BUSINESS REPORT

PREDICTIVE MODELING

Srinidhi Devan

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.

Objective:

1. 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.

2. To provide them with the best 5 attributes that are most important.

About the Data:

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	Cubic zirconia Clarity refers to the absence of the Inclusions and Blemishes. (In order from Best to Worst, FL = flawless, I3= level 3 inclusions) FL, IF, VVS1, VVS2, VS1, VS2, SI1, SI2, I1, I2, I3
Depth	The Height of a 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	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). Perform Univariate and Bivariate Analysis.**

carat	cut	color	clarity	depth	table	x	y	z	price
0.3	Ideal	E	SI1	62.1	58	4.27	4.29	2.66	499
0.33	Premium	G	IF	60.8	58	4.42	4.46	2.7	984
0.9	Very Good	E	VVS2	62.2	60	6.04	6.12	3.78	6289
0.42	Ideal	F	VS1	61.6	56	4.82	4.8	2.96	1082
0.31	Ideal	F	VVS1	60.4	59	4.35	4.43	2.65	779

The **dimension/shape** of the Data is: **(26967, 10)**
Number of **Duplicate Rows** in the dataset: **34**

Missing Values:

From the below table, we see that 'depth' has 697 missing values. Imputations are done in the subsequent steps.

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

Data Types:

'cut', 'color' and 'clarity' are of Object type. We need to convert them into numerical values for further analysis.

Variables	Data Types
carat	float64
cut	object
color	object
clarity	object
depth	float64
table	float64
x	float64
y	float64
z	float64
price	int64

Measures of Central Tendency:

Variables	Mean	Median	Mode
carat	0.80	0.7	0.3
depth	61.75	61.8	62
table	57.46	57	56
x	5.73	5.69	4.38
y	5.73	5.71	4.35
z	3.54	3.52	2.69
price	3939.52	2375	544

- The average 'carat' of cubic zirconia is 0.8 and 'depth' is 61.75.

- The average ‘price’ stood around 3939.52.
- From the inspection of the above table, the variables more or less seems to be symmetric. However, ‘price’ seems to be right-skewed.

Measures of Dispersion:

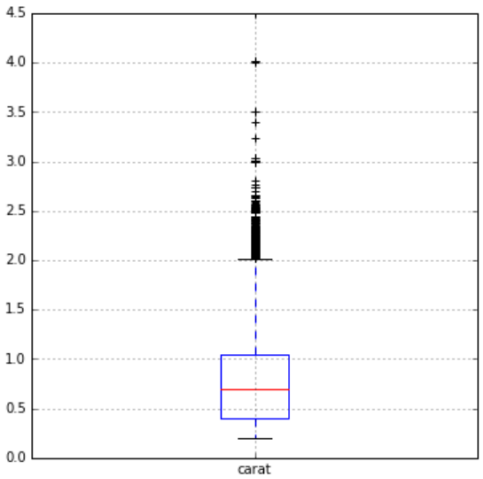
Variables	Variance	IQR	CV
carat	0.23	0.65	0.60
depth	2.00	1.5	0.02
table	4.98	3	0.04
x	1.27	1.84	0.20
y	1.36	1.83	0.20
z	0.52	1.14	0.20
price	16199540.00	4415	1.02

- ‘price’ amongst all the variables has the highest variance implying it is highly spread out from the mean.
- The lowest variance, on the other hand, is observed for ‘carat’ and ‘z’ implying data points are identical to one another.

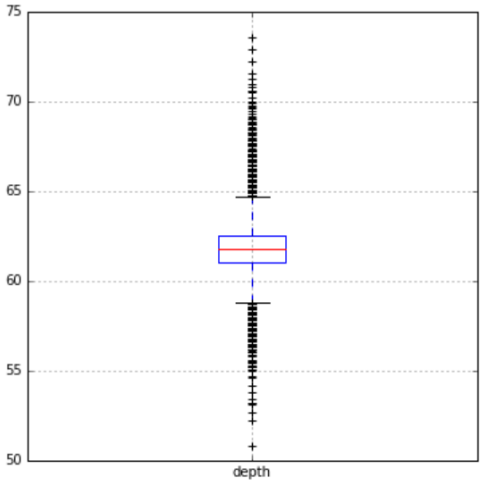
Five Number Summary:

The five number summary helps describe the center, spread and shape of data.

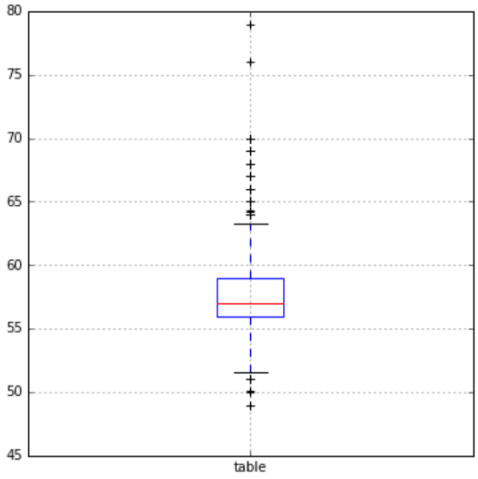
1. Carat



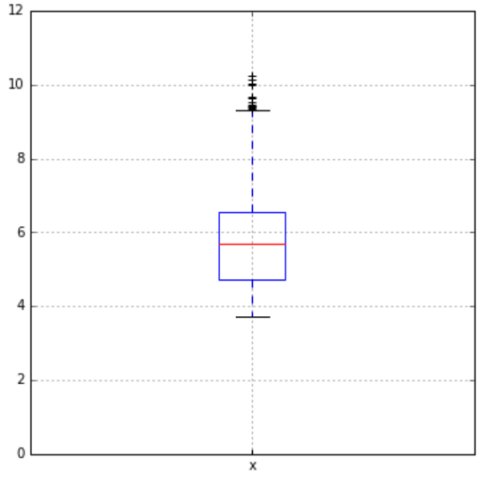
2. Depth



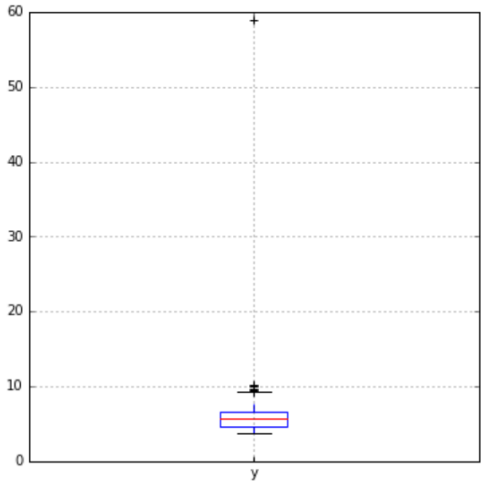
3. Table



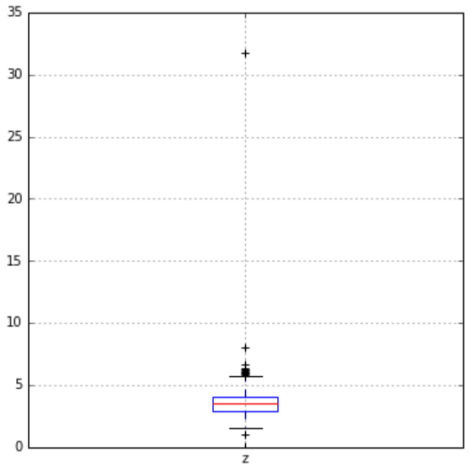
4. x



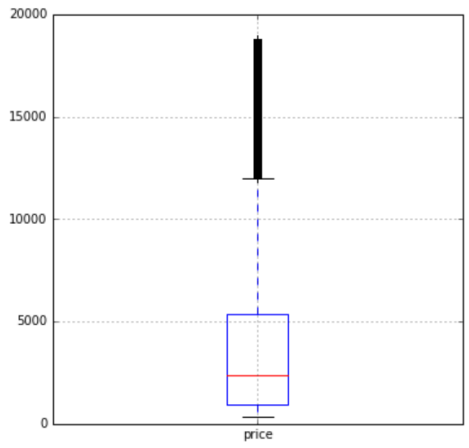
5. y


60
50
40
30
20
10
0
y

6. z


35
30
25
20
15
10
5
0
z

7. Price


20000
15000
10000
5000
0
price

All the variables under consideration contains outliers and needs to be treated for further analysis.

