PREDICTIVE MODELLING

PROJECT

RATEESH UPENDRAN

JULY 2020

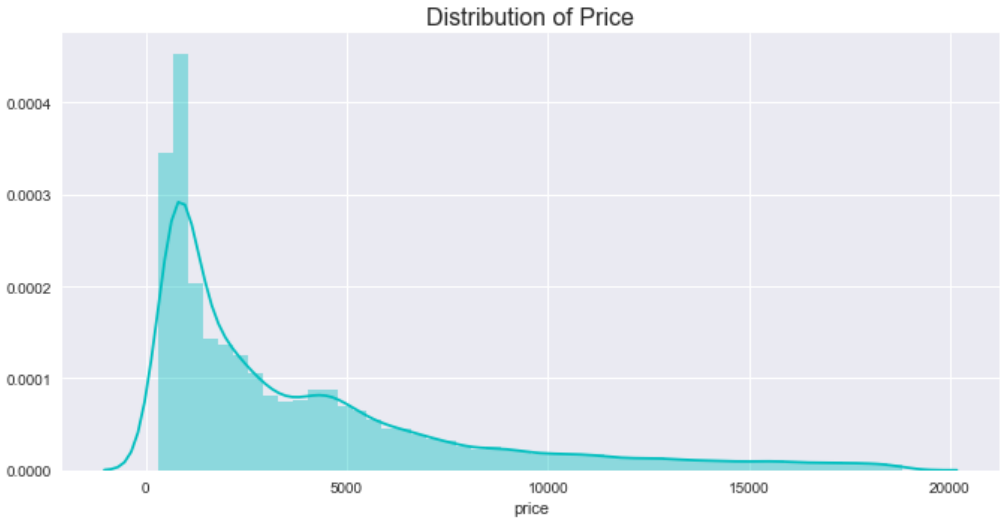
PROBLEM 1: LINEAR REGRESSION

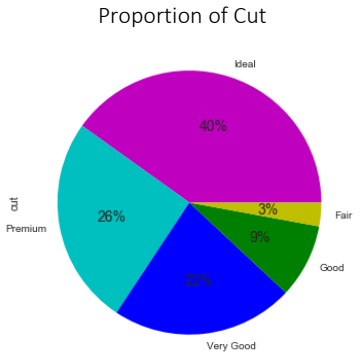
You are hired by a company Gem Stones co ltd, which is a cubic zirconia manufacturer. You are provided with the dataset containing the prices and other attributes of almost 27,000 cubic zirconia (which is an inexpensive diamond alternative with many of the same qualities as a diamond). The company is 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.

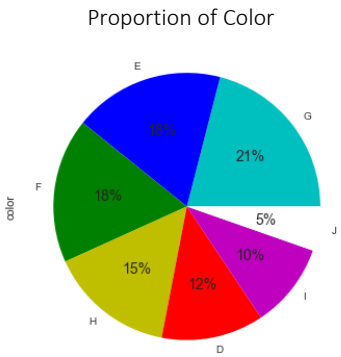
1.1. EXPLORATORY DATA ANALYSIS

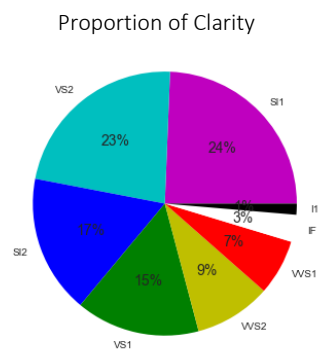
- The dataset contain 26,967 observations and ten variables, including the dependent variable ‘price’ of the zirconia stones
- There are three categoric ordinal variables such as cut, color and clarity, which represent the respective quality of the stone from lower to higher order
- All other six variables are continuous numeric types, where x, y and z indicates dimension of the cubic stones, length, breath and width
- The below descriptive statistical summary provides a high-level view of all the variables
- We can infer from the below summary that average price of a zirconia stone is around 3,940, and a price range of 326 to 18,818, which indicates presence of outliers and a skewed distribution
- There are 697 null values in the variable ‘depth’ which are imputed with the mean of the variable

