28 November 2021

Predictive Modeling Project

PGP-DSBA June-Batch

SOOREJ KUNNOOL BALAN

2021

Table of Contents

[1. Problem 1: Linear Regression 4](#_Toc89025996)

[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. 4](#_Toc89025997)

[Data Description 4](#_Toc89025998)

[Sample of the dataset: 4](#_Toc89025999)

[Exploratory Data Analysis: 5](#_Toc89026000)

[1.2 Impute null values if present, also check for the values which are equal to zero. Do they have any meaning, or do we need to change them or drop them? Check for the possibility of combining the sub levels of an ordinal variables and take actions accordingly. Explain why you are combining these sub levels with appropriate reasoning. 18](#_Toc89026001)

[1.3 Encode the data (having string values) for Modelling. Split the data into train and test (70:30). Apply Linear regression using scikit learn. Perform checks for significant variables using appropriate method from statsmodel. Create multiple models and check the performance of Predictions on Train and Test sets using Rsquare, RMSE & Adj Rsquare. Compare these models and select the best one with appropriate reasoning. 21](#_Toc89026002)

[1.4 Inference: Basis on these predictions, what are the business insights and recommendations. 26](#_Toc89026003)

[2. Problem 2: Logistic Regression and LDA 27](#_Toc89026004)

[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. 27](#_Toc89026005)

[Data Description 27](#_Toc89026006)

[Sample of the dataset: 27](#_Toc89026007)

[Exploratory Data Analysis: 28](#_Toc89026008)

[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). 40](#_Toc89026009)

[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. 41](#_Toc89026010)

[2.4 Inference: Basis on these predictions, what are the insights and recommendations. 44](#_Toc89026011)

[THE END 45](#_Toc89026012)

List of Tables

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Table-1.1 Dataset Sample | 4 |  |  | |  |
| Table-1.2 Concise data summary | 5 |  |  | |  |
| Table-1.3 Duplicate records | 5 |  |  | |  |
| Table-1.4 Updated Concise data summary | 6 |  |  | |  |
| Table-1.5 Data Summary | 6 |  |  | |  |
| Table-1.6 Correlation matrix | 15 |  |  | |  |
| Table-1.7 Null values | 18 |  |  | |  |
| Table-1.8 Data summary | 19 |  |  | |  |
| Table-1.9 Data with 0 value | 19 |  |  | |  |
| Table-1.10 Concise Data summary | 19 |  |  | |  |
| Table-1.11 ‘cut’ distribution | 20 |  |  | |  |
| Table-1.12 ‘color’ distribution | 21 |  |  | |  |
| Table-1.13 ‘clarity’ distribution | 21 |  |  | |  |
| Table-1.14 scaled data- summary | 22 |  |  | |  |
| Table-1.15 concise data- summary | 22 |  |  | |  |
| Table-1.16 concise data- summary train/test | 22 |  |  | |  |
| Table-1.17 model coefficient | 23 |  |  | |  |
| Table-1.18 ols summary | 23 |  |  | |  |
| Table-1.19 model1 vif | 24 |  |  | |  |
| Table-1.20 model coefficient | 24 |  |  | |  |
| Table-1.21 ols model summary | 25 |  |  | |  |
| Table-1.22 model2 vif | 25 |  |  | |  |
| Table-1.23 comparison of model metrics | 26 |  |  | |  |
| Table-2.1 Dataset Sample | 27 |  |  | |  |
| Table-2.2 Concise data summary | 28 |  |  | |  |
| Table-2.3 Data Summary | 28 |  |  |  | |
| Table-2.4 Correlation matrix | 34 |  |  |  | |
| Table-2.5 Crosstab: foreign vs no\_young\_children | 39 |  |  |  | |
| Table-2.6 Crosstab: domestic vs no\_young\_children | 39 |  |  |  | |
| Table-2.7 Crosstab: foreign vs no\_older\_children | 40 |  |  |  | |
| Table-2.8 Crosstab: domestic vs no\_older\_children | 40 |  |  |  | |
| Table-2.9 Data head of encoded data | 40 |  |  |  | |
| Table-2.10 Train/Test target variable data | 41 |  |  |  | |
| Table-2.11 Logistic Regression – confusion matrix(training data) | 42 |  |  |  | |
| Table-2.12 Logistic Regression – confusion matrix(test data) | 42 |  |  |  | |
| Table-2.13 Classification Report: Logistic Regression -training data | 42 |  |  |  | |
| Table-2.14 Classification Report: Logistic Regression -test data | 42 |  |  |  | |
| Table-2.15 LDA – confusion matrix(training data) | 42 |  |  |  | |
| Table-2.16 LDA – confusion matrix(test data) | 42 |  |  |  | |
| Table-2.17 Classification Report: LDA-training data | 43 |  |  |  | |
| Table-2.18 Classification Report: LDA -test data | 43 |  |  |  | |
| Table-2.19 Model comparison | 43 |  |  |  | |