Skewness:

Variables	Skewness
carat	1.116
depth	-0.029
table	0.766
x	0.388
y	3.850
z	2.568
price	1.619

From the above table, 'depth' seems to have negative skewness while 'y' has the highest positive skewness.

Covariance:

	carat	depth	table	x	y	z	price
carat	0.228	0.024	0.194	0.526	0.524	0.324	1773.68
depth	0.024	1.996	-0.939	-0.030	-0.041	0.100	-14.60
table	0.194	-0.939	4.982	0.494	0.475	0.240	1140.42
x	0.526	-0.030	0.494	1.274	1.267	0.778	4025.45
y	0.524	-0.041	0.475	1.267	1.360	0.781	4018.54
z	0.324	0.100	0.240	0.778	0.781	0.519	2466.91
price	1773.678	-14.597	1140.420	4025.446	4018.538	2466.906	16199540.00

- Covariance tells us to what extent two variables change together.
- A positive value indicates that the two variables move in the same direction and vice versa.
- 'depth' varies inversely to 'table', 'x', 'y' and 'price' i.e., 'depth' and these variables does not move in the same direction.
- All other variables have a positive values with each other.

Correlation:

- While covariance can only determine the relationship between the two variables, correlation helps us determine the strength between the variables.
- In comparison to covariance, correlation is independent of scale. However, correlation does not imply causation.
- As observed from the above heatmap, there is a high positive correlation between ‘x’ and ‘y’; ‘y’ and ‘z’; ‘x’ and ‘z’.
- Negative correlation is observed between: ‘depth’ and ‘table’; ‘depth’ and ‘x’; ‘depth’ and ‘y’.

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? Do you think scaling is necessary in this case?**

- From the initial data exploration of the dataset, we saw that there were around 697 missing values in 'depth'. This has been imputed using 'Median' as it is the most stable when compared to 'Mean' imputation which is highly affected by the extreme values in the dataset.

Variables	Missing Values
carat	0
cut	0
color	0
clarity	0
depth	0
table	0
x	0
y	0
z	0
price	0

Now we see that there are no missing values in the data.

- Also, there were a total of 34 rows of duplicates which were dropped from the data.

- Dropping zeros from the dataset makes more sense in this case because the zero values were present in the 'x', 'y' and 'z' variables which represent 'length', 'width' and 'height' (in mm). Imputing them might highly affect the accuracy and that in turn might affect the business decision taken (if imputed). If the cubic zirconia has a zero length, width, and height, it might imply that the gem(s) does not exist. Including them in the further analysis will also inflate the number of observations when the actual gem(s) itself might not be present and since the data is large, dropping them is not a problem.

- The variables under consideration have different range of values. 'carat', 'x', 'y' and 'z' have a small range compared to 'depth', 'table' and 'price' (which ranges in '00s). In other words, the variables/features under consideration have different scales, hence scaling will make the model more accurate. Hence scaling in such scenario is preferable to obtain accurate results for taking business decisions. Also, if we were to interpret the coefficients of the data, it is better to scale the data.

Note:

Outliers are treated using IQR method.

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

- The variables 'cut', 'color' and 'clarity' are encoded using Label Encoding method as these categorical variables are Ordinal in nature i.e.,

a. Cut- Quality is increasing order Fair, Good, Very Good, Premium, and Ideal which is encoded as 0, 1, 2, 3 and 4 respectively.
b. Colour- With D being the best and J the worst, which is encoded as 6, 5, 4, 3, 2, 1 and 0 (Best to Worst) respectively.
c. Clarity- In order from Best to Worst, (FL = flawless, I3= level 3 inclusions) FL, IF, VVS1, VVS2, VS1, VS2, SI1, SI2, I1, I2, I3, encoded as 10, 9, 8, 7, 6, 5, 4, 3, 2, 1 and 0 respectively.
   This will also help using in avoiding the problem of Dummy-Variable Trap.

- After this step, we convert these categorical variables into 'int64'.

- The data is scaled using 'Standard Scalar', as the main purpose is to interpret the coefficients.

- Here, the dependent variable is: 'price'

  Independent variables: 'carat', 'color', 'cut', 'clarity', 'depth', 'table', 'x', 'y' and 'z'.

- The data is then split in 70:30 ratio i.e., 70% of the data goes to the Train set and the remaining 30% to the Test set (where rows are selected randomly).

  Random State is set to 1 (this ensures the generated splits are reproducible).

	Rows	Variables
X_train	18847	9
X_test	8078	9
Y_train	18847	1
Y_test	8078	1

- We then combine 'X_train' and 'Y_train' as 'trainset' and 'X_test' and 'Y_test' as 'testset' for regression purpose.

- Two ways to perform Linear Regression are: Linear Model (Scikit-Learn) and StatsModel (OLS). The difference between the two are as follows:

	linear model (Scikit-learn)	statsmodel (OLS)
Purpose	Mainly used for prediction purpose	Mainly used for exploratory purpose
Constant	May not necessarily include intercept/constant (unless specified)	Always adds the intercept/constant term

OLS Regression Results:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.931
Model:                            OLS   Adj. R-squared:                  0.931
Method:                 Least Squares   F-statistic:                 2.834e+04
Date:                Fri, 30 Oct 2020   Prob (F-statistic):               0.00
Time:                        18:31:11   Log-Likelihood:                -1472.3
No. Observations:               18847   AIC:                             2965.
Df Residuals:                   18837   BIC:                             3043.
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -0.0003      0.002     -0.153      0.878      -0.004       0.003
carat          1.1845      0.011    107.642      0.000       1.163       1.206
cut            0.0364      0.002     15.498      0.000       0.032       0.041
color          0.1345      0.002     66.558      0.000       0.131       0.138
clarity        0.2075      0.002     97.724      0.000       0.203       0.212
depth          0.0125      0.004      3.196      0.001       0.005       0.020
table         -0.0094      0.002     -3.851      0.000      -0.014      -0.005
x             -0.4379      0.044     -9.963      0.000      -0.524      -0.352
y              0.5035      0.043     11.738      0.000       0.419       0.588
z             -0.1948      0.028     -6.979      0.000      -0.249      -0.140

