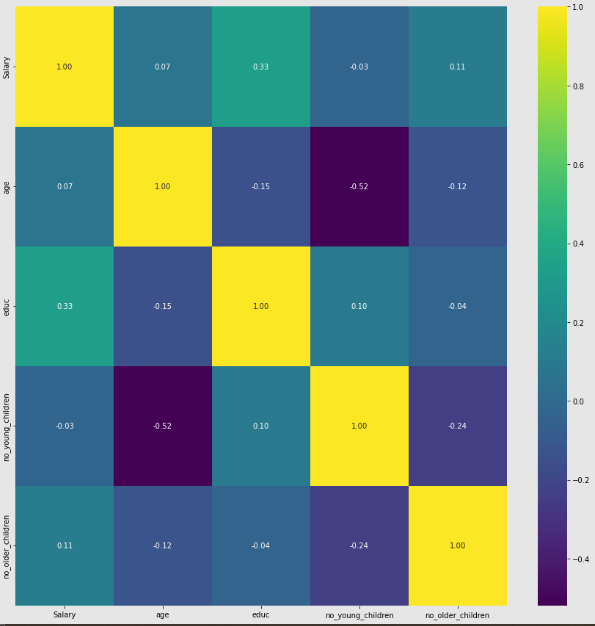


Fig 20. Correlation Heatmap Holiday Package

The above heatmap shows that there is absolutely no correlation between the attributes.

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).

Foreign and Holliday\_Package attributes have been encoded to 0’s and 1’s for No and Yes values respectively.

Train Test Split

X\_train (610, 6)

X\_test (262, 6)

y\_train (610, 1)

y\_test (262, 1)

After applying Logistic Regression and LDA, the coefficients of the attributes are as follows:

Coefficients of Logistic Regression

The coefficient for Salary is -1.749001897075807e-05

The coefficient for age is -0.054807870552790135

The coefficient for educ is 0.0781270667793458

The coefficient for no\_young\_children is -1.5489852417257366

The coefficient for no\_older\_children is -0.05724887090716496

The coefficient for foreign is 1.587936588582022

Coefficients of Linear Discriminant Analysis

The coefficient for Salary is -1.4754954809881397e-05

The coefficient for age is -0.05430378306113369

The coefficient for educ is 0.07596537387390198

The coefficient for no\_young\_children is -1.4285464350098698

The coefficient for no\_older\_children is -0.04635929801474014

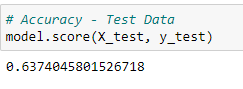
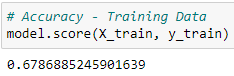
The coefficient for foreign is 1.6239034671206722

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

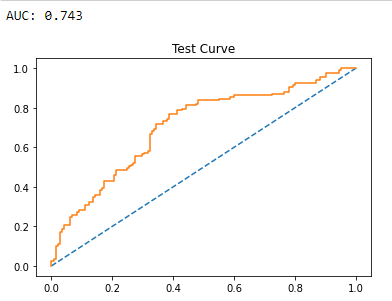
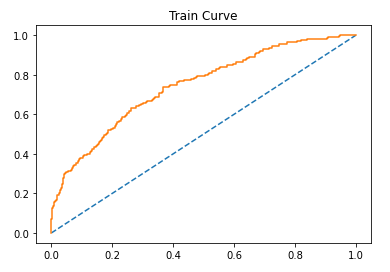
***Logistic Regression***

Newton-cg solver

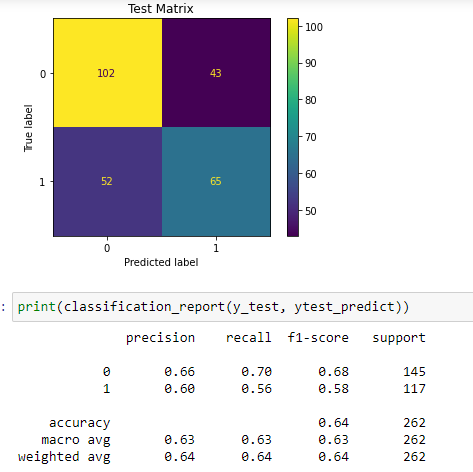
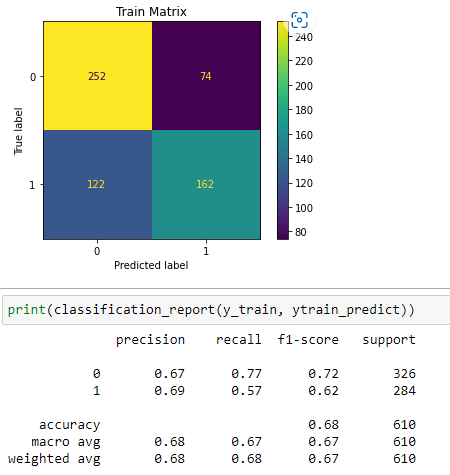
Accuracy:



ROC Curve: AUC – 0.743



*Fig 21. ROC Curve on Train Data Fig 22. ROC Curve on Test Data*

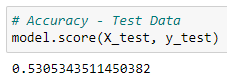
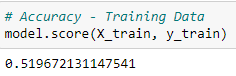
**

*Fig 23. Classification Report and Fig 24. Classification Report and*

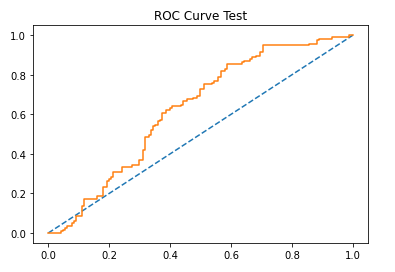
*Confusion Matrix on Train Data Confusion Matrix on Test Data*