List of Figures

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Figure-1.1 Histogram & Boxplot : carat | 7 |  |  | |  |
| Figure-1.2 Histogram & Boxplot : depth | 8 |  |  | |  |
| Figure-1.3 Histogram & Boxplot : table | 9 |  |  | |  |
| Figure-1.4 Histogram & Boxplot : x | 10 |  |  | |  |
| Figure-1.5 Histogram & Boxplot : y | 11 |  |  | |  |
| Figure-1.6 Histogram & Boxplot : z | 12 |  |  | |  |
| Figure-1.7 Histogram & Boxplot : price | 13 |  |  | |  |
| Figure-1.8 Count-plot & Pie-chart : price | 13 |  |  | |  |
| Figure-1.9 Count-plot & Pie-chart : color | 14 |  |  | |  |
| Figure-1.10 Count-plot & Pie-chart : clarity | 15 |  |  | |  |
| Figure-1.11 Pairplot | 16 |  |  | |  |
| Figure-1.12 cut vs price | 17 |  |  | |  |
| Figure-1.13 color vs price | 17 |  |  | |  |
| Figure-1.14 clarity vs price | 18 |  |  | |  |
| Figure-1.15 cut vs color/clarity | 20 |  |  | |  |
| Figure-1.16 color vs cut/clarity | 20 |  |  | |  |
| Figure-1.17 clarity vs cut/color | 21 |  |  | |  |
| Figure-2.1 Histogram & Boxplot : Salary | 29 |  |  | |  |
| Figure-2.2 Histogram & Boxplot : age | 30 |  |  | |  |
| Figure-2.3 Histogram & Boxplot : educ | 31 |  |  | |  |
| Figure-2.4 Histogram & Boxplot : no\_young\_children | 32 |  |  | |  |
| Figure-2.5 Histogram & Boxplot : no\_older\_children | 33 |  |  | |  |
| Figure-2.6 Count-plot & Pie-chart : Holliday\_Package | 33 |  |  | |  |
| Figure-2.7 Count-plot & Pie-chart : foreign | 34 |  |  | |  |
| Figure-2.8 Pairplot | 35 |  |  | |  |
| Figure-2.9 Countplot : foreign vs Holliday\_Package | 36 |  |  |  | |
| Figure-2.10 Barplot : foreign vs Salary/Holliday\_Package | 36 |  |  |  | |
| Figure-2.11 Scatterplot: age vs Salary/Holliday\_Package | 37 |  |  |  | |
| Figure-2.12 Barplot: foreign vs educ/Holliday\_Package | 37 |  |  |  | |
| Figure-2.13 Scatterplot: age vs educ/Holliday\_Package | 38 |  |  |  | |
| Figure-2.14 Countplot: no\_young\_children | 38 |  |  |  | |
| Figure-2.15 Countplot: no\_young\_children | 39 |  |  |  | |
| Figure-2.16 Countplot: no\_older\_children | 39 |  |  |  | |
| Figure-2.17 Countplot: no\_young\_children | 40 |  |  |  | |
| Figure-2.18 ROC Curve: Logistic Regression -training data | 42 |  |  |  | |
| Figure-2.19 ROC Curve: Logistic Regression -test data | 42 |  |  |  | |
| Figure-2.20 ROC Curve: LDA -training data | 43 |  |  |  | |
| Figure-2.21 ROC Curve: LDA -test data | 43 |  |  |  | |

# Problem 1: Linear Regression

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.



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

### Data Description

1. carat : Carat weight of the cubic zirconia (continuous variable)
2. cut : Describe the cut quality of the cubic zirconia. Quality is increasing order Fair, Good, Very Good, Premium, Ideal.
3. color : Color of the cubic zirconia. With D being the worst and J the best. Values present are D,E,F,G,H,I and J
4. clarity : Clarity refers to the absence of the Inclusions and Blemishes. (In order from Worst to Best in terms of avg price) IF, VVS1, VVS2, VS1, VS2, Sl1, Sl2, l1
5. depth : The Height of cubic zirconia, measured from the Culet to the table, divided by its average Girdle Diameter.(continuous variable)
6. table : The Width of the cubic zirconia's Table expressed as a Percentage of its Average Diameter (continuous variable).
7. price : the Price of the cubic zirconia (continuous variable).
8. x : Length of the cubic zirconia in mm (continuous variable).
9. y : Width of the cubic zirconia in mm (continuous variable).
10. z : Height of the cubic zirconia in mm (continuous variable).

### Sample of the dataset:

We can remove the first column in the csv file, which is a serial number column. The dataset sample is as shown below:

A screenshot of a computer

Description automatically generated with medium confidence

Table-1.1 Dataset Sample

### Exploratory Data Analysis:

Text

Description automatically generated

Table-1.2 Concise data summary

#### Let us check the type of variables in the data frame

There are a total of 26967 observations and 10 columns in the dataset. We have 3 columns of object type, 6 of float64 and 1 of int64 type.

#### Check for missing values in the dataset

From Table-1.2 we can see that all the columns except depth have 26967 non-null values. We have 697 null columns in depth variable.

#### Check for duplicate observations in the dataset

There are 34 duplicate records as shown below:

Table

Description automatically generated with medium confidence

Table-1.3 Duplicate records

We can delete these 34 duplicate records. The updated concise data summary is as follows:

Text

Description automatically generated

Table-1.4 Updated Concise data summary

#### Data summary

Table

Description automatically generated

Table-1.5 Data Summary

The above data summary will be further explained in the univariate analysis section below.

#### Univariate Analysis:

Let’s check the central measures of tendency, quartiles, histogram, and boxplot of all 7 continuous columns.

1. carat

Carat weight of the cubic zirconia is a continuous variable with the below stats (refer Table 1.5):

Mean = 0.798010

Standard Deviation = 0.477237

Min value in dataset = 0.2

Max value in dataset = 4.50

Range = Min – Max = 4.30

Q1(1st Quartile) = 0.40

Q2(2nd Quartile)/Median = 0.70

Q3(3rd Quartile) = 1.05

IQR(Inter-Quartile Range) = Q3- Q1 = 0.65

Quartile Min value = max(Qmin,Q1 – 1.5 \* IQR) = 0.2

Quartile Max value = min(Qmax,Q3 + 1.5 \* IQR) = 2.025

Chart

Description automatically generated

Figure-1.1 Histogram & Boxplot : carat

Figure-1.1 depicts the histogram and boxplot of “carat” which shows positive skewness in the data. The data depicts the presence of multi-modes as we can see multiple peaks in the histogram.

From the boxplot we can see that there are outliers present in ‘carat’. This does not seem to point to erroneous data and is expected.

1. depth

The height of cubic zirconia, measured from the culet to the table, divided by its average girdle diameter is a continuous variable with the below stats (refer Table 1.5):

Mean = 61.745285

Standard Deviation = 1.412243

Min value in dataset = 50.8

Max value in dataset = 73.6

Range = Min – Max = 22.8

Q1(1st Quartile) = 61

Q2(2nd Quartile)/Median = 61.8

Q3(3rd Quartile) = 62.5

IQR(Inter-Quartile Range) = Q3- Q1 = 1.5

Quartile Min value = max(Qmin,Q1 – 1.5 \* IQR) = 58.75

Quartile Max value = min(Qmax,Q3 + 1.5 \* IQR) = 64.75

Chart, box and whisker chart

Description automatically generated

Figure-1.2 Histogram & Boxplot : depth

Figure-1.2 depicts the histogram and boxplot of “depth” which indicates normal distribution.

From the boxplot we can see that there are outliers present in ‘depth’. This does not seem to point to erroneous data and is expected.