==============================================================================
Omnibus:                     2652.224   Durbin-Watson:                   2.004
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             9564.804
Skew:                           0.690   Prob(JB):                         0.00
Kurtosis:                       6.206   Cond. No.                         62.9
==============================================================================
```

- **R-square/R^2/Coefficient of Determination** ranges between 0 and 1. It is the measure of goodness-of-fit of the model.
  R^2 = 0.931 implies that all the independent variables (combined) explains 93.1% variation in the dependent variable (Price).

However, it is always better to look at the **Adjusted R^2** value as it does not change with the increase in number of explanatory variables, which is same as the R^2 value in this case.

- **F-test** is useful in Regression model to see whether the model has prediction ability or not. The null and alternate hypothesis for F-test is:

  H_0: $\beta_1 = \beta_2 = \beta_3 = \text{---} = \beta_k$ (i.e., all the coefficients of the model are 0)

  H_1: at least one of the β ('s) is/are not 0

  In other words,

  H_0: There is no predictive ability of the model

  H_1: The model has predictive ability

  If the **Prob(F-statistic) < Level of Significance (α)**, then the model has the predictive ability.

  Here, **Prob(F-statistic) = 0.00 < α = 0.05**, hence the model seems to have the predictive capability.

The **Multiple Linear Regression equation**:

price = -0.00 + (1.18)*carat + (0.04)*cut + (0.13)*color + (0.21)*clarity + (0.01)*depth + (-0.01)*table + (-0.44)*x + (0.5)*y + (-0.19)*z										
s.e.	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.04)	(0.04)	(0.02)

- **Standard Error** of the coefficient measures the distance of the data points from the best fit line. It is always a positive value and measures the estimate's precision. Lower the value, better is the precision of the coefficient.
  From the above equation, we see that the standard error for all the variables are less implying better precision.

- **T-statistic** shows whether the coefficient is statistically significant or not.
  If **(P > |t|) < α = 0.05**, then the variable is statistically significant.
  Here, from all variables except the intercept, (P > |t|) < α = 0.05, implying these variables are statistically significant.

- **[0.025, 0.975]** shows the lower and upper limit of the variables at 95% confidence interval.

- **Skewness** shows the symmetricity in the data and it shows the distribution of residual. The skewness for normal distribution is 0 and value close to 0 implies that the residual is normally distributed. Here, Skew = 0.69 implying the residual are moderately skewed.

- **Kurtosis** is the measure of peak of the curve. If Kurtosis = 3 implies Normal Distribution/Mesokurtic; Kurtosis < 3 implies Platykurtic- Higher it is heavier the tail; Kurtosis > 3 implies Leptokurtic- Lower it is, lighter the tail. Here, the kurtosis value is 6.2 implying the tail(s) are lighter.

- **Jarque-Bera (JB)** test helps us identify whether the residuals are normally distributed or not (Using Skewness and Kurtosis).

  H_0: The residuals are normally distributed

  H_0: The residuals are not normally distributed

  If **Prob(JB) < α = 0.05,** then we reject the null of normality.

  Here, the **Prob(JB) = 0.00 < α = 0.05,** implying the residuals are not normally distributed.

Performance Metrics:

	R-squared	RMSE
Training Set	0.9312	0.2616
Testing Set	0.9316	0.2628

1. **R-square/R^2/Coefficient of Determination** ranges between 0 and 1. It is the measure of goodness-of-fit of the model.

   R^2 for the Training Set = 0.9312 implies that all the independent variables (combined) explains 93.12% variation in the dependent variable (Price).

   R^2 for the Testing Set = 0.9316 implies that all the independent variables (combined) explains 93.16% variation in the dependent variable (Price).

   The R^2 for both Training and Testing Set is almost close to each other, implying the model is good as the independent variables explains the majority of variation in the dependent variable.

2. **Root Mean Square Error** predicts how are the data points concentrated around the best fit line. It is also called as the prediction error. The lower the RMSE, the better is the model.

$$RMSE = \sqrt{\sum \frac{(Predicted - Actual)}{n}}$$

   If 0.2 < RMSE < 0.5, the better is the model.

   RMSE >= 0.5, the model performs poorly in terms of prediction

   The RMSE values for Training and Testing set are both lower implying most of the data points are correctly predicted by the model.

R-square being high and RMSE being low is an ideal situation. Here, with the given dataset, the business decisions can be highly accurate.

