Vinay Santosh
PGP Data Science and Business Analytics
March 2021 B

Predictive Modelling Project

29th August 2021

OVERVIEW

This project provides analyzed solutions to the given problem statements by implementing the concepts covered in the Predictive Modelling module. It covers the conceptual framework of the following models: Linear Regression, Logistic Regression, and Linear Discriminant Analysis.

GOALS

1. To read and understand the problem statement. Perform exploratory data analysis on the given datasets.
2. To gain useful insights from EDA, visualize them and perform required imputation or scaling when necessary.
3. Splitting datasets into train and test sets and build models using them. Evaluate model performance on test sets and report the comparison of model performance using the various metrics used for evaluation like Accuracy, Recall, Precision, etc.
4. To construct confusion matrices and draw conclusions from AUC and ROC plots.
5. To report the various insights from the analysis and provide suitable business recommendations/ suggestions.

Contents:

List of Figures:

List of Tables:

Problem 1: Predicting the prices for Cubic Zirconia

1.1 Exploratory Data Analysis:

- The problem statement given requires a model that is able to predict the prices for Cubic Zirconia based on the various attributes of the stone such as dimensions, weight, depth, etc.
- Each stone is valued based on the various outlined attributes and occupies a specific price slot.
- To predict the prices of stones that are most profitable to the GemStone company, we will build a linear regression model that distinguishes lower profitable stones from higher profitable stones.
- The given dataset contains 26967 records of cubic zirconia stones along with 10 features or attributes such as clarity, cut, and color of the stone.

	carat	cut	color	clarity	depth	table	x	y	z	price
0	0.30	Ideal	E	SI1	62.1	58.0	4.27	4.29	2.66	499
1	0.33	Premium	G	IF	60.8	58.0	4.42	4.46	2.70	984
2	0.90	Very Good	E	VVS2	62.2	60.0	6.04	6.12	3.78	6289
3	0.42	Ideal	F	VS1	61.6	56.0	4.82	4.80	2.96	1082
4	0.31	Ideal	F	VVS1	60.4	59.0	4.35	4.43	2.65	779

Table 1.1 a

- There are three attributes that are of object datatype and the remaining are numeric entries.
- Upon checking for and removing duplicates, we have a total of 26933 records in our data.

Univariate Analysis:

- By performing univariate analysis on each of the numeric attributes, we can understand its distribution in the dataset and make observations on its spread.

```
Int64Index: 26933 entries, 0 to 26966
Data columns (total 10 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   carat    26933 non-null  float64
 1   cut      26933 non-null  object
 2   color    26933 non-null  object
 3   clarity  26933 non-null  object
 4   depth    26236 non-null  float64
 5   table    26933 non-null  float64
 6   x        26933 non-null  float64
 7   y        26933 non-null  float64
 8   z        26933 non-null  float64
 9   price    26933 non-null  int64
dtypes: float64(6), int64(1), object(3)
```

Figure 1.1 a

- From the distribution plots given in figure 1.1 b, we notice that most of the features have a bit of skewness in them.
- “Carat”, “table”, “y” and “z” have their data skewed to its left where are “x” leans more to the right.
- The feature “depth”, which is the height of the stone from its table to the culet divided by the average diameter of its girdle, has a more centered distribution.

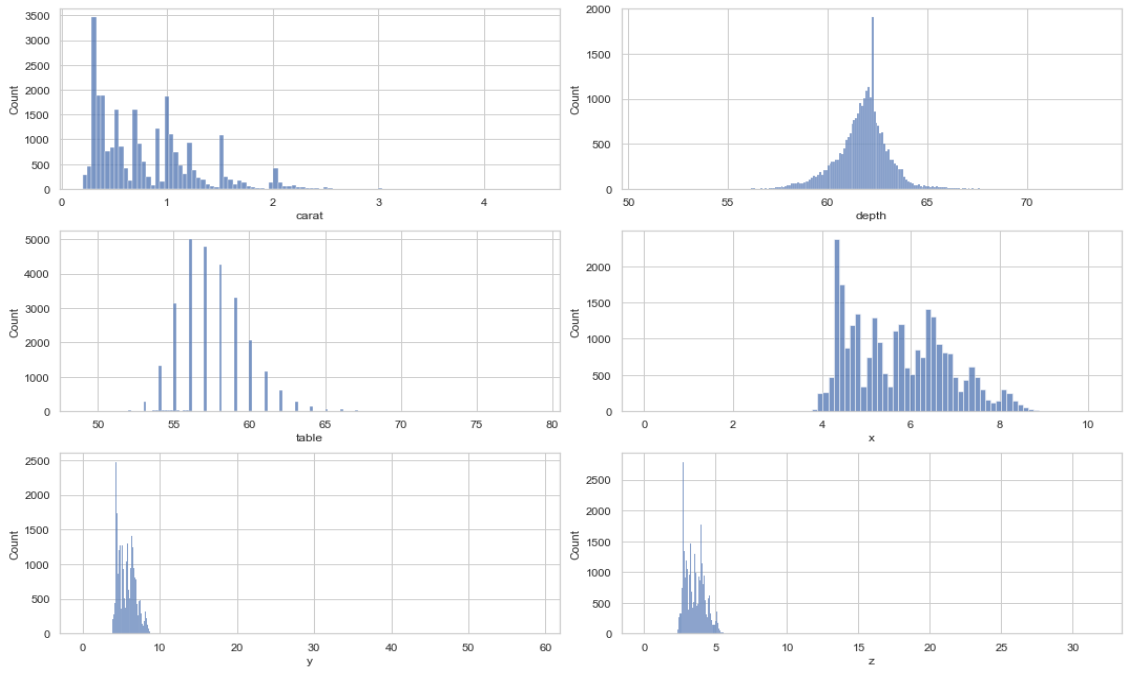


Figure 1.1 b

- By plotting the boxplots for the given numeric variables (Figure 1.1 c), we observe the presence of outliers in most of the features.
- In order to build a good Linear Regression model, we will have to treat the outliers present.