1. table

The width of the cubic zirconia's table expressed as a percentage of its average diameter, is a continuous variable with the below stats (refer Table 1.5):

Mean = 57.455950

Standard Deviation = 2.232156

Min value in dataset = 49

Max value in dataset = 79

Range = Min – Max = 30

Q1(1st Quartile) = 56

Q2(2nd Quartile)/Median = 57

Q3(3rd Quartile) = 59

IQR(Inter-Quartile Range) = Q3- Q1 = 3

Quartile Min value = max(Qmin,Q1 – 1.5 \* IQR) = 51.5

Quartile Max value = min(Qmax,Q3 + 1.5 \* IQR) = 63.5

Chart, box and whisker chart

Description automatically generated

Figure-1.3 Histogram & Boxplot : table

Figure-1.1 depicts the histogram and boxplot of “depth” which shows positive skewness in the data. The data depicts the presence of multi-modes as we can see multiple peaks in the histogram.

From the boxplot we can see that there are outliers present in ‘depth’. This does not seem to point to erroneous data and is expected.

1. x

Length of the cubic zirconia in mm is a continuous variable with the below stats (refer Table 1.5):

Mean = 5.729346

Standard Deviation = 1.127367

Min value in dataset = 0

Max value in dataset = 10.23

Range = Min – Max = 10.23

Q1(1st Quartile) = 4.71

Q2(2nd Quartile)/Median = 5.69

Q3(3rd Quartile) = 6.55

IQR(Inter-Quartile Range) = Q3- Q1 = 1.84

Quartile Min value = max(Qmin,Q1 – 1.5 \* IQR) = 1.950

Quartile Max value = min(Qmax,Q3 + 1.5 \* IQR) = 9.310

Chart, histogram, box and whisker chart

Description automatically generated

Figure-1.4 Histogram & Boxplot : x

Figure-1.4 depicts the histogram and boxplot of “x” which shows positive skewness in the data. The data depicts the presence of multi-modes as we can see multiple peaks in the histogram.

From the boxplot we can see that there are outliers present in ‘x’. This does not seem to point to erroneous data and is expected.

1. y

Width of the cubic zirconia in mm is a continuous variable with the below stats (refer Table 1.5):

Mean = 5.733102

Standard Deviation = 1.165037

Min value in dataset = 0

Max value in dataset = 58.9

Range = Min – Max = 58.9

Q1(1st Quartile) = 4.71

Q2(2nd Quartile)/Median = 5.70

Q3(3rd Quartile) = 6.54

IQR(Inter-Quartile Range) = Q3- Q1 = 1.83

Quartile Min value = max(Qmin,Q1 – 1.5 \* IQR) = 1.965

Quartile Max value = min(Qmax,Q3 + 1.5 \* IQR) = 9.285

Graphical user interface, chart

Description automatically generated

Figure-1.5 Histogram & Boxplot : y

Figure-1.5 depicts the histogram and boxplot of “y” which shows positive skewness in the data. The data depicts the presence of multi-modes as we can see multiple peaks in the histogram.

From the boxplot we can see that there are outliers present in ‘y’, but less in number. This does not seem to point to erroneous data and is expected, though one observation has a suspiciously high value.

1. z

Height of the cubic zirconia in mm is a continuous variable with the below stats (refer Table 1.5):

Mean = 3.537769

Standard Deviation = 0.719964

Min value in dataset = 0

Max value in dataset = 31.80

Range = Min – Max = 31.80

Q1(1st Quartile) = 2.9

Q2(2nd Quartile)/Median = 3.52

Q3(3rd Quartile) = 4.04

IQR(Inter-Quartile Range) = Q3- Q1 = 1.14

Quartile Min value = max(Qmin,Q1 – 1.5 \* IQR) = 1.190

Quartile Max value = min(Qmax,Q3 + 1.5 \* IQR) = 5.750

Graphical user interface, application

Description automatically generated

Figure-1.6 Histogram & Boxplot : z

Figure-1.6 depicts the histogram and boxplot of “z” which shows positive skewness in the data. The data depicts the presence of multi-modes as we can see multiple peaks in the histogram.

From the boxplot we can see that there are outliers present in ‘z’. This does not seem to point to erroneous data and is expected, though one observation has a suspiciously high value.

1. price

The price of the cubic zirconia is a continuous variable with the below stats (refer Table 1.5):

Mean = 3937.526120

Standard Deviation = 4022.551862

Min value in dataset = 326

Max value in dataset = 18818

Range = Min – Max = 18492

Q1(1st Quartile) = 945

Q2(2nd Quartile)/Median = 2375

Q3(3rd Quartile) = 5356

IQR(Inter-Quartile Range) = Q3- Q1 = 4411

Quartile Min value = max(Qmin,Q1 – 1.5 \* IQR) = 326

Quartile Max value = min(Qmax,Q3 + 1.5 \* IQR) = 11972.5

Chart, box and whisker chart

Description automatically generated

Figure-1.7 Histogram & Boxplot : price

Figure-1.7 depicts the histogram and boxplot of “price” which shows positive skewness in the data. The data depicts the presence of multi-modes as we can see multiple peaks in the histogram.

From the boxplot we can see that there are outliers present in ‘price’. This does not seem to point to erroneous data and is expected.

Let’s look at the categorical variables:

1. cut

Describes the cut quality of the cubic zirconia. Quality is increasing order Fair, Good, Very Good, Premium, Ideal. Let’s look at the count-plot for cut variable and the pie-chart for % distribution.

Chart, pie chart

Description automatically generated

Figure-1.8 Count-plot & Pie-chart : price

The above plot shows that the cut of products offered have the below counts.

‘Ideal’ (10805) > ‘Premium’ (6886) > ‘Very Good’ (6027) > ‘Good’ (2435) > ‘Fair’ (780)

We can see that the company has positioned to provide the best quality cuts, with Ideal and Premium accounting for 65% of the products.

1. color

Color of the cubic zirconia, with D being the worst and J the best. Let’s look at the count-plot for color variable and the pie-chart for % distribution.

Chart, pie chart

Description automatically generated

Figure-1.9 Count-plot & Pie-chart : color

The above plot shows that the color of products offered have the below counts.

‘G’ (5653) > ‘E’ (4916) > ‘F’ (4723) > ‘H’ (4095) > ‘D’ (3341) > ‘I’ (2765) > ‘J’ (1440)

We can see that the best color ‘J’ has the least count, and intermediate quality color ‘G’ has the maximum count among the provided data.