Note:

- Multicollinearity is a situation where the independent/explanatory variable is highly correlated with the other.
- Variance Inflation Factor (VIF) helps us to see to what extent the explanatory variables are correlated with the other explanatory variable.
    - ✓ VIF = 1 implies that the explanatory variable(s) are not correlated: Here, it is 'color' and 'clarity'
    - ✓ 1 < VIF < 5 implies moderate correlation: 'cut', 'depth', and 'table'
    - ✓ VIF >= 5 implies highly correlated variables: 'carat', 'x', 'y' and 'z' and needs treatment to avoid multicollinearity problem.

Variables	VIF
carat	32.89
cut	1.51
color	1.12
clarity	1.24
depth	4.45
table	1.62
x	417.37
y	398.58
z	234.84

Price Prediction:

	Scaled Data		Unscaled Data	
Row No.	Predicted	Actual	Predicted2	Actual2
0	1.3134	1.4492	8287.1588	8758
1	0.3240	0.2837	4857.5967	4718
2	2.8106	2.3744	13477.0029	11965
3	2.2816	2.3744	11643.2817	11965
4	1.4089	1.2782	8618.0200	8165
...	...	...	...	...
8073	0.1343	0.2618	4199.9509	4642
8074	0.1693	0.0876	4321.2703	4038
8075	-1.0855	-0.9005	-28.1929	613
8076	-0.9595	-0.8339	408.6273	844
8077	0.4339	0.4222	5238.4442	5198

From the above, we see that the 'price' prediction for both scaled and unscaled data (done using Linear Model (Scikit-Learn)), the predicted and actual values are approximately close to each other (with a small error rate). This shows that the model is quite accurate and business decision based on this might be highly useful for Gem Stones Co Ltd.,

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

From the above analysis, the top five important variables are:

1. **carat**
2. **y**
3. **x**
4. **clarity**
5. **z**

The above five variables mainly influence the price of cubic zirconia (CZ).

- A one unit increase in 'carat' will lead to 1.18 units increase in the price of cubic zirconia.

- A one unit increase in 'y' (Width of cubic zirconia in mm) will lead to 0.50 units increase in the price of cubic zirconia.

- A one unit increase in 'x' (Length of cubic zirconia in mm) will lead to 0.44 units decrease in the price of cubic zirconia.

- A one unit increase in 'clarity' will lead to 0.21 units increase in the price of cubic zirconia.

- A one unit increase in 'z' (Height of cubic zirconia in mm) will lead to 0.19 units decrease in the price of cubic zirconia.

Accordingly, Gem Stones Co Ltd., should concentrate more on increasing the 'carat', improving on its 'y' (Width) and enhancing the 'clarity' as these variables have a positive impact on the price. On the other hand, they should concentrate on decreasing 'x' (Length) and 'z' (Height) as they are inversely related to the price of cubic zirconia. These will help the company in pricing the CZ in such a way that they can extract maximum profit.

Note:

If we see the individual significance of each variable, all the independent variables are significant.

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.

Objective:

1. To predict whether an employee will opt for the package or not on the basis of the information given in the data set.

2. To provide them with the important factors on the basis of which the company will focus on particular employees to sell their packages.

About the Data:

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.**

Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
no	48412	30	8	1	1	no
yes	37207	45	8	0	1	no
no	58022	46	9	0	0	no
no	66503	31	11	2	0	no
no	66734	44	12	0	2	no

- The **dimension/shape** of the dataset: **(872, 7)**
- The number of **missing values** in the dataset = **0**
- The number of **duplicate rows** in the dataset = **0**

Data Type:

'Holliday_Package' and 'foreign' are of Object type, we need to convert them into numerical values for further analysis.

Variables	Data Types
Holliday_Package	object
Salary	int64
age	int64
educ	int64
no_young_children	int64
no_older_children	int64
foreign	object

Measures of Central Tendency:

Variables	Mean	Median
Salary	47729.17	41903.50
age	39.96	39.00
educ	9.31	9.00
no_young_children	0.31	0.00
no_older_children	0.98	1.00

The average 'salary' of the dataset is 47729.17 and the average 'age' is 40 with the average of 9.3 years of 'education'.

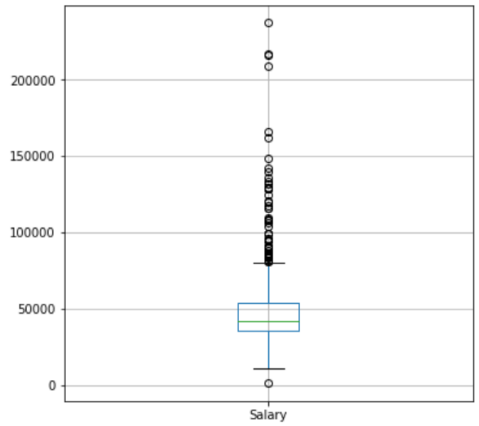
Measures of Dispersion:

Variables	Variance	IQR	CV
Salary	548434000.00	18145.50	0.49
age	111.34	16.00	0.26
educ	9.22	4.00	0.33
no_young_children	0.38	0.00	1.96
no_older_children	1.18	2.00	1.11

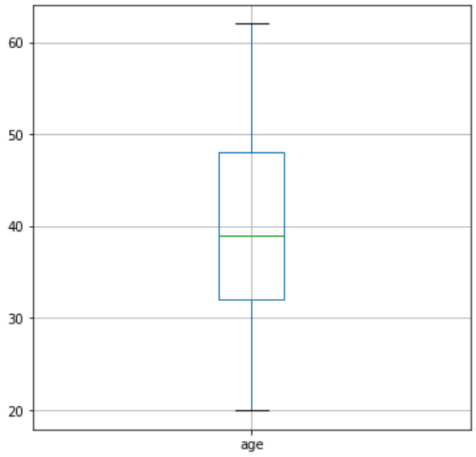
- 'salary' amongst all the variables has the highest variance implying it is highly spread out from the mean.
- The lowest variance, on the other hand, is observed for 'no_young_children' implying data points are identical to one another.