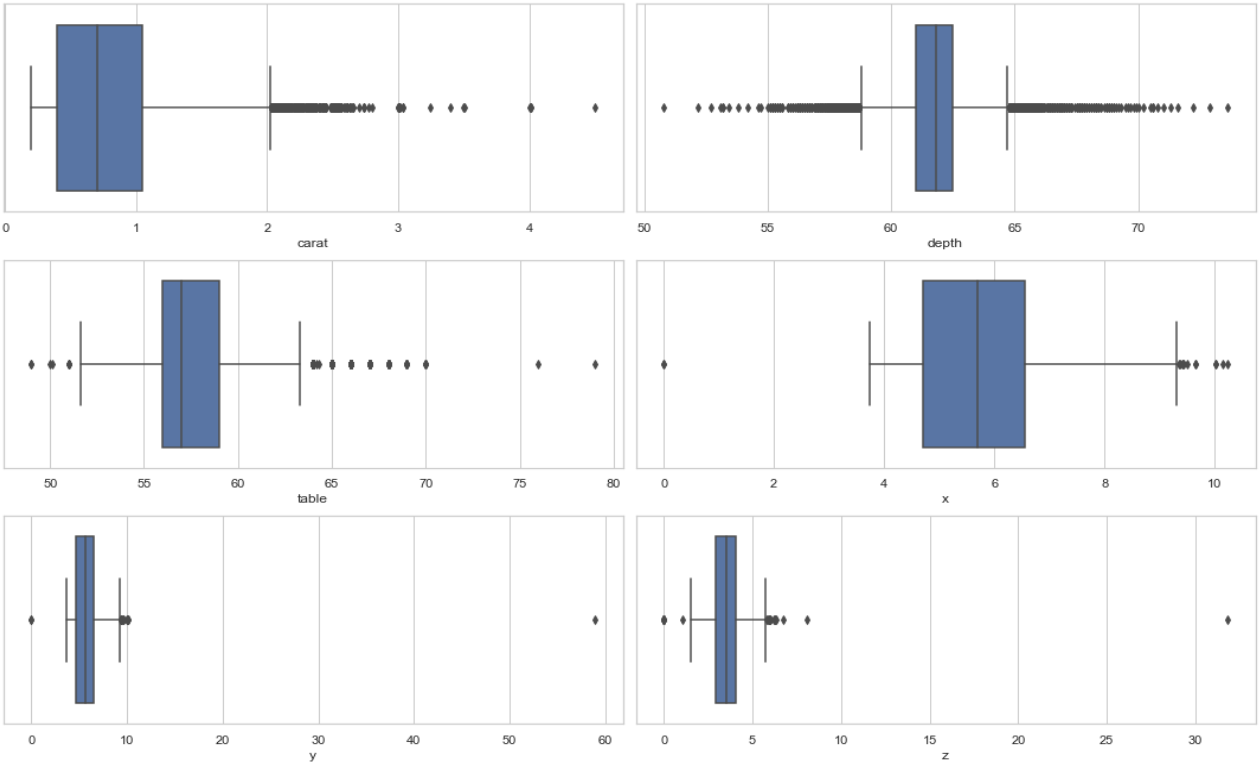


Figure 1.1 c

Multivariate Analysis:

- By performing multivariate analysis, we are able to observe the relationships between the various features.
- From the figure below (Figure 1.1 d), we notice that “carat” has a linear relationship with the dimensional features such as “x”, “y”, “z”.
- The features “Depth” and “table” have the least correlation and can be significant features in building a good model.
- We can observe that the “price” variable increases with the increase in “carat”, “depth” and “table”.

- The “carat” variable increases with the length of the stone, however, the other-dimensional variables “y” and “z” seem to have a constant lower value.
- This tells us that most stones have a smaller width and height but can still be priced high and can weigh more.

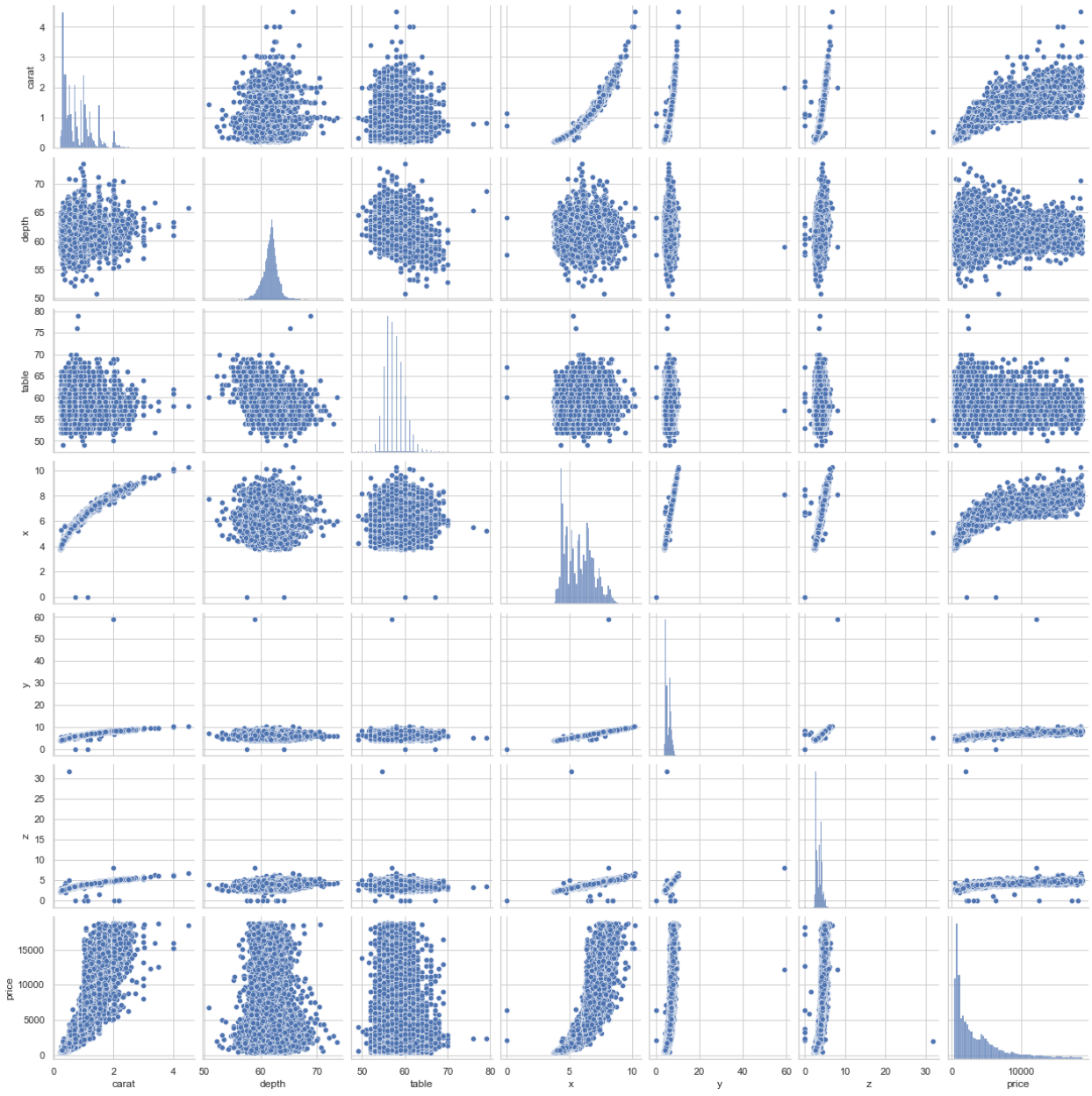


Figure 1.1 d

- From the heatmap given below in Figure 1.1 e, we can observe that the variable “carat” has a high correlation with most of the variables.

- The features “depth” and “table” have the least correlation with all the other variables which can be ideal for building a good model.