1. clarity

Clarity refers to the absence of the Inclusions and Blemishes. (In order from Worst to Best in terms of avg price) IF, VVS1, VVS2, VS1, VS2, Sl1, Sl2, l1.

The below plot shows that the clarity of products offered have the below counts.

‘SI1’ (6565) > ‘VS2’ (6093) > ‘SI2’ (4564) > ‘VS1’ (4087) > ‘VVS2’ (2530) > ‘VVS1’ (1839) > ‘IF’ (891) > ‘I1’ (364)

We can see that the best clarity ‘I1’ has the least count, and intermediate clarity ‘SI1’ has the maximum count among the provided data.

Chart, pie chart

Description automatically generated

Figure-1.10 Count-plot & Pie-chart : clarity

#### Outlier Treatment:

We have outliers present in almost all the 7 numerical variables. But these do not seem to be erroneous data and removing them might result in overfitting, causing the model to perform abysmally against new test data. Ideally for the suspiciously high observation in y and z column, needs to be further clarified with business for veracity.

#### Bivariate/Multivariate Analysis:

Let us plot a heat map for the correlation matrix of given data frame.

Chart, histogram

Description automatically generated

Table-1.6 Correlation matrix

We can see very strong positive correlation between the following variables:

* carat against price/x/y/z

The above data shows that we could have the issue of multi-collinearity between carat and x, y and z variables. This does follow reason as carat is an indication of weight and x, y, z indicates the size of the cubic zirconium. We know that gemstones of higher carats are expensive, and the data reflects the same.

* x, y and z

Strong correlation exists between x, y and z. This shows that the shapes of products are sized almost uniformly in scale. When length increases width and depth increases in scale.

* Table and depth

There exists a moderate negative correlation between table and depth.

All other correlations are weak.

Let’s also check the Pairplot:

A picture containing diagram

Description automatically generated

Figure-1.11 Pairplot

Pairplot substantiates the heatmap and displays the correlation between variables as stated above.

Let us also look at interaction of categorical variables against target variable price and check for inferences.

* cut vs price

Let’s look at the below Countplot, barplot and stripplot of cut vs price.

Chart, bar chart

Description automatically generated

Figure-1.12 cut vs price

Interestingly we can see that irrespective of the type of cut, there is an offering at each price point. This should indicate that the sizes of the product must differ between these cuts. Ideal Premium and Very Good shows higher representation at higher price levels. Premium and Fair cut commands highest average price whereas Ideal cut commands the least.

We can see that Ideal, Premium and Very Good cut makes up for the bulk of the products and the average price of Premium and Very Good is substantially higher than that of Ideal. Hence business must have skilled workers/tools to ensure that maximum of the cut is targeted towards Premium/Very Good. Also, we can see that the average price between Very Good/Good is almost the same.

* color vs price

Let’s look at the below Countplot, barplot and stripplot of color vs price.

Chart, bar chart

Description automatically generated

Figure-1.13 color vs price

We can see that color J has the highest average price and color E has the least. We can see that irrespective of color, the products are available at all price points across the colors. If the list of product indicates consumer demand, we can say that color G is preferred by majority due to the lower price. Business can look at ways to entice consumers preferring G/H colors to move to I/J which could lead to increased revenue.

* clarity vs price

Let’s look at the below Countplot, barplot and stripplot of clarity vs price.

Chart, bar chart

Description automatically generated

Figure-1.14 clarity vs price

We can see that Sl2 clarity has the highest average price and VVS1 has the least.

Apart from l1 we have all other clarity stones across the price band.

The average price of SI1,VS2,VS1 and l1 clarity are nearly same.

## 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.

Let’s check for the presence of null values.

A picture containing text

Description automatically generated

Table-1.7 Null values

We can see that there are 697 null values in the depth column. Let’s impute these null values with the median of depth (if we choose, we can delete these records too as they are only about 2.5% of the total records 26933).

After imputing these null records let’s check the data summary.

Table

Description automatically generated

Table-1.8 Data summary

Let’s check the number of records with values as 0.

Text

Description automatically generated

Table-1.9 Data with 0 value

Zeroes are in x, y and z columns. This is erroneous data as these columns depict the size measures of the stones and cannot be zero. Since the occurrence is 8 observations out of 26933, let us delete these observations and proceed.

Let us check the updated concise summary to verify the 8 records deleted.

Text

Description automatically generated

Table-1.10 Concise Data summary

We can see that 8 records have been deleted, count has reduced from 26933 to 26925 and all columns show non-null data.

Let’s look at the possibility of combing sub levels within ordinal variables.

* cut

In Figure 1.12 cut vs price, we have seen that the average price for ‘Very Good’ and ‘Good’ was about same. Similarly, ‘Premium’ and ‘Fair’ too has nearly same average price.

Let’s see the distribution of ‘clarity’ and ‘color’ across cut.

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated

Figure-1.15 cut vs color/clarity

We can see that color and clarity distributed across the cuts ‘Very Good’ and ‘Good’ is comparable as well as ‘Premium’ and ‘Fair’ is also comparable( but the ordinal order of Premium and Fair is not consecutive, so we won’t combine this).

Hence let’s combine ‘Very Good’ and ‘Good’ into ‘Good’ cut.

After combining we have the following data in ‘cut’:

Text, table

Description automatically generated

Table-1.11 ‘cut’ distribution

* color

From figure 1.13 we can see that the average price for ‘D’ and ‘E’ as well as for ‘I’ and ‘J’ are comparable.

Let’s look at how the cut and clarity are distributed among D/E and I/J

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated

Figure-1.16 color vs cut/clarity

ZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZ

We can see that the distribution of cut/clarity follows the same trend between D/E and I/J. So, let us combine ‘D’ and ‘E’ color to ‘E’ and ‘I’,’J’ color to ‘I’.

After combining we have the below color distribution:

Text

Description automatically generated

Table-1.12 ‘color’ distribution

* clarity

The ordinal values in terms of clarity are IF, VVS1, VVS2, VS1, VS2, SI1, SI2, I1.

From figure 1.14 we can see that the average price of VS1,VS2,SI1 and I1. We can see that VS1,VS2 and SI1are adjacent in the ordinal order.

Let’s look at how the cut and color are distributed within VS1,VS2 and SI1.

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated

Figure-1.17 clarity vs cut/color

The trend of cut and color is comparable between VS1,VS2 and Sl1. Hence let’s combine these 3 values into Sl1.