- However, Coefficient of Variation shows a different picture. It allows to check for the dispersion around the mean when the scales of variables under consideration are different from each other. It shows that the highest dispersion around mean is seen in 'no_older_children' and lowest dispersion in 'age'.

Five Number Summary:

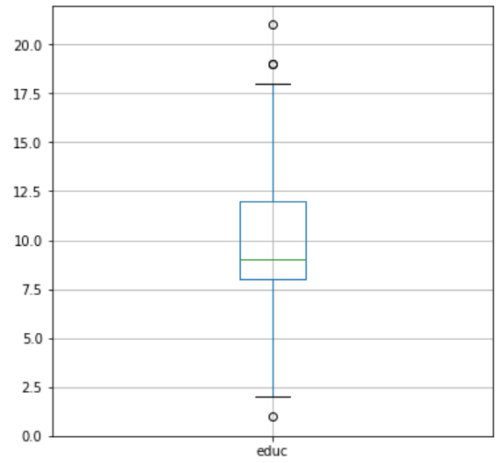
1. Salary



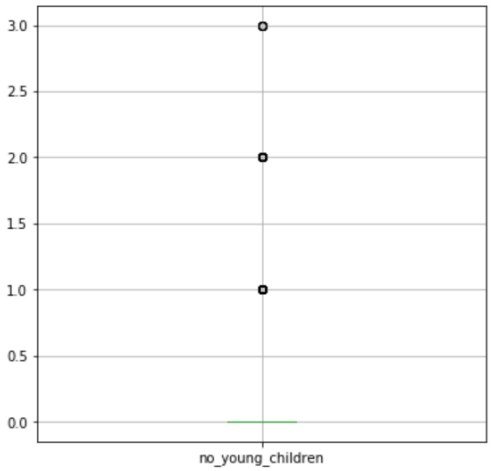
2. Age



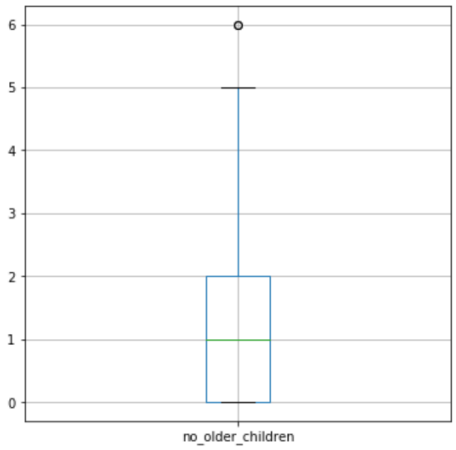
3. Education



4. Number of Young Children



5. Number of Older Children



Here, all the variables except 'age' contains outlier. Since these outliers seems 'genuine', we shall proceed with these for further analysis.

Skewness:

Variables	Skewness
Salary	3.103
age	0.146
educ	-0.046
no_young_children	1.947
no_older_children	0.954

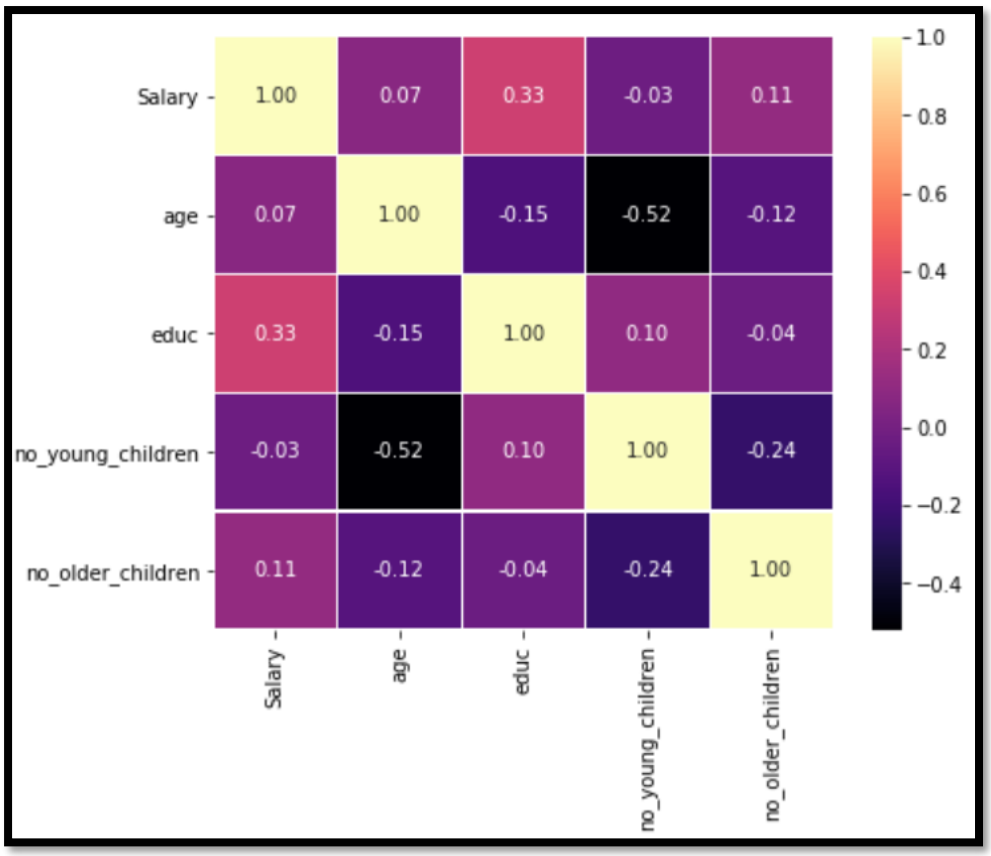
From the above table, 'educ' seems to have negative skewness while 'Salary' has the highest positive skewness.

Covariance:

	Salary	age	educ	no_young_children	no_older_children
Salary	548434000.00	17719.78	23218.66	-425.75	2895.61
age	17719.78	111.34	-4.78	-3.36	-1.33
educ	23218.66	-4.78	9.22	0.18	-0.12
no_young_children	-425.75	-3.36	0.18	0.38	-0.16
no_older_children	2895.61	-1.33	-0.12	-0.16	1.18

- Covariance tells us to what extent two variables change together.
- A positive value indicates that the two variables move in the same direction and vice versa.
- 'age' varies inversely to 'educ', 'no_young_children', 'no_older_children' i.e., 'age' and these variables does not move in the same direction.
- 'salary' variers inversely 'no_young_children' ; 'no_young_children' and 'no_older_children' also vary inversely
- All other variables have a positive values with each other.

Correlation:



- While covariance can only determine the relationship between the two variables, correlation helps us determine the strength between the variables.
- In comparison to covariance, correlation is independent of scale. However, correlation does not imply causation.

- As observed from the above heatmap, there is a high negative correlation between ‘age’ and ‘no_young_children’
- Positive correlation is observed between: ‘educ’ and ‘salary’.

Pair Plot:

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).**

- Here, since the 'Object' type is not ordinal, we can use 'pd.Categorical' (Converting all objects to categorical codes) wherein the binary values (in this case) are assigned as follows:

Holliday_Package	
No	Yes
0	1

foreign	
No	Yes
0	1

Now the data looks like the following (After Encoding the Categorical Data)

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

Checking for Class Balance/Imbalance

Holliday_Package	Proportion
0	0.54
1	0.46

From the above, we see that there is no problem of class imbalance (to some extent it seems balanced).

Data Types (After Encoding)

Variables	Data Types
Holliday_Package	int8
Salary	int64
age	int64
educ	int64
no_young_children	int64
no_older_children	int64
foreign	int8

- Here, the Dependent variable is: 'Holliday_Package'

Independent variables: 'Salary', 'age', 'educ', 'no_young_children', 'no_older_children', 'foreign'.

- The data is then split in 70:30 ratio i.e., 70% of the data goes to the Train set and the remaining 30% to the Test set (where rows are selected randomly).

Random State is set to 1 (this ensures the generated splits are reproducible).

'stratify = y' will ensure that the random split will contain 54% of 0's and 46% of 1's (as in the categorical dependent variable viz 'Holliday_Package' in this case).

	Rows	Variables
X_train	610	6
X_test	262	6
y_train	610	1
y_test	262	1

Logistics Regression: LogisticRegression()

Logit Regression Result:

Logit Regression Results

Dep. Variable:	Holliday_Package	**No. Observations:**	872
Model:	Logit	**Df Residuals:**	865
Method:	MLE	**Df Model:**	6
Date:	Sun, 01 Nov 2020	**Pseudo R-squ.:**	0.1281
Time:	11:27:32	**Log-Likelihood:**	-524.53
converged:	True	**LL-Null:**	-601.61
Covariance Type:	nonrobust	**LLR p-value:**	1.023e-30

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	2.3259	0.554	4.199	0.000	1.240	3.411
Salary	-1.814e-05	4.35e-06	-4.169	0.000	-2.67e-05	-9.61e-06
age	-0.0482	0.009	-5.314	0.000	-0.066	-0.030
educ	0.0392	0.029	1.337	0.181	-0.018	0.097
no_young_children	-1.3173	0.180	-7.326	0.000	-1.670	-0.965
no_older_children	-0.0204	0.074	-0.276	0.782	-0.165	0.124
foreign	1.3216	0.200	6.601	0.000	0.929	1.714

- If **Pseudo R^2** is high, it implies that the model can predict the outcomes in a better manner.
  If 0.2 < Pseudo R^2 < 0.4 implies it is a good model
  Pseudo R^2 < 0.2 then the model might need a different strategy
  Here, Pseudo R^2 = 0.13, which might not be considered as best. However, we further proceed with the model and try using hyperparameters.

- **Log-likelihood**: This is used to derive MLE of the parameter.
  Log-Likelihood (Full Model): Here all the variables including the intercept is considered.