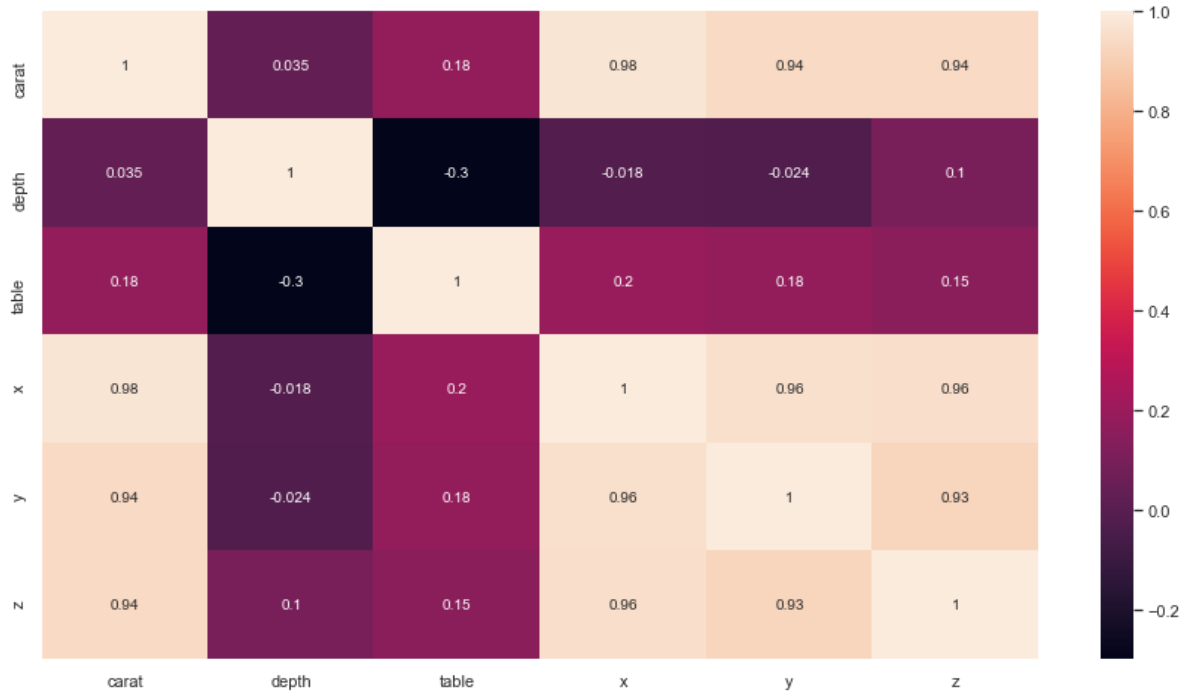


Figure 1.1 e

1.2 Data Pre-Processing:

- The given dataset has a few null values in the “depth” variable as shown in the table (Table 1.2 a) below.

carat	0
cut	0
color	0
clarity	0
depth	697
table	0
x	0
y	0
z	0
price	0

Table 1.2 a

- Since the median depth of the stones in the dataset is 61.8, we impute the missing values with the median value.
- We can observe all the entries with the value zero in them from the table (Table 1.2 b) below.

	carat	cut	color	clarity	depth	table	x	y	z	price
5821	0.71	Good	F	SI2	64.1	60.0	0.00	0.00	0.0	2130
6034	2.02	Premium	H	VS2	62.7	53.0	8.02	7.95	0.0	18207
10827	2.20	Premium	H	SI1	61.2	59.0	8.42	8.37	0.0	17265
12498	2.18	Premium	H	SI2	59.4	61.0	8.49	8.45	0.0	12631
12689	1.10	Premium	G	SI2	63.0	59.0	6.50	6.47	0.0	3696
17506	1.14	Fair	G	VS1	57.5	67.0	0.00	0.00	0.0	6381
18194	1.01	Premium	H	I1	58.1	59.0	6.66	6.60	0.0	3167
23758	1.12	Premium	G	I1	60.4	59.0	6.71	6.67	0.0	2383

Table 1.2 b

- The two entries (index: 5821, 17506) have the value zero in all of their dimensional values.
- We know that a stone cannot exist with these values, hence we can drop them.
- Upon treating the outliers, the rest of the zero values get capped.
- The scaling of the data does not impact the accuracy of the model. However, for easier interpretation of coefficients and the difference in magnitude of the variables, we can scale the data.
- We can observe from the description of the data that there is not much variance in the data that will affect the model.

	count	mean	std	min	25%	50%	75%	max
carat	26931.0	0.798001	0.477250	0.20	0.400	0.70	1.05	4.50
cut	26931.0	3.909881	1.113004	1.00	3.000	4.00	5.00	5.00
color	26931.0	4.394787	1.705941	1.00	3.000	4.00	6.00	7.00
clarity	26931.0	4.946381	1.646752	1.00	4.000	5.00	6.00	8.00
depth	26931.0	61.746771	1.393613	50.80	61.100	61.80	62.50	73.60
table	26931.0	57.455501	2.231428	49.00	56.000	57.00	59.00	79.00
x	26931.0	5.729772	1.126327	3.73	4.710	5.69	6.55	10.23
y	26931.0	5.733528	1.164032	3.71	4.715	5.70	6.54	58.90
z	26931.0	3.538032	0.719345	0.00	2.900	3.52	4.04	31.80
price	26931.0	3937.502506	4022.658593	326.00	945.000	2375.00	5355.50	18818.00

Table 1.2 c

- For this example, we apply a z-score technique of scaling which centralizes the data so that the mean of all the variables tends to zero and the standard deviation becomes 1.
- The description of the scaled data is given in the table (Table 1.2 d) below.

	count	mean	std	min	25%	50%	75%	max
carat	26931.0	-1.737746e-16	1.000019	-1.253037	-0.833962	-0.205349	0.528033	7.757083
cut	26931.0	-3.840516e-16	1.000019	-2.614488	-0.817515	0.080971	0.979457	0.979457
color	26931.0	4.899146e-17	1.000019	-1.990016	-0.817621	-0.231423	0.940972	1.527170
clarity	26931.0	-7.934935e-17	1.000019	-2.396508	-0.574706	0.032561	0.639828	1.854362
depth	26931.0	4.293761e-17	1.000019	-7.855103	-0.464105	0.038195	0.540496	8.505552
table	26931.0	-1.466459e-15	1.000019	-3.789349	-0.652286	-0.204134	0.692170	9.655208
x	26931.0	-7.238154e-16	1.000019	-1.775514	-0.905413	-0.035312	0.728247	3.995566
y	26931.0	1.067259e-15	1.000019	-1.738410	-0.875016	-0.028803	0.692839	45.675243
z	26931.0	-4.828157e-16	1.000019	-4.918500	-0.886979	-0.025068	0.697826	39.289210
price	26931.0	2.443802e-17	1.000019	-0.897807	-0.743925	-0.388433	0.352509	3.699239

Table 1.2 d

1.3 Data Encoding and Building Linear Regression Model:

- To build the Linear Regression Model, we encode the categorical variables with numerical values that represent the information it carries.

	cut	color	clarity
0	Ideal	E	SI1
1	Premium	G	IF
2	Very Good	E	VVS2
3	Ideal	F	VS1
4	Ideal	F	VVS1

Figure 1.3 a

- "Clarity" refers to the absence of inclusions and blemishes. (In order from best to worst: IF, VVS1, VVS2, VS1, VS2, SI1, SI2, I1
- "Cut" refers to the quality of the cut, ranging in increasing order Fair, Good, Very Good, Premium, Ideal.
- "Color" indicates the quality of the stone with D being the worst and J the best.