After combining we have the below clarity distribution:

Table

Description automatically generated

Table-1.13 ‘clarity’ distribution

## 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.

Since the continuous variables have different range, lets scale the data using standard scaler.

After scaling let’s check the data summary:

We can see that the mean approaches zero and standard deviation is 1.

Text

Description automatically generated

Table-1.14 scaled data- summary

Let’s get dummy variables for the 3 ordinal columns, and then check the concise summary:

Table

Description automatically generated

Table-1.15 concise data- summary

Let’s split the data with 70% as train and 30% for test and check concise data summary.

Table

Description automatically generated Table

Description automatically generated

Table-1.16 concise data- summary train/test

#### Model 1:

Now we will apply linear regression on the training data and check the coefficient for predictor variables and intercept value.

Text

Description automatically generated

Table-1.17 model coefficient

The intercept for the model is -0.98884601253559.

Let us also create a model via ordinary least square method of statsmodel. The model summary for the

training data is given below:

Table

Description automatically generated

Table-1.18 ols summary

From the above summary we can see that predictors y and z have p-value much higher than 0.05 which indicates that these 2 columns don’t contribute enough towards the determination of target variable. (Model 2 we will create by eliminating y and z columns).

The R2 for the model on training data is 0.9132552823265415, i.e., the model can explain 91.33% of the variation in price.

The R2 for the model on testing data is 0.9170890039419749 i.e., the model can explain 91.71% of

The variation in price.

The RMSE for the model on training data is 0.2935838537381607

The RMSE for the model on training data is 0.29007336898159625

The RMSE is on the lower side and hence prediction accuracy is high.

The Adjusted R2 for the model on training data is 0.9131723523861272.

The Adjusted R2 for the model on testing data is 0.9182084928556183.

The Adjusted R2 values are high, and the model will perform well in accurately predicting price.

Let’s look at the multi-collinearity by checking the variance inflation factor.

Text

Description automatically generated

Table-1.19 model1 vif

#### Model 2:

For this model we will remove ‘y’ and ‘z’ column from the predictors (as was indicated in Table 1.18) as their contribution to target variable is insignificant.

We will apply linear regression on this dataset and check the coefficient for predictor variables and intercept value.

Text

Description automatically generated

Table-1.20 model coefficient

The intercept for the model is -0.988749704920028.

Let us also create a model via ordinary least square method of statsmodel. The model summary for the

training data is given below:

Table

Description automatically generated

Table-1.21 ols model summary

The R2 for the model on training data is 0.9132527153890151, i.e., the model can explain 91.33% of the variation in price.

The R2 for the model on testing data is 0.9170863988659084 i.e., the model can explain 91.71% of

The variation in price.

The RMSE for the model on training data is 0.29358819755215027

The RMSE for the model on training data is 0.2900779260204173

The RMSE is on the lower side and hence prediction accuracy is high.

The Adjusted R2 for the model on training data is 0.9131790055348581.

The Adjusted R2 for the model on testing data is 0.9175072052200024.

The Adjusted R2 values are high, and the model will perform well in accurately predicting price.

Let’s look at the multi-collinearity by checking the variance inflation factor.

Text

Description automatically generated

Table-1.22 model2 vif

We can see high vif values for ‘carat’ and column ‘x’, which is highly probable as the heat map shows

high collinearity between carat and ‘x’. We need to understand from business how carat is related to x

and if we could remove x and create a model.

Let’s look at the performance parameters of the 2 models from the table given below:

Table

Description automatically generated

Table-1.23 comparison of model metrics

In the above table ‘lm1’ and ‘lm2’ in column names refer to the model 1 and model 2 respectively.

Both models have comparable performance metrics, the changes are minimal.

We can see that RMSE values are lesser in model1 compared to model 2, and for the test dataset the

model 1’s adjusted R2 value is slightly higher than that of model2. So, we can see that model1 has lower

root mean square error and accuracy slightly higher than model2.

So, we can decide that model 1 is better choice, as it has better accuracy to predict the target price variable.

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

Gem Stones co ltd, which is a cubic zirconia manufacturer have provided us with a 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 and needs help in predicting the price for the stone on the bases of the details given in the dataset so it can distinguish between higher profitable stones and lower profitable stones to have better profit share.

To achieve this objective, we started with exploratory data analysis of the given dataset. We performed univariate and bivariate analysis to determine how the predictors influence the target variable. We performed linear regression on the data and have come up with 2 models. Both the models had very comparable performance metrics, but we have decided to go with model 2, for reasons stated in section 1.3.

Insight & Recommendations:

We can see that Ideal, Premium and Very Good cut makes up for the bulk of the products and the average price of Premium and Very Good is substantially higher than that of Ideal. Hence business must have skilled workers/tools to ensure that maximum of the product cut is targeted towards Premium/Very Good.

If the list of product indicates consumer demand, we can say that color G/H is preferred by majority due to the lower price. Business can look at ways to entice consumers preferring G/H colors to move to I/J which could lead to increased revenue. Bulk of the revenue is made by E/F/G/H colors.

VS1,VS2,SI1 and SI2 clarity brings maximum revenue to the business. Among these clarity categories we can see that SI2 has a much higher average price than others. Business should explore feasibility to increase the clarity of VS1,VS2 and SI1 to SI2. This will lead to higher inflow of revenue.

The linear regression model can be expressed as (where all variables are normalized):

price = -0.989 + (1.389 \* carat)

-(0.0336 \* depth) – (0.0217 \* table) – (0.344 \* x)

-(0.002 \* y) – (0.005 \* z) + (0.122 \* cut\_Good)

+(0.172 \* cut\_Ideal) + (0.140 \* cut\_Premium)

-(0.025 \* color\_F) – (0.064 \* color\_G)

-(0.205 \* color\_H) – (0.401 \* color\_I)

+(1.266 \* clarity\_IF) +(0.985 \* clarity\_SI1)

+(0.640 \* clarity\_SI2) + (1.199 \* clarity\_VVS1)

+ (1.195 \* clarity\_VVS2)

The 5 best attributes that determines the price in descending order are:

* Carat
* Clarity
* Color
* X
* cut

# Problem 2: Logistic Regression and LDA

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



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

### Data Description

1. Holliday\_Package : Opted for Holiday Package (discrete column with value yes/no).
2. Salary : Employee salary (continuous variable).
3. age : Age in years (continuous variable).
4. educ : Years of formal education (continuous variable).
5. no\_young\_children : The number of young children -younger than 7 years (continuous variable).
6. no\_older\_children : Number of older children (continuous variable).
7. foreign : foreigner (discrete column with value yes/no).