Log-Likelihood (Null Model): Intercept-only model. The rest of the variables at 95% confidence interval is not statistically significant.
If **Log-Likelihood (Full Model) > Log-Likelihood (Null Model)** implies full model is the better model. It is the case in this model that the full model is better than null.

The **Logistic Regression equation** is

logit(p) = log(p/(1-p)) = 2.3 - (1.8e-05) * Salary - (0.04) * age + (0.04) * educ – (1.32) * no_young_children - (0.02) * no_older_children + (1.32) * foreign								
	s.e.	(0.6)	(4.3e-06)	(0.009)	(0.029)	(0.18)	(0.07)	(0.20)

- The coefficient for **Salary** = -1.814e-05 which is interpreted as the expected change in log odds for a one-unit increase in the Salary. The odds ratio can be calculated by exponentiating this value to get 0.9999 which means we expect to see about 0.01% decrease in the odds of the employee opting for Holiday Package, for a one-unit increase in Salary.

- The coefficient for **age** = -0.0482 which is interpreted as the expected change in log odds for a one-unit increase in the age. The odds ratio can be calculated by exponentiating this value to get 0.95 which means we expect to see about 5% decrease in the odds of the employee opting for Holiday Package, for a one-unit increase in age.

- The coefficient for **educ** = 0.0392 which is interpreted as the expected change in log odds for a one-unit increase in the educ. The odds ratio can be calculated by exponentiating this value to get 1.04 which means we expect to see about 4% increase in the odds of the employee opting for Holiday Package, for a one-unit increase in educ.

- The coefficient for **no_young_children** = -1.3173 which is interpreted as the expected change in log odds for a one-unit increase in the no_young_children. The odds ratio can be calculated by exponentiating this value to get 0.27 which means we expect to see about 73% decrease in the odds of the employee opting for Holiday Package, for a one-unit increase in no_young_children.

- The coefficient for **no_older_children** = -0.0204 which is interpreted as the expected change in log odds for a one-unit increase in the no_older_children. The odds ratio can be calculated by exponentiating this value to get 0.98 which means we expect to see about 2% decrease in the odds of the employee opting for Holiday Package, for a one-unit increase in no_older_children.

- The coefficient for **foreign** = 1.3216 which corresponds to the log of odds ratio between foreigner and non-foreigner. The odds ratio equals 3.75 which means the odds for foreigner taking the Holiday Package are about 275% higher than the odds for non-foreigner.

GRID SEARCH:

For GridSearchCV the following were used:

penalty	
none l2	Penalization norm
solver	
newton-cg sag lbfgs	Handles multinomial loss for multiclass problems
tol	
0.0001 0.00001	Stopping criteria
max_iter	
1000	For solvers to converge
n_jobs	
-1 2	Number of simultaneously running workers
cv	
10	Validation set on which the model is not trained and is performed 10 times

- Best Parameters: {'penalty': 'l2', 'solver': 'newton-cg', 'tol': 0.0001}
- Best estimators: {max_iter=1000, n_jobs=2, solver='newton-cg'}

Coefficients after Grid Search CV

Intercept: 2.4277372
Salary: -1.64102826e-05
Age: -5.54454155e-02
Educ: 5.59478398e-02
No_young_children: -1.28781182e+00
No_older_children: -3.98294754e-02
Foreign: 1.19359256e+00

- The coefficient for **Salary** = -1.641e-05 which is interpreted as the expected change in log odds for a one-unit increase in the Salary. The odds ratio can be calculated by exponentiating this value to get 0.9987 which means we expect to see about 0.13% decrease in the odds of the employee opting for Holiday Package, for a one-unit increase in Salary.

- The coefficient for **age** = -0.055 which is interpreted as the expected change in log odds for a one-unit increase in the age. The odds ratio can be calculated by

exponentiating this value to get 0.95 which means we expect to see about 5% decrease in the odds of the employee opting for Holiday Package, for a one-unit increase in age.

- The coefficient for **educ** = 0.056 which is interpreted as the expected change in log odds for a one-unit increase in the educ. The odds ratio can be calculated by exponentiating this value to get 1.06 which means we expect to see about 6% increase in the odds of the employee opting for Holiday Package, for a one-unit increase in educ.

- The coefficient for **no_young_children** = -1.2878 which is interpreted as the expected change in log odds for a one-unit increase in the no_young_children. The odds ratio can be calculated by exponentiating this value to get 0.28 which means we expect to see about 72% decrease in the odds of the employee opting for Holiday Package, for a one-unit increase in no_young_children.