Value Counts Before Encoding:	Value Counts AfterEncoding:
Ideal 10805 Premium 6886 Very Good 6027 Good 2434 Fair 779 Name: cut, dtype: int64	5 10805 4 6886 3 6027 2 2434 1 779 Name: cut, dtype: int64

G 5652 E 4916 F 4722 H 4095 D 3341 I 2765 J 1440 Name: color, dtype: int64	4 5652 6 4916 5 4722 3 4095 7 3341 2 2765 1 1440 Name: color, dtype: int64
SI1 6565 VS2 6093 SI2 4563 VS1 4086 VVS2 2530 VVS1 1839 IF 891 I1 364 Name: clarity, dtype: int64	6 6565 5 6093 7 4563 4 4086 3 2530 2 1839 1 891 8 364 Name: clarity, dtype: int64

Table 1.3 a

- From the table, we notice that the majority of the stones have a fairly good cut and a small proportion of stones with the best color.
- The average color quality and clarity of most of the stones seems to be below average standards.
- The given dataset is split into train and test sets in the ratio of 70:30. By creating a Linear Regression model, we are able to derive the following scores:

	Train RMSE	Test RMSE	Training Score	Test Score
Linear Regression	0.303002	0.301255	0.907797	0.910135

Figure 1.3 b

- The linear regression model finds the best fit line that is able to distinguish between more profitable and less profitable stones.

- The RMSE (Root Mean Square Error) is the measure of spread of the residuals around the best fit line. We can see that both the train and test sets have a similar RMSE.
- The model score for both sets is close to each other which indicates that the model performs equally well on the test set and there is no overfitting in the model.

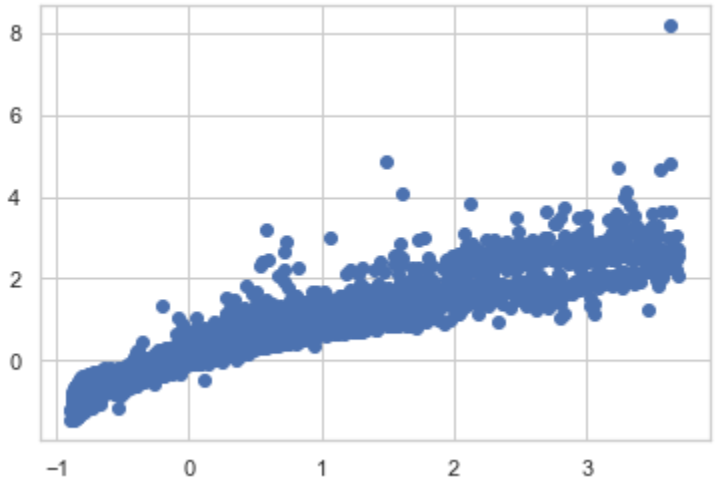


Figure 1.3 c

- On plotting the values of the predicted values on the test set with the actual values, we can observe that the data points are close to the best fit line with some amount of scattering.
- The model explains 91% of the variability in predicting the price of the stone with the given independent variables and its coefficients.

- The figure above gives us a summary of statistics that tells us how well the model is able to predict the “price” variable given the independent variable.

- We can observe the coefficients of the independent variables along with the intercept. From the derived coefficients, we can determine that there is a presence of collinearity in the dataset.

OLS Regression Results

Dep. Variable:	price	R-squared:	0.908
Model:	OLS	Adj. R-squared:	0.908
Method:	Least Squares	F-statistic:	2.061e+04
Date:	Sat, 28 Aug 2021	Prob (F-statistic):	0.00
Time:	14:52:51	Log-Likelihood:	-4240.0
No. Observations:	18851	AIC:	8500.
Df Residuals:	18841	BIC:	8579.
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.747e-05	0.002	0.008	0.994	-0.004	0.004
carat	1.3012	0.011	117.987	0.000	1.280	1.323
cut	0.0289	0.003	10.717	0.000	0.024	0.034
color	0.1378	0.002	58.974	0.000	0.133	0.142
clarity	-0.2055	0.002	-84.028	0.000	-0.210	-0.201
depth	-0.0292	0.003	-10.667	0.000	-0.035	-0.024
table	-0.0172	0.003	-6.144	0.000	-0.023	-0.012
x	-0.2655	0.014	-18.911	0.000	-0.293	-0.238
y	0.0030	0.007	0.433	0.665	-0.011	0.017
z	-0.0016	0.007	-0.233	0.816	-0.015	0.012

Omnibus:	4235.479	Durbin-Watson:	1.999
Prob(Omnibus):	0.000	Jarque-Bera (JB):	202981.015
Skew:	-0.138	Prob(JB):	0.00
Kurtosis:	19.073	Cond. No.	15.8

Figure 1.3 c

- The R-squared and adjusted R-squared are similar and closer to 1 which indicates that the given features are good enough to predict the dependent variable.
- The p-value is < 0.05 which means that there is a relationship between the “price” variable and the independent variables. Hence this model, minus the less important features, is good for predicting the future profitability of the stone.
- We calculate the variance inflation factor for each of the variables, which is a measurement of the multicollinearity between the independent variables.

Variance Inflation Factor:

```
carat ---> 24.9792832820242
cut ---> 1.4939070171396887
color ---> 1.1208192931535599
clarity ---> 1.2304345213907233
depth ---> 1.572847193224135
table ---> 1.5938807880178072
x ---> 46.85866904936365
y ---> 13.84084424311968
z ---> 14.107227076730016
```

Table 1.3 b

1.4 Inference, Insights, and Recommendations:

From the linear regression model, we can infer the following:

- Most of the given independent variables are significant in predicting the profitability of the stone.
- The variables “cut”, “color”, “clarity”, “depth”, “table” appear to be the best predictors of price.
- From a business perspective, stones that are of high quality ie with the best clarity, color, and cut are more profitable than stones that are bigger.
- The appearance of the stone is more valuable than its size. However, the weight of the stone is a key factor in pricing the stone and make it more profitable.
- Stones that are smaller in size with higher quality are more profitable as bigger stones will require more processing and yield smaller profits as they are more expensive to make.
- The gemstone company must invest in procedures to ensure that the process stones meet higher standards to yield better profits.