### Sample of the dataset:

We can remove the first column in the csv file, which is a serial number column. The dataset sample is as shown below:

Graphical user interface

Description automatically generated with medium confidence

Table-2.1 Dataset Sample

### Exploratory Data Analysis:

Text

Description automatically generated

Table-2.2 Concise data summary

#### Let us check the type of variables in the data frame

There are a total of 872 observations and 7 columns in the dataset. We have 2 columns of object type and remaining 5 of int64 type.

#### Check for missing values in the dataset

From Table-2.2 we can see that all the columns have 872 non-null values, hence there are no null columns in the dataset.

#### Check for duplicate observations in the dataset

There are no duplicate records in the dataset.

Data summary

Graphical user interface, text, application

Description automatically generated

Table-2.3 Data Summary

The above data summary will be further explained in the univariate analysis section below.

#### Univariate Analysis:

Let’s check the central measures of tendency, quartiles, histogram, and boxplot of all 5 continuous columns.

1. Salary

Employee salary is a continuous variable with the below stats (refer Table 2.3):

Mean = 47729.172018

Standard Deviation = 23418.668531

Min value in dataset = 1322

Max value in dataset = 236961

Range = Min – Max = 235639

Q1(1st Quartile) = 35324

Q2(2nd Quartile)/Median = 41903.5

Q3(3rd Quartile) = 53469.5

IQR(Inter-Quartile Range) = Q3- Q1 = 18145.5

Quartile Min value = max(Qmin,Q1 – 1.5 \* IQR) = 8105.75

Quartile Max value = min(Qmax,Q3 + 1.5 \* IQR) = 80687.75

Chart, box and whisker chart

Description automatically generated

Figure-2.1 Histogram & Boxplot : Salary

The Shapiro wilk test on Salary gives a p-value of 2.48e-34, since p-value is less than 0.05 we have enough evidence to reject the null hypothesis, hence Salary does not follow a normal distribution.

Figure-2.1 depicts the histogram and boxplot of “Salary” which shows positive skewness in the data. From the boxplot we can see that there are outliers present in ‘salary’.

1. age

Age in years is a continuous variable with the below stats (refer Table 2.3):

Mean = 39.955275

Standard Deviation = 10.551675

Min value in dataset = 20

Max value in dataset = 62

Range = Min – Max = 42

Q1(1st Quartile) = 32

Q2(2nd Quartile)/Median = 39

Q3(3rd Quartile) = 48

IQR(Inter-Quartile Range) = Q3- Q1 = 16

Quartile Min value = max(Qmin,Q1 – 1.5 \* IQR) = 20

Quartile Max value = min(Qmax,Q3 + 1.5 \* IQR) = 72

Chart, histogram

Description automatically generated

Figure-2.2 Histogram & Boxplot : age

The Shapiro wilk test on age gives a p-value of 5.25e-11, since p-value is less than 0.05 we have enough evidence to reject the null hypothesis, hence age does not follow normal distribution.

Figure-2.2 depicts the histogram and boxplot of “age” which shows positive skewness in the data. From the boxplot we can see that there are no outliers present in ‘age’.

1. educ

Years of formal education is a continuous variable with the below stats (refer Table 2.3):

Mean = 9.307339

Standard Deviation = 3.036259

Min value in dataset = 1

Max value in dataset = 21

Range = Min – Max = 20

Q1(1st Quartile) = 8

Q2(2nd Quartile)/Median = 9

Q3(3rd Quartile) = 12

IQR(Inter-Quartile Range) = Q3- Q1 = 4

Quartile Min value = max(Qmin,Q1 – 1.5 \* IQR) = 2

Quartile Max value = min(Qmax,Q3 + 1.5 \* IQR) = 18

Chart, histogram

Description automatically generated

Figure-2.3 Histogram & Boxplot : educ

The Shapiro wilk test on educ gives a p-value of 1.12e-11, since p-value is less than 0.05 we have enough evidence to reject the null hypothesis, hence educ does not follows normal distribution.

Figure-2.3 depicts the histogram and boxplot of “educ” which shows positive skewness in the data. From the boxplot we can see that there are outliers present in ‘educ’.

1. no\_young\_children

The number of young children (younger than 7 years) is a continuous variable with the below stats (refer Table 2.3):

Mean = 0.311927

Standard Deviation = 0.612870

Min value in dataset = 0

Max value in dataset = 3

Range = Min – Max = 3

Q1(1st Quartile) = 0

Q2(2nd Quartile)/Median = 0

Q3(3rd Quartile) = 0

IQR(Inter-Quartile Range) = Q3- Q1 = 0

Quartile Min value = max(Qmin,Q1 – 1.5 \* IQR) = 0

Quartile Max value = min(Qmax,Q3 + 1.5 \* IQR) = 0

The Shapiro wilk test on no\_young\_children give a p-value of 1.12e-11, since p-value is less than 0.05 we have enough evidence to reject the null hypothesis, hence no\_young\_children does not follow normal distribution.

Figure-2.4 depicts the histogram and boxplot of “no\_young\_children” which shows positive skewness in the data. From the boxplot we can see that there are outliers present in ‘no\_young\_children’.

Chart, histogram

Description automatically generated

Figure-2.4 Histogram & Boxplot : no\_young\_children

1. no\_older\_children

Number of older children is a continuous variable with the below stats (refer Table 2.3):

Mean = 0.982798

Standard Deviation = 1.086786

Min value in dataset = 0

Max value in dataset = 6

Range = Min – Max = 6

Q1(1st Quartile) = 0

Q2(2nd Quartile)/Median = 1

Q3(3rd Quartile) = 2

IQR(Inter-Quartile Range) = Q3- Q1 = 2

Quartile Min value = max(Qmin,Q1 – 1.5 \* IQR) = 0

Quartile Max value = min(Qmax,Q3 + 1.5 \* IQR) = 5

The Shapiro wilk test on no\_older\_children give a p-value of 1.16e-30, since p-value is less than 0.05 we have enough evidence to reject the null hypothesis, hence no\_older\_children does not follow normal distribution.

Figure-2.5 depicts the histogram and boxplot of “no\_older\_children” which shows positive skewness in the data. From the boxplot we can see that there are outliers present in ‘no\_older\_children’.