- The coefficient for **no_older_children** = -0.039 which is interpreted as the expected change in log odds for a one-unit increase in the no_older_children. The odds ratio can be calculated by exponentiating this value to get 0.96 which means we expect to see about 4% decrease in the odds of the employee opting for Holiday Package, for a one-unit increase in no_older_children.

- The coefficient for **foreign** = 1.194 which corresponds to the log of odds ratio between foreigner and non-foreigner. The odds ratio equals 3.30 which means the odds for foreigner taking the Holiday Package are about 230% higher than the odds for non-foreigner.

The coefficients of the variables after Grid Search did not change significantly than from the original Logit Regression Results.

Linear Discriminant Analysis: LinearDiscriminantAnalysis()*

*We then use '.fit' on X_train and y_train

*Train & Test Data Class Prediction with a cut-off value of 0.5

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

Performance metrics helps us to evaluate how the models under consideration has performed overall using different criteria.

I. **ACCURACY**: Number of correct predictions made by the model divided by the total number of predictions. Lesser the false prediction, more the accuracy.

II. **CONFUSION MATRIX**: A 2 X 2 matrix reflecting the performance of the model in 4 blocks and is specifically designed for Classification techniques. The y-axis represent Actual Label and x-axis represent Predicted Label.

Here,

- True Negative: The 'no' claimed data which is classified as 'no' claimed.
- False Positive: The 'no' claimed data which is classified as 'yes'. Also known as Type I error.
- True Positive: The claimed data actually classified as claimed.
- False Negative: The claimed data that is classified as 'no'. Also known as a "Type II error.")

a. Accuracy: How accurately does the model classify the data points.

b. Sensitivity: How many of the actual true data points are identified as true data points by the model.

c. Specificity: How many of the actual negative data points are identified as negative by the model.

d. Precision: Among the positives identified as positives by the model, how many are really positive.

III. **CLASSIFICATION REPORT**: This is also specifically used for Classification algorithm. This helps us identifying the quality of the model using: Precision, Recall, f1-Score, Support and Accuracy.

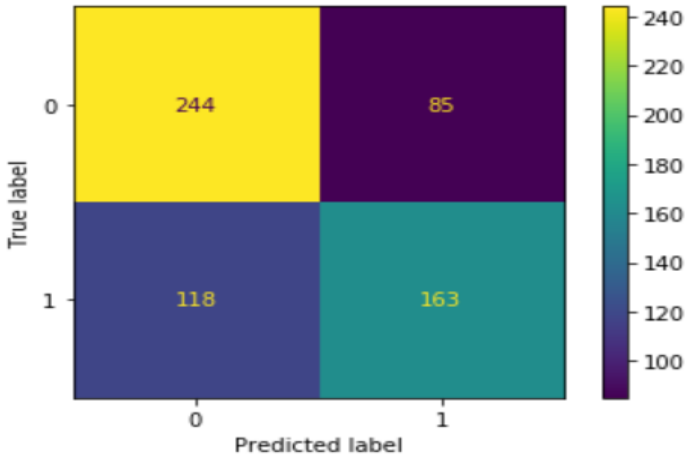
IV. **AREA UNDER THE R.O.C. CURVE**: At various threshold levels, how does the model perform is evaluated by AUC ROC Curve as well as AUC Score are useful to compare the performance of classifier method. The y-axis represents True Positive Rate and x-axis represents False Positive Rate. Graph is a trade-off between Benefits (True Positive) and Costs (False Positive). Larger the area under the curve, better is the model.

LOGISTIC REGRESSION

TRAIN DATASET:

A. Accuracy: 66.72%

B. Confusion Matrix:



True Negative = 244

False Positive = 85

False Negative = 118

True Positive = 163

C. Classification Report:

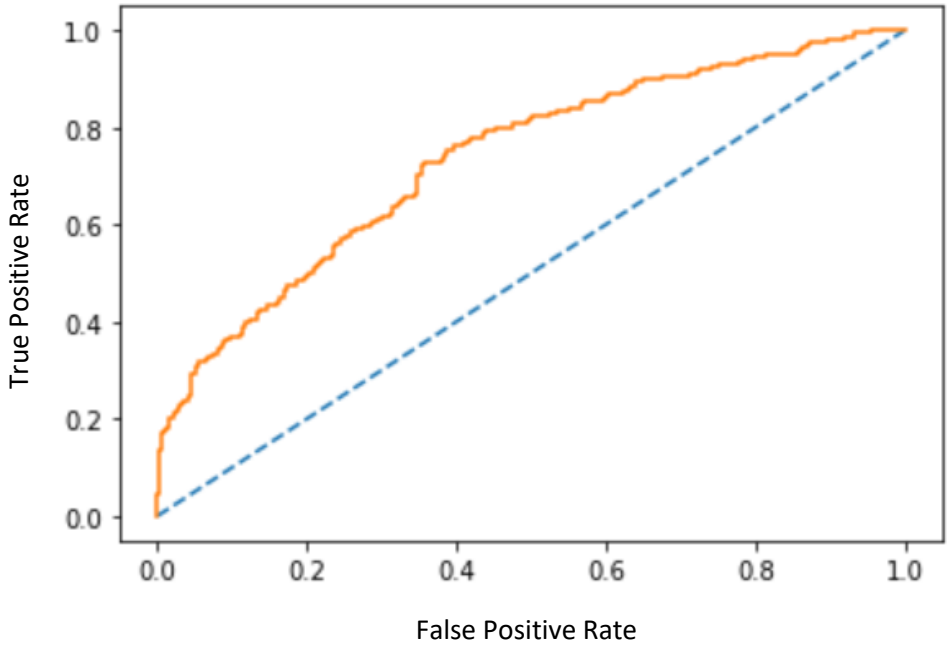
	precision	recall	f1-score	support
0	0.67	0.74	0.71	329
1	0.66	0.58	0.62	281
accuracy			0.67	610
macro avg	0.67	0.66	0.66	610
weighted avg	0.67	0.67	0.66	610

- PRECISION: Percentage of 0s correctly predicted is 67%
  Percentage of 1s correctly predicted is 66%

- RECALL: Percentage of positive cases in 0s is 74%
  Percentage of positive cases in 1s is 58%

- F1-SCORE: Percentage of positive predictions in 0s which were correct is 71%
  Percentage of positive predictions in 1s which were correct is 62%

D. Area Under Curve/Receiver Operating Characteristics Curve:

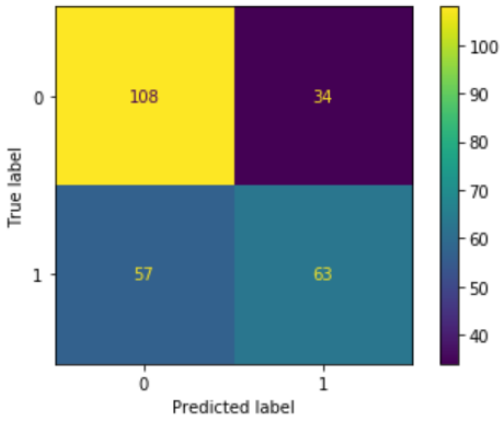


AUC Score = 0.735

TEST DATASET:

A. Accuracy: 65.26%

B. Confusion Matrix:



True Negative = 108

False Positive = 34

False Negative = 57

True Positive = 63

C. Classification Report:

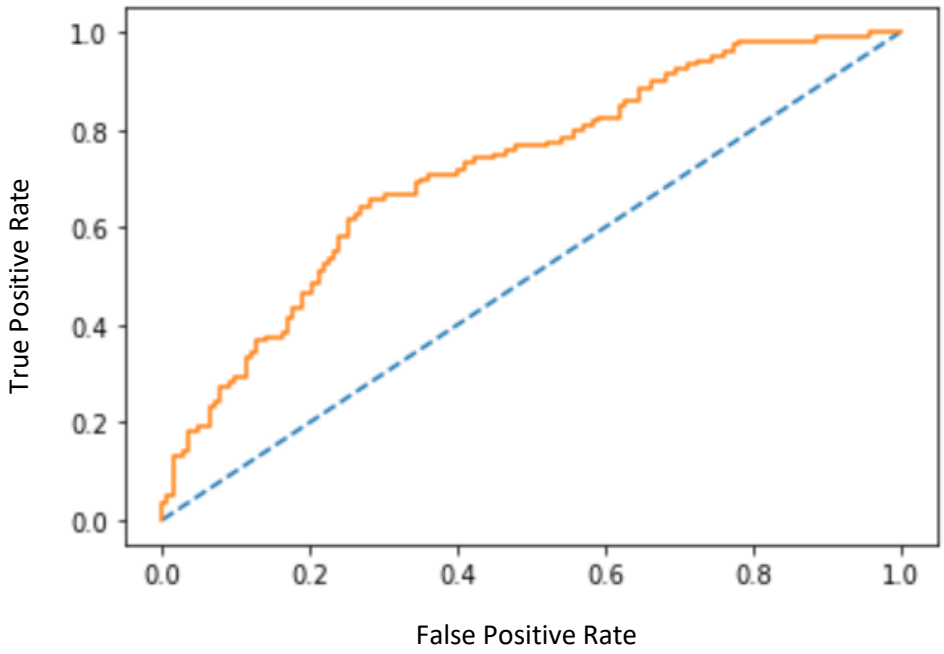
	precision	recall	f1-score	support
0	0.65	0.76	0.7	142
1	0.65	0.53	0.58	120
accuracy			0.65	262
macro avg	0.65	0.64	0.64	262
weighted avg	0.65	0.65	0.65	262

- PRECISION: Percentage of 0s correctly predicted is 65%
  Percentage of 1s correctly predicted is 65%

- RECALL: Percentage of positive cases in 0s is 76%
  Percentage of positive cases in 1s is 53%

- F1-SCORE: Percentage of positive predictions in 0s which were correct is 70%
  Percentage of positive predictions in 1s which were correct is 58%

D. Area Under Curve/Receiver Operating Characteristics Curve:



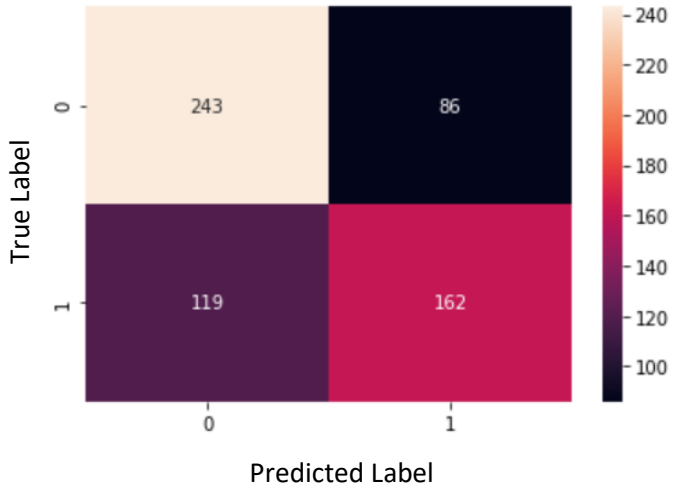
AUC Score = 0.718

LINEAR DISCRIMINANT ANALYSIS

TRAIN DATASET:

A. Accuracy: 66.39%

B. Confusion Matrix:



True Negative = 243

False Positive = 86

False Negative = 119

True Positive = 162

C. Classification Report:

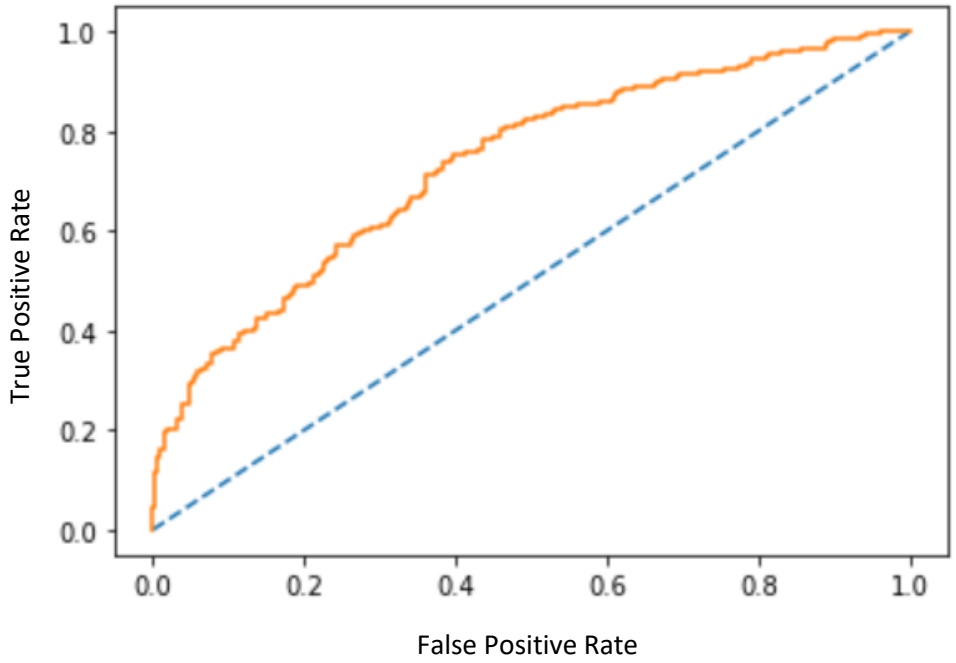
	precision	recall	f1-score	support
0	0.67	0.74	0.7	329
1	0.65	0.58	0.61	281