Problem 2: Predicting Employees who chose of Tour Package

2.1 Exploratory Data Analysis:

- The given dataset consists of 872 records of employees with their information such as age, number of children, salary, etc.
- The feature to be predicted in this problem statement is “Holiday Package” which has binary values of whether or not an employee chooses the holiday package offered by the company.
- The dataset has no duplicates or null values.
- It consists of a total of 7 attributes. Two features are of “object” datatype and five numeric features.

```
RangeIndex: 872 entries, 0 to 871
Data columns (total 7 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Holliday_Package  872 non-null    object
 1   Salary            872 non-null    int64
 2   age               872 non-null    int64
 3   educ              872 non-null    int64
 4   no_young_children 872 non-null    int64
 5   no_older_children 872 non-null    int64
 6   foreign           872 non-null    object
dtypes: int64(5), object(2)
```

Figure 2.1 a

- The majority of the employees are of the same nationality as observed in the plot given below.
- The number of employees in the dataset that opt out of the offered holiday package (54%) is higher than those that choose the package.

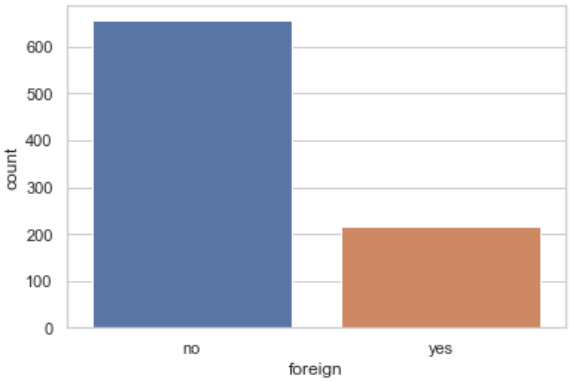


Figure 2.1 b

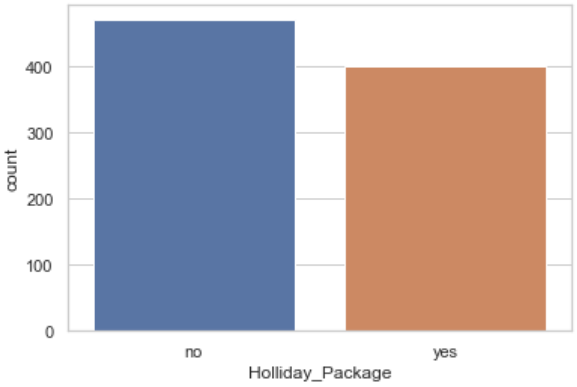


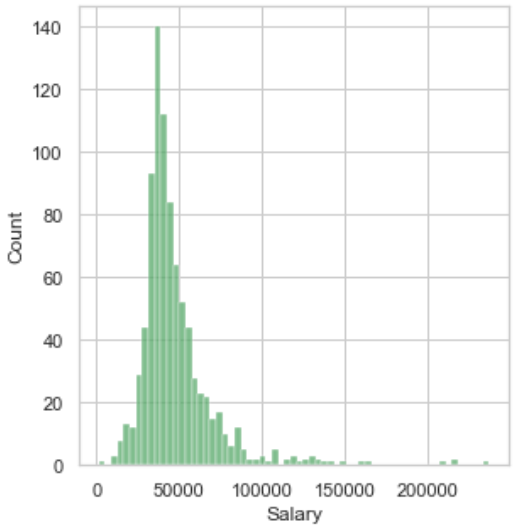
Figure 2.1 c

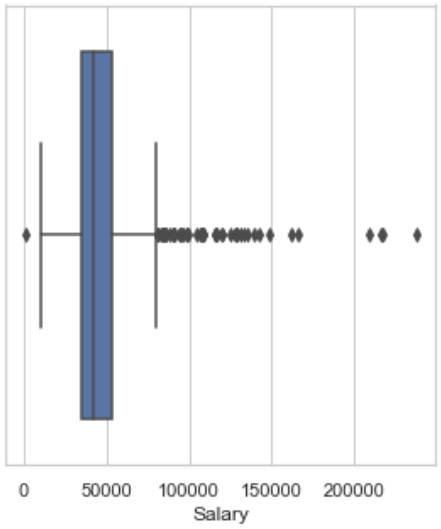
- The description of the dataset provides us information about the mean values, the frequency of variation, max and min values.
- The employees are of different age groups ranging from 20 to 62 years.
- 31% of the employees have no children.

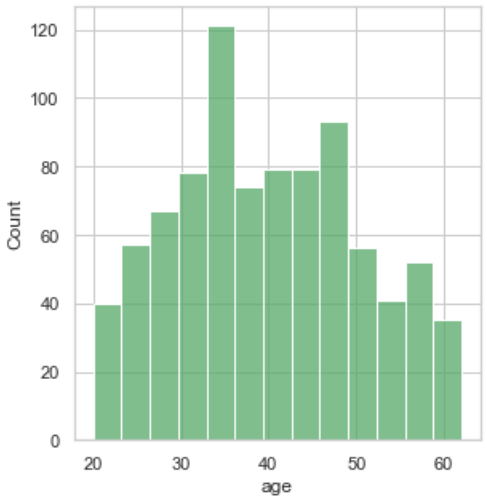
	Salary	age	educ	no_young_children	no_older_children
count	872.000000	872.000000	872.000000	872.000000	872.000000
mean	47729.172018	39.955275	9.307339	0.311927	0.982798
std	23418.668531	10.551675	3.036259	0.612870	1.086786
min	1322.000000	20.000000	1.000000	0.000000	0.000000
25%	35324.000000	32.000000	8.000000	0.000000	0.000000
50%	41903.500000	39.000000	9.000000	0.000000	1.000000
75%	53469.500000	48.000000	12.000000	0.000000	2.000000
max	236961.000000	62.000000	21.000000	3.000000	6.000000

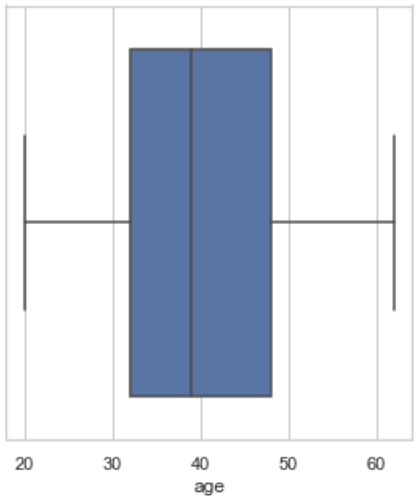
Table 2.1 a

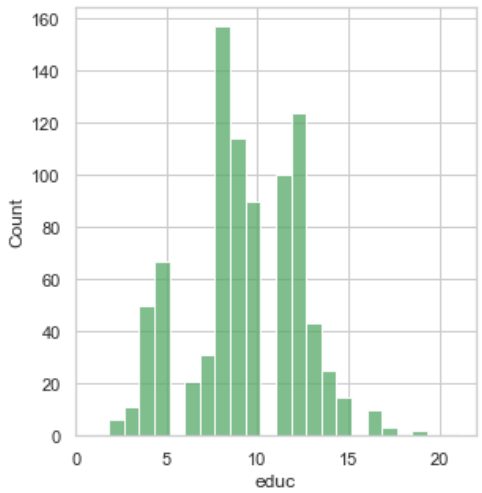
- The table below visualizes the univariate analysis of each feature. We can observe the presence of outliers in the data set (except for “age”) which will be treated in order to build a good model.
- We also observe skewness in the data with irregularities in the distribution with an exception of the variable “age”.