Chart, histogram

Description automatically generated

Figure-2.5 Histogram & Boxplot : no\_older\_children

Let’s look at the categorical variables:

1. Holliday\_Package

Describes whether the employee has opted or not for any holiday packages.

Let’s look at the count-plot for Holliday\_Package variable and the pie-chart for % distribution.

Chart, bar chart, pie chart

Description automatically generated

Figure-2.6 Count-plot & Pie-chart : Holliday\_Package

From figure 2.6 we can see that 45.99% of the employees(401 employees) have opted for the package and 54.01% (471 employees) have not.

1. foreign

Describes whether the employee is a foreigner or not.

Let’s look at the count-plot for foreign variable and the pie-chart for % distribution.

Chart, treemap chart

Description automatically generated

Figure-2.7 Count-plot & Pie-chart : foreign

The above figure shows that 24.77% (216 employees) are foreigners and 75.23% (656 employees) are not.

#### Outlier Treatment:

We have outliers present in 4 of the numerical variables. But these do not seem to be erroneous data and are real world samples. For example, the number of outliers in salary column accounts for 56 observations out of total 872 observations, i.e., about 6.4%. Removing them might result in overfitting, causing the model to perform abysmally against new test data. Hence let us not impute values for outliers.

#### Bivariate/Multivariate Analysis:

Let us plot a heat map for the correlation matrix of given data frame.

Chart, bar chart

Description automatically generated

Table-2.4 Correlation matrix

We can see moderate negative correlation between age and no\_young\_children.

There exists a weak positive correlation between educ and Salary, which is the norm of the world.

All other correlations are weak.

Let’s also check the Pairplot:

Table

Description automatically generated with medium confidence

Figure-2.8 Pairplot

Pairplot substantiates the heatmap and displays the correlation between variables as stated above.

Let us also look at interaction of variables towards target variable Holliday\_Package and check for inferences.

Chart, bar chart

Description automatically generated

Figure-2.9 Countplot : foreign vs Holliday\_Package

From figure 2.9 we can see that out of 216 foreign employees 147 (68%) have opted for Holliday package whereas for domestic employees only 254 out of 656 (39%) have opted in.

Let’s check if higher salary can explain for the more foreign subscribers.

Chart, bar chart

Description automatically generated

Figure-2.10 Barplot : foreign vs Salary/Holliday\_Package

We can see that foreign employees have a lower average salary than domestic employees and overall, we can see that the employees who chose for holiday package have lower average salary than their counterparts who did not opt. So, it looks like salary does not play a huge factor in the decision to opt.

Let’s look at the salary distribution (not just average) and age for foreign and domestic workers and ensure that our inference made above is correct.

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

Figure-2.11 Scatterplot: age vs Salary/Holliday\_Package

For foreign employees we can see that across age and across Salary, more people have opted in for holiday package.

For domestic employees we can see some specific trends. Counter-intuitively we can see that almost all employees earning more than 100,000 have not opted in for the package, across the ages. Interestingly, salary 100,000 is almost the upper threshold for foreign employees. The concentration of employees who have opted in are more in the age band 25-50. So, we can surmise that for foreign employees age and salary does not really influence the choice to opt or not. For domestic employees age is playing a factor and employees earning more than 100,000 is not generally opting in.

Let’s check how education plays a role.

Chart, bar chart

Description automatically generated

Figure-2.12 Barplot: foreign vs educ/Holliday\_Package

The average education for domestic employees is higher than that of foreign employees.

Among domestic employees we can see that those who choose the holiday package has more education than those who did not, the increase is marginal and hence we can say for domestic employees this might not be a contributing factor. For foreign employees we can see that those who chose the package have lesser average education (less by 1 year) than those who did not.

Let’s look at the scatterplot of employees w.r.t age and education.

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

Figure-2.13 Scatterplot: age vs educ/Holliday\_Package

We can see that among foreign employees across education level, majority of employees have opted in than not. Whereas among domestic employees we can see that between education years of 7.5 to 12.5, there is a clear majority of employees who did not opt in for the holiday package and for other education levels the count is around same.

Let’s check how no\_young\_children affect Holliday\_Package:

Chart, bar chart

Description automatically generated

Figure-2.14 Countplot: no\_young\_children

We can see that majority of the employees have no young children and among these employees, the number of opt ins for holiday packages is higher than those who didn’t opt. Among employees who have at least 1 young children majority have opted not to subscribe for the holiday package.

Let’s look at the distribution among foreign/domestic employees.

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated

Figure-2.15 Countplot: no\_young\_children

Among foreign and domestic employee’s, we can see a marked difference.

Among foreign employees within each band of count of young children, we can see that majority has opted for the package, whereas in domestic employee’s majority has opted out.

The number of employees in each band of young children for foreign/domestic employees are given below:

Table

Description automatically generated Table

Description automatically generated

Table-2.5 Crosstab: foreign vs no\_young\_children Table-2.6 Crosstab: domestic vs no\_young\_children

Let’s check how no\_older\_children affect Holliday\_Package:

Chart, bar chart

Description automatically generated

Figure-2.16 Countplot: no\_older\_children

We can see that among employees having none or one older child, majority have opted out of the holiday package. Among employees who have more than 1 older children majority have chosen to opt in for the package.

Let’s look at the distribution among foreign/domestic employees.

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated

Figure-2.17 Countplot: no\_young\_children

Among foreign and domestic employee’s, we can see a marked difference.

Among foreign employees within each band of count of older children, we can see that majority has opted for the package, whereas in domestic employee’s majority has opted out.

The number of employees in each band of older children for foreign/domestic employees are given below:

Table

Description automatically generated Table

Description automatically generated

Table-2.7 Crosstab: foreign vs no\_older\_children Table-2.8 Crosstab: domestic vs no\_older\_children

## 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).

Let us encode the categorical data via dummy and see the transformed data:

Graphical user interface, text, application, chat or text message

Description automatically generated

Table-2.9 Data head of encoded data

We can see the transformed dummy variable Holliday\_Package\_yes and foreign\_yes.

Let us split the data into test and test and then check the distribution of target variable across test/train datasets.

Graphical user interface, text, application

Description automatically generated Graphical user interface, text, application

Description automatically generated

Table-2.10 Train/Test target variable data

So, we can see that the % distribution of people who have opted for Holliday\_Package is nearly same within the train/test datasets.