accuracy			0.66	610
macro avg	0.66	0.66	0.66	610
weighted avg	0.66	0.66	0.66	610

- PRECISION: Percentage of 0s correctly predicted is 67%
  Percentage of 1s correctly predicted is 65%

- RECALL: Percentage of positive cases in 0s is 74%
  Percentage of positive cases in 1s is 58%

- F1-SCORE: Percentage of positive predictions in 0s which were correct is 70%
  Percentage of positive predictions in 1s which were correct is 61%

D. Area Under Curve/Receiver Operating Characteristics Curve:

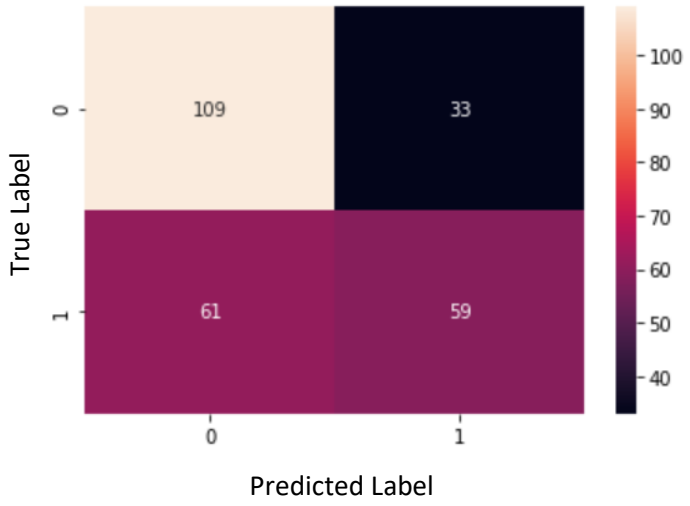


AUC Score = 0.733

TEST DATASET:

A. Accuracy: 64.12%

B. Confusion Matrix:



True Negative = 109

False Positive = 33

False Negative = 61

True Positive = 59

C. Classification Report:

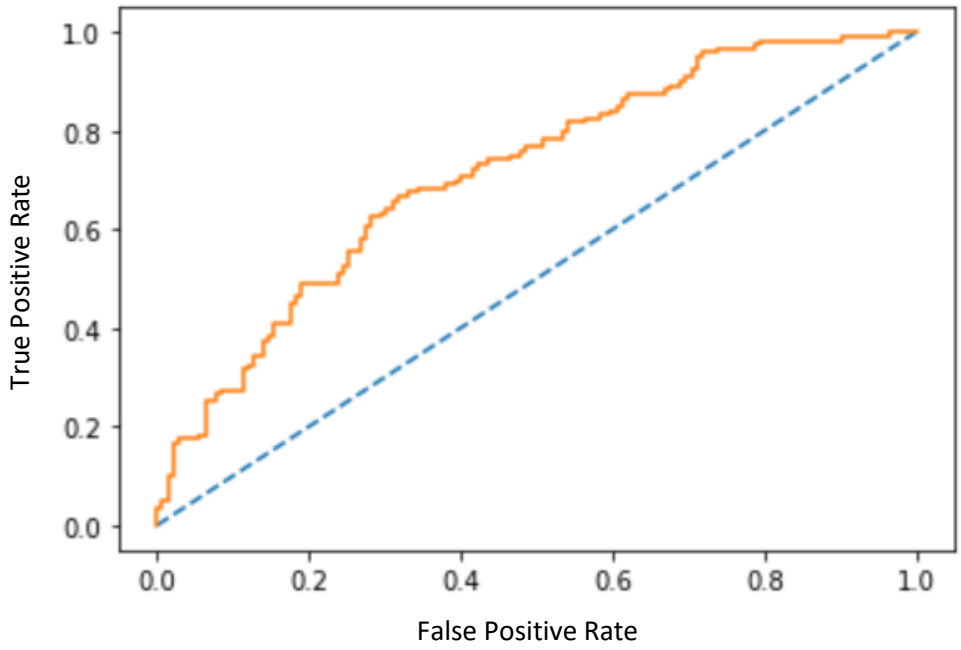
	precision	recall	f1-score	support
0	0.64	0.77	0.7	142
1	0.64	0.49	0.56	120
accuracy			0.64	262
macro avg	0.64	0.63	0.63	262
weighted avg	0.64	0.64	0.63	262

- PRECISION: Percentage of 0s correctly predicted is 64%
  Percentage of 1s correctly predicted is 64%

- RECALL: Percentage of positive cases in 0s is 77%
  Percentage of positive cases in 1s is 49%

- F1-SCORE: Percentage of positive predictions in 0s which were correct is 70%
  Percentage of positive predictions in 1s which were correct is 56%

D. Area Under Curve/Receiver Operating Characteristics Curve:

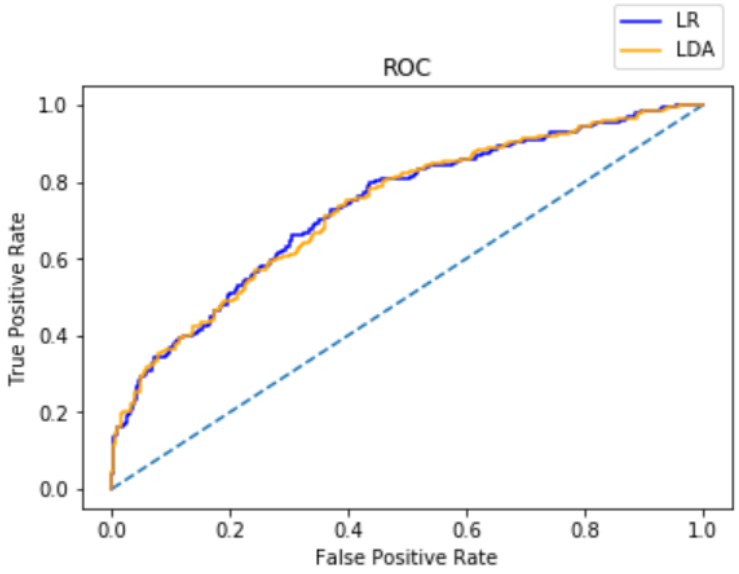


AUC Score = 0.714

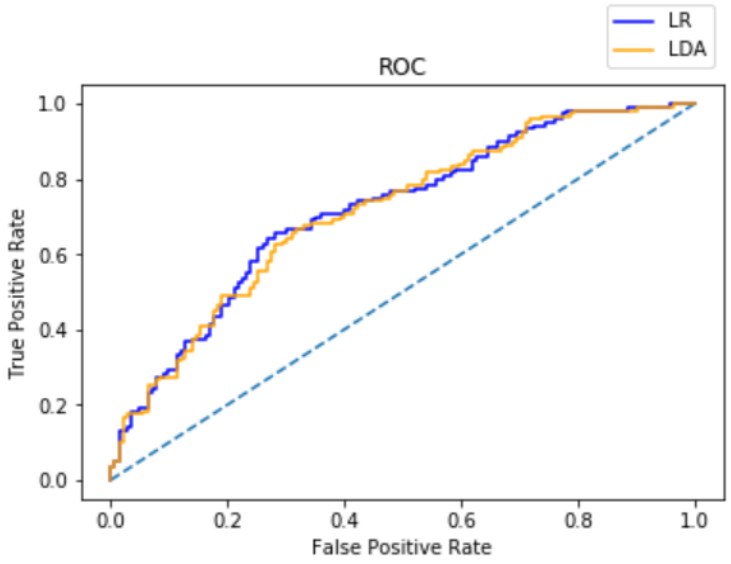
Comparing the Two Models

	LR Train	LR Test	LDA Train	LDA Test
Accuracy	0.67	0.66	0.66	0.64
AUC	0.73	0.72	0.73	0.71
Precision	0.66	0.67	0.67	0.67
Recall	0.57	0.52	0.52	0.52
F1 Score	0.61	0.59	0.59	0.59

1. Logistic Regression:



2. Linear Discriminant Analysis:



- In terms of accuracy, Logistic Regression model is better off than Linear Discriminant Analysis model. However, considering only accuracy (how accurately does the model classify the data points) may not give us a full picture of the solution.

- Looking at the AUC score, we can see that both the model has performed well. However, the LDA's test data AUC Score lesser compared to Logistics'.

- Further, from Precision, which helps us identify how many are really positive among the positives identified as positives by the model LDA model has performed well both on train and test data as compared to Logistic model.

- It is strenuous to compare two models if there is low recall and high precision and vice versa. For this purpose, we can use f1-Score to compare the models. The highest f1-Score is observed for Logistic model.

In conclusion, Logistic Regression model has performed well under almost all the metrics (Accuracy, AUC and f1-Score) as compared to LDA model (with a small difference).

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

The most important variables are:

1. **Foreign**
2. **No_young_children**
3. **Educ**
4. **Age**
5. **No_older_chilren**

- Foreigner employees of the company are more likely to opt for the holiday package as compared to the domestic employees. We can further investigate as to what could plausibly be the reason for same and formulate few strategies to attract more domestic employees of the company (as they might be larger in proportion compared to foreigners in the company). This might further help stimulating the profits for the Tour and Travel agency.

- With the increase in the no of young children in the employees' family lesser are the chances that they are opting for the holiday package. It is very intuitive in the sense that spending on children for education and saving up for them might cause the employees to opt out from the package.

- With the increase in the years of the formal education, higher are the chances they will opt for the package. Hence to some extent, the agency can also take into consideration the no of years of education of the employees while segmenting the customers.

- With increase in the age, people are observed to have lesser and lesser interest in opting for the holiday package. This will help the agency in identifying the segment they have to target on, hence can maximise the profit.

- With higher number of older children in the family, chances that the employees choose for the package will also reduce. This again might be influenced by the fact that children needs investments on their education and families try to save up on that instead of opting for the holiday package. This factor again needs attention as this along with the no of young children in the employees' family may have severe effects on the profitability of the agency. So along with the holiday package provided to such category of the employees', the agency can also provide them with the online class coupons or discounts on few course enrolment for their children to attract them into opting the package. This will help enhancing the profitability of the company.