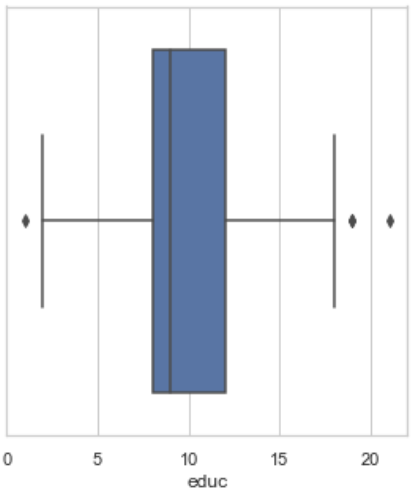

Count
Salary
0
20
40
60
80
100
120
140
0
50000
100000
150000
200000


Salary
0
50000
100000
150000
200000


Count
age
0
20
40
60
80
100
120
20
30
40
50
60


age
20
30
40
50
60


Count
educ
0
20
40
60
80
100
120
140
160
0
5
10
15
20


educ
0
5
10
15
20

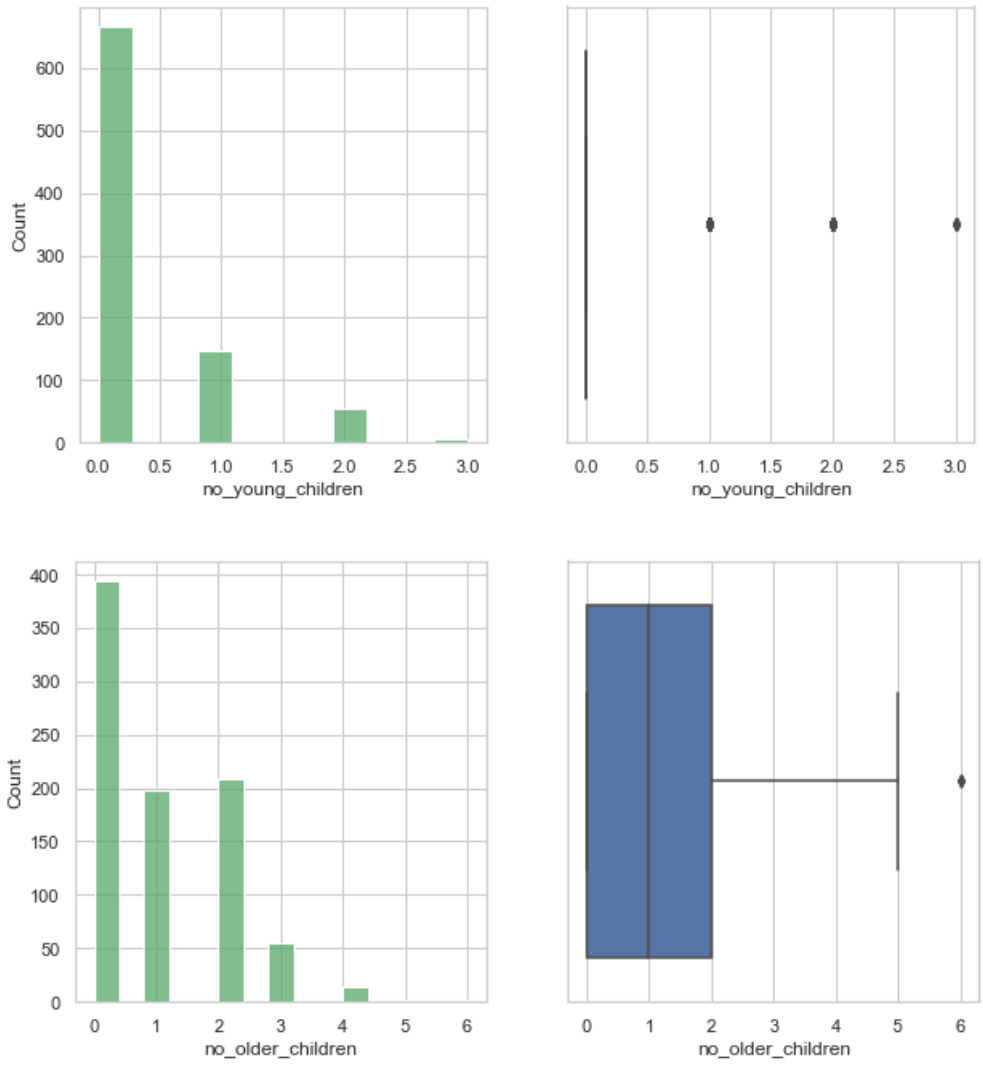


Table 2.1 b

- The heatmap below shows the correlation between the independent variables ranging from high (1) to low(0).
- We can observe that all the variables are not highly correlated in the entire dataset.
- There is a negative correlation between variables “age” and the number of children/ years of education of the employees.
- We see a relatively high correlation between “Salary” and “edu” which indicates that the higher the years of education, the higher the salary of the employee is.