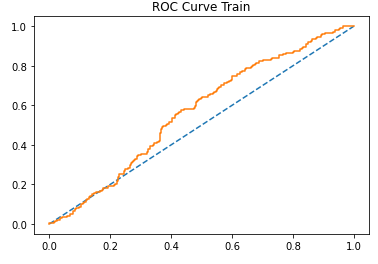
lbfgs solver

Accuracy:

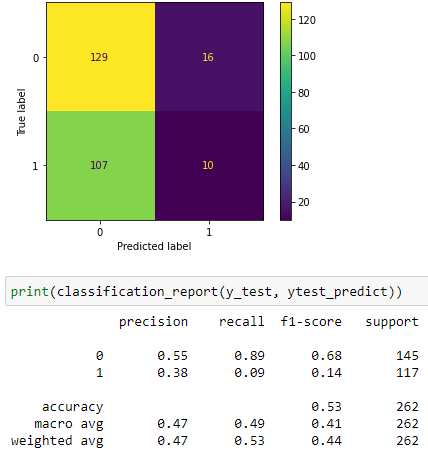
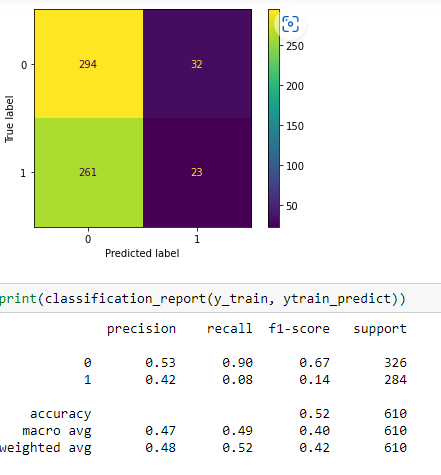


ROC Curve: AUC – 0.567





*Fig 25. ROC Curve on Train Data* l*bfgs solver Fig 26. ROC Curve on Test Data lbfgs solver*

**

*Fig 27. Classification Report and Fig 28. Classification Report and*

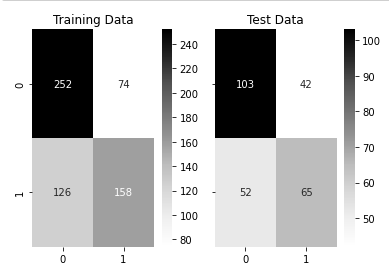
*Confusion Matrix on Train Data Confusion Matrix on Test Data*

*lbfgs solver lbfgs solver*

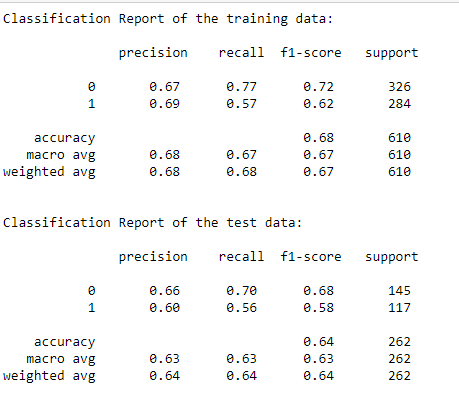
From the above two solver models – newton-cg and lbfgs solver, we can interpret that the newton-cg model has performed better with an accuracy of 67% and 63% on train and test datarespectively when compared to 51% and 53% on lbfgs solver model. Same goes well with the AUC score also.

***Linear discriminant Analysis***

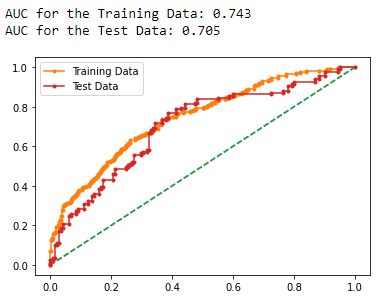
Confusion Matrix



*Fig 29. LDA confusion Matrix*

Classification report – Accuracy – 0.68

*Fig 30. LDA Classification Report*

**

*Fig 31. ROC Curve for LDA*

Observation:

On comparing logistic regression newton-cg model and LDA, both have performed almost comparably similar. But with just over 65% accuracy, we cannot conclude that either of the models have given the best output. The rest 35% is where the model needs to improve.

Treating the outliers can probably produce better results, but with such a small dataset, treating outliers may remove a chunk of the data and that is not advisable. Hence we are not treating outliers in this case.

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

Please explain and summarise the various steps performed in this project. There should be proper business interpretation and actionable insights present.

* From the correlation chart, it is evident that employees with young children do not prefer to go on holidays, and hence the business can try to provide special packages for families travelling with small kids. This way they can attract more families.
* Age of the employee does not seem to have much impact on employee choosing a holiday package, and hence this variable can be avoided in the future.
* Foreign employees with no children seem to opt more for this holiday package.
* The model predicts 65% if an employee will choose a holiday package or not. 35% of the model still needs improvement. This shows that collection of more data is required. This can also be done from other customer organizations. This can help to predict the models better.
* The company is probably way too expensive with their holiday packages, so an offer or discount once in a while may improve their business.
* The coefficient of salary is very low when compared to other attributes. This probably does not play a major role in the employees choosing a holiday package.