* Logistic Regression

Let’s apply logistic regression using sklearn’s linear\_model library. We will set random\_state to 1, rest all parameters default (solver = ‘lbfgs’, tolerance = 0.0001, max\_iteration =100) and fit logistic regression model on the training set.

The model score is coming to 0.5197 or 51.97% accuracy. We will use gridsearch to run the regression model with multiple input parameters and see the best model.

Based on iterations the best model has been derived as:

LogisticRegression(max\_iter=50, random\_state=1, solver='liblinear', tol=1e-06)

The score for the above model on the training data is 0.6705 or 67.05% accuracy.

* Linear Discriminant Analysis (LDA)

Let’s apply LDA using sklearn’s discriminant\_analysis library. We will set all parameters to default (solver = ‘svd’,tolerance = 0.0001) and fit LDA model on the training set.

The model score is coming to 0.6721 or 67.21% accuracy. We will use gridsearch to run the regression model with multiple input parameters and see the best model.

Based on iterations the best model has been derived as:

LinearDiscriminantAnalysis('solver': 'svd', 'tol': 0.0001), which is the default settings.

Hence, the score for LDA on training set is 0.6721 or 67.21% accuracy.

## 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.

Let’s check the model performance of Logistic Regression & LDA as shown below:

#### Logistic Regression

Accuracy:

The accuracy of Logistic regression model on training set is 0.670, i.e., 67%

The accuracy of Logistic regression model on test set is 0.645, i.e., 64.5%

Confusion Matrix:

The below tables display the confusion matrix for training/test data set.

 Text

Description automatically generated

Table-2.11 Logistic Regression – confusion matrix(training data) Table-2.12 Logistic Regression – confusion matrix(test data)

ROC Curve/ROC\_AUC score:

The ROC\_AUC score for Logistic Regression model on training data is 0.7429901494858723.

The ROC\_AUC score for Logistic Regression model on test data is 0.7019746536987916

Chart, line chart

Description automatically generated Chart, line chart

Description automatically generated

Figure-2.18 ROC Curve: Logistic Regression -training data Figure-2.19 ROC Curve: Logistic Regression -test data

Classification Report:

Let’s us look at the classification report of Logistic Regression model against train/test data:

A screenshot of a computer

Description automatically generated with medium confidence A screenshot of a computer

Description automatically generated with medium confidence

Table-2.13 Classification Report: Logistic Regression -training data Table-2.14 Classification Report: Logistic Regression -test data

#### Linear Discriminant Analysis(LDA)

Accuracy:

The accuracy of LDA model on training set is 0.6721, i.e., 67.21%

The accuracy of LDA model on test set is 0.6412, i.e., 64.12%

Confusion Matrix:

The below tables display the confusion matrix for training/test data set.

 

Table-2.15 LDA – confusion matrix(training data) Table-2.16 LDA – confusion matrix(test data)

ROC Curve/ROC\_AUC score:

The ROC\_AUC score for LDA model on training data is 0.7421152682968979.

The ROC\_AUC score for LDA model on training data is 0.7029177718832891. Chart, line chart

Description automatically generated Chart, line chart

Description automatically generated

Figure-2.20 ROC Curve: LDA -training data Figure-2.21 ROC Curve: LDA -test data

Classification Report:

Let’s us look at the classification report of Logistic Regression model against train/test data:

A screenshot of a computer

Description automatically generated with medium confidence A screenshot of a computer

Description automatically generated with medium confidence

Table-2.17 Classification Report: LDA-training data Table-2.18 Classification Report: LDA -test data

#### Model Comparison:

On a high level we can see that both models have similar performance parameters for both training and testing data sets. Let us look at a comparison of the metrics:

A screenshot of a computer

Description automatically generated with medium confidence

Table-2.19 Model comparison

In the above table column LGR refers to Logistic Regression and LDA refers to Linear Discriminant Analysis. We can see that LDA has performed better than LGR in the training dataset for accuracy and precision, whereas in the test dataset LGR has a better score in all parameters except roc\_auc\_score.

This could be attributed to the assumption by LDA that all predictor variables follow a normal distribution, which is not the case with this dataset. LDA is very sensitive to outliers, and this could also be the cause.

Hence, we can consider Logistic Regression model to be the better model for the given dataset.

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

The tour & travel agency have provided information of 872 employees of a company, of which some have opted for the holiday package, and some have not. The business needed help in predicting whether an employee will opt for the package or not based on the information given in the data set.

To achieve this objective, we started with exploratory data analysis of the given dataset. We performed univariate and bivariate analysis to determine how the predictors influence the target variable. We performed logistic regression and linear discriminant analysis of the data. Both the models had very comparable performance metrics, but we have decided to go with logistic regression, for reasons stated in section 2.3.

Insights:

* For foreign employees we can see that across age and across Salary, more people have opted in for holiday package than not.
* For domestic employees almost all employees earning more than 100,000 have not opted in for the package, across the ages.
* Among domestic employees the concentration of employees who have opted in are more in the age band 25-50.
* We can see that among foreign employees across education level, majority of employees have opted in than not.
* Whereas among domestic employees we can see that between education years of 7.5 to 12.5, there is a clear majority of employees who did not opt in for the holiday package.
* Among employees who have at least 1 young children majority have opted not to subscribe for the holiday package.
* Among foreign employees within each band of count of older children, we can see that majority has opted for the package, whereas in domestic employee’s majority has opted out.
* The importance of factors deciding the predictions in descending order are:

foreign, no\_young\_children, educ, age, no\_older\_children, Salary

Recommendations:

* Domestic employees, even though they earn a higher average salary than foreign employees, are mostly opting out of the holiday package. Business needs to determine if there is a major difference between the work functions of domestic/foreign employees ( for e.g., domestic employees are those working on shop floors and hence vacation needs to be planned months in advance and foreign employees are blue collared employees who can manage vacations easily).
* Business needs to have specific holiday packages for employees above 50, such as packages for religious destinations, beach resorts, wine tasting etc.
* Business needs to tie up with holiday resorts who are equipped to deal with young children (presence of in-house creche etc.), which will entice all employees with young children to consider it.
* Business needs to offer specific holiday tours such as adventure, nature, wildlife tours for parents with older children.
* Business needs to conduct a survey among all employees seeking the nature and period of holiday tours they are interested in. This information will give avenues of higher revenue and more data for accurate prediction.

## THE END