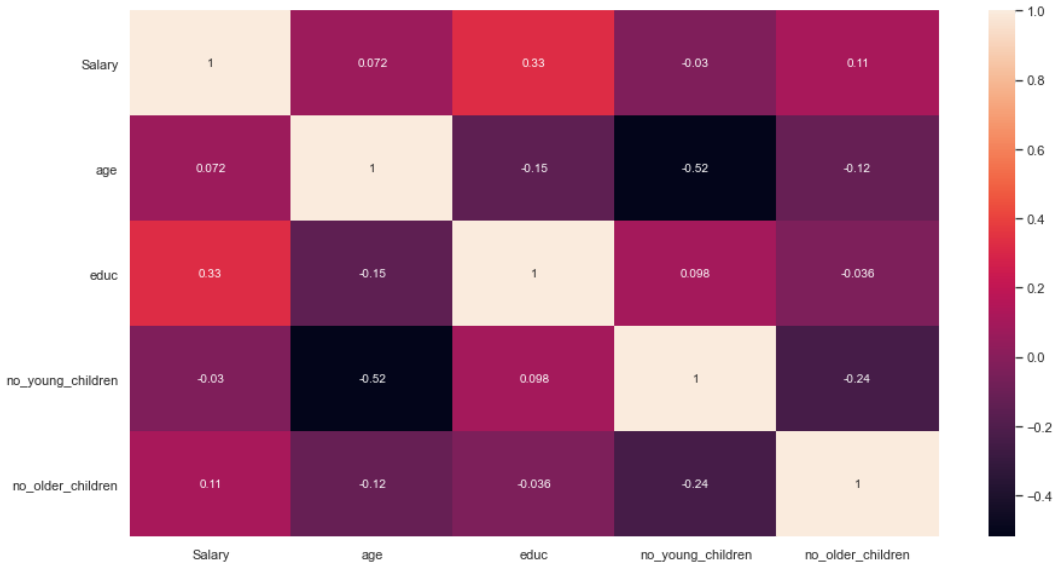


Figure 2.1 d

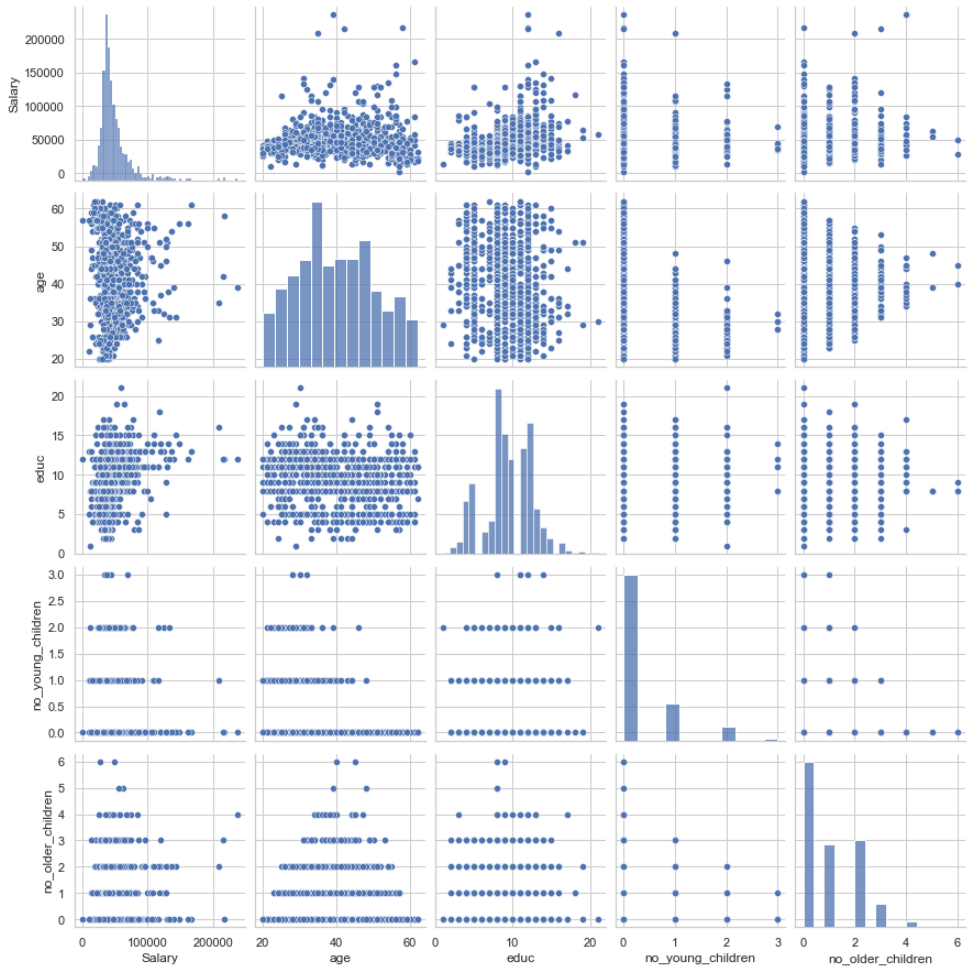


Table 2.1 e

2.2 Data Encoding and Pre-Processing:

- The outliers found in the numeric variables are treated and imputed to the appropriate ranges however, this treating outliers does not affect the accuracy of the model.
- The upper and lower ranges for the independent variables are as given in the figure (Figure 2.2 a) below.
- The categorical variables are encoding using the categorical codes function available in the pandas library.

	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
0	0	48412	30	8	1	1	0
1	1	37207	45	8	0	1	0
2	0	58022	46	9	0	0	0
3	0	66503	31	11	2	0	0
4	0	66734	44	12	0	2	0

Figure 2.2 a

```
Lower Range Values:
 Salary              8105.75
age                    8.00
educ                   2.00
no_young_children      0.00
no_older_children     -3.00
dtype: float64

Upper Range Values:
 Salary             80687.75
age                   72.00
educ                  18.00
no_young_children      0.00
no_older_children      5.00
dtype: float64
```

Figure 2.2 b

- The logistic regression and linear discriminant models are classification models that can help predict the employees who opt for the holiday package.
- For logistic regression we use the following parameters for improved accuracy : solver='newton-cg',max_iter=10000,penalty='none',verbose=True,n_jobs=2.

2.3 Model Evaluation and Performace Metrics:

Logistic Regression Model:

Train Set:

	precision	recall	f1-score	support
0	0.67	0.77	0.72	326
1	0.68	0.56	0.61	284
accuracy			0.67	610
macro avg	0.67	0.66	0.66	610
weighted avg	0.67	0.67	0.67	610

Test Set:

```
Accuracy for Logistic Regression model is
 0.6374045801526718
```

Classification report for Logistic Regression model is

	precision	recall	f1-score	support
0	0.66	0.70	0.68	145
1	0.60	0.56	0.58	117
accuracy			0.64	262
macro avg	0.63	0.63	0.63	262
weighted avg	0.64	0.64	0.64	262

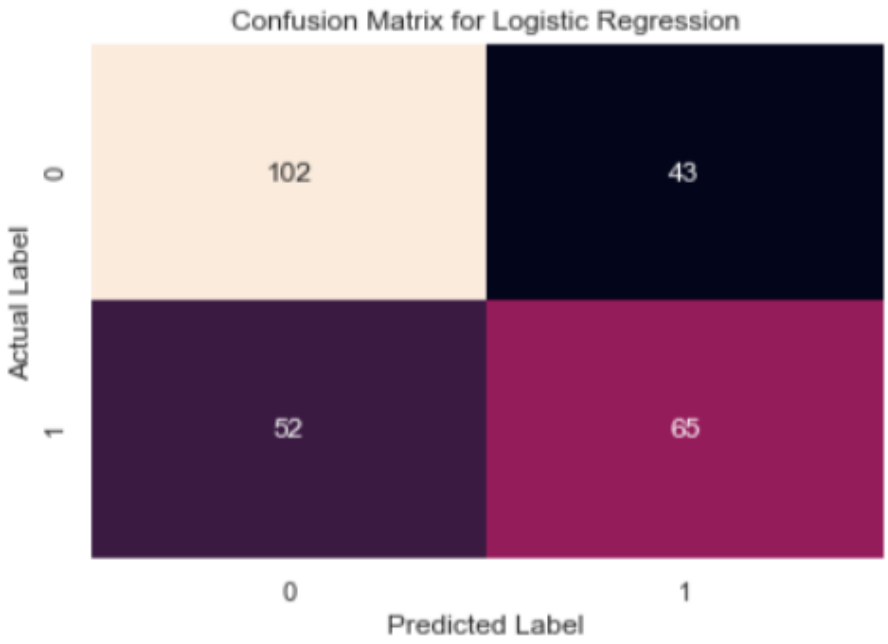


Figure 2.3 a

LDA:

Train Set:

	precision	recall	f1-score	support
0	0.67	0.77	0.72	326
1	0.68	0.56	0.61	284
accuracy			0.67	610
macro avg	0.67	0.66	0.66	610
weighted avg	0.67	0.67	0.67	610

Test Set:

```
Accuracy for LDA model is
 0.6412213740458015


Classification report for LDA model is
              precision    recall  f1-score   support

           0       0.66      0.71      0.69       145
           1       0.61      0.56      0.58       117

    accuracy                           0.64       262
   macro avg       0.64      0.63      0.63       262
weighted avg       0.64      0.64      0.64       262


Confusion Matrix for LDA model is
```

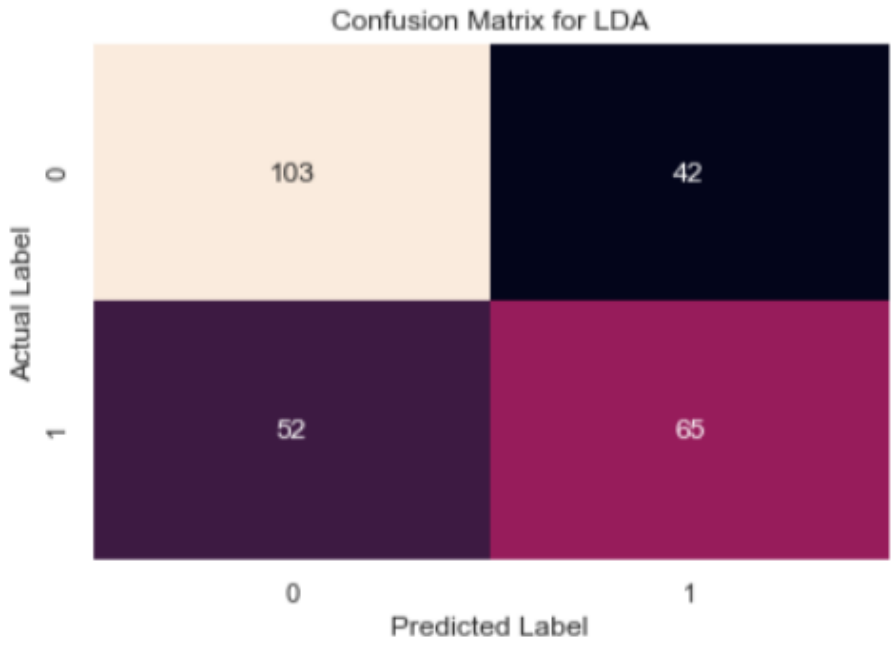


Figure 2.3b

- From the above analysis of the performance metrics, we can observe that both the models perform better on train sets but drops in performance on the train sets.

- We observe that both the models have an accuracy of 64% on the test sets. This indicates that the models do not do a good job in predicting the chance of an employee opting for a holiday package.
- Upon checking the area under the curve and plotting the ROC for the test sets of both models, we arrive with the following plot:
- Area under the curve for Logistic Regression Model is 0.7046861184792218
- Area under the curve for LDA Model is 0.7029177718832891

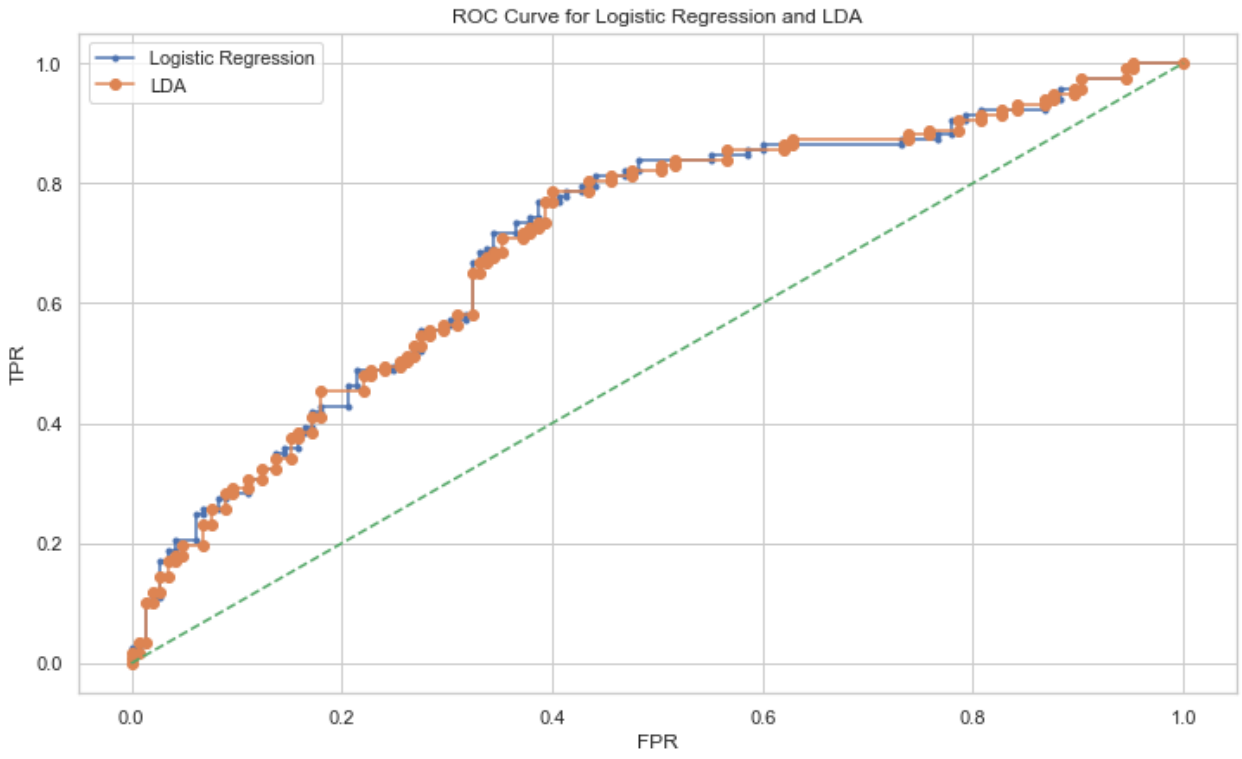


Figure 2.3 c

- We can observe that the results for both the models are similar and either of them can be used to solve this classification problem.

2.4 Inference, Insights, and Recommendations:

- From the analysis of the given dataset, we conclude that the given variables are not sufficient to build a model to predict the desired class.
- We can determine that both the models are able to predict the target class with the same precision and the model overfits the data as it performs poorly in the test sets.

- It is evident that the available features have no correlation between them.
- Since the proportion of the target variable is not too spread out, it is more difficult to make a prediction with the given set of data.
- However, from the evaluation of the models, we conclude that both the models do a similar job in making a prediction.
- The business must collect further information of the employees to determine the key factors that play a role in helping us predict the class.
- It should also provide a discounted rate for employees of the company inorder to increase the chances for them to opt for the package.