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
carat	26967	NaN	NaN	NaN	0.798375	0.477745	0.2	0.4	0.7	1.05	4.5
cut	26967	5	Ideal	10816	NaN	NaN	NaN	NaN	NaN	NaN	NaN
color	26967	7	G	5661	NaN	NaN	NaN	NaN	NaN	NaN	NaN
clarity	26967	8	SI1	6571	NaN	NaN	NaN	NaN	NaN	NaN	NaN
depth	26270	NaN	NaN	NaN	61.7451	1.41286	50.8	61	61.8	62.5	73.6
table	26967	NaN	NaN	NaN	57.4561	2.23207	49	56	57	59	79
x	26967	NaN	NaN	NaN	5.72985	1.12852	0	4.71	5.69	6.55	10.23
y	26967	NaN	NaN	NaN	5.73357	1.16606	0	4.71	5.71	6.54	58.9
z	26967	NaN	NaN	NaN	3.53806	0.720624	0	2.9	3.52	4.04	31.8
price	26967	NaN	NaN	NaN	3939.52	4024.86	326	945	2375	5360	18818





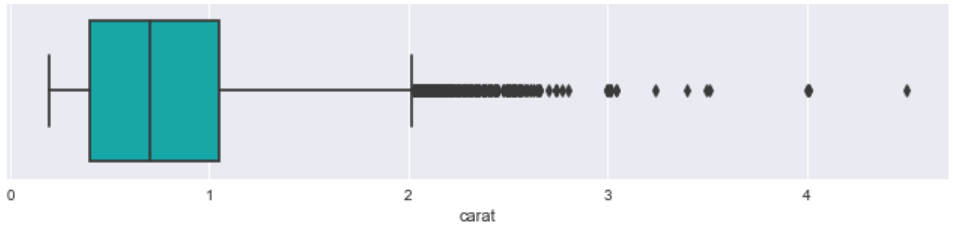




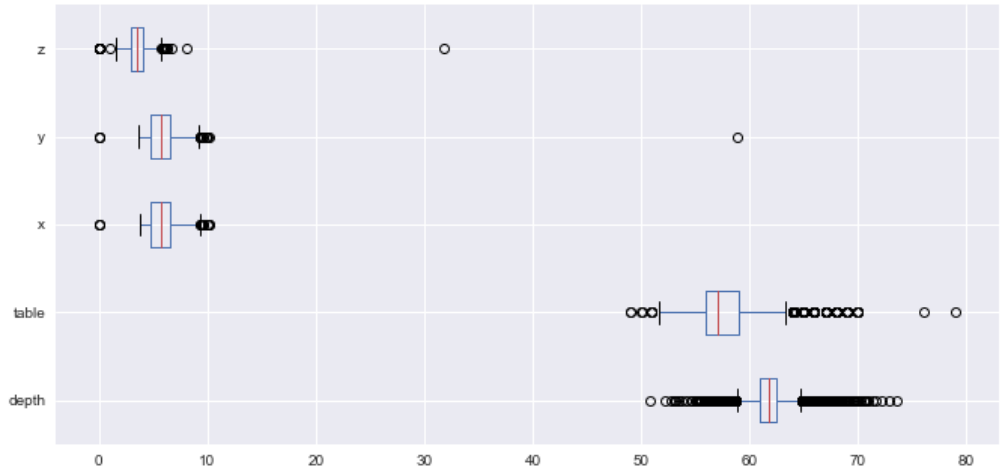
1.1. EXPLORATORY DATA ANALYSIS

- The dimension variables, x, y & z shows a min value of zero, which are dropped as shape of the stone can't be zero and there are only 3 such observations
- The categorical quality variables has a proportion as shown in the pie charts. There are 10,816 stones in the 'Ideal' category, which is of the highest grade, 5661 stones are graded 'G' in color quality and 6571 stones are observed as 'Sl1' in clarity grading
- All the continuous variables in the dataset got significant number of outliers and these variables are in different scales of values, which are removed so that the outliers does not influence the regression line towards them
- From the below plots we can infer that the response variable 'price' got a colinear relationship with variables such as 'carat', 'x', 'y' and 'z'. Whereas there is no significant relation seen with 'depth' and 'table'
- From the heatmap of Pearson correlation of the continuous variables, we can infer that the response variable 'price' got high correlation with 'carat' and the dimension variables (x, y & z)
- From the pairplot and the heatmap we can also see the collinearity of 'carat' is very high with the dimension variables, which indicates that 'carat' is influenced by the size of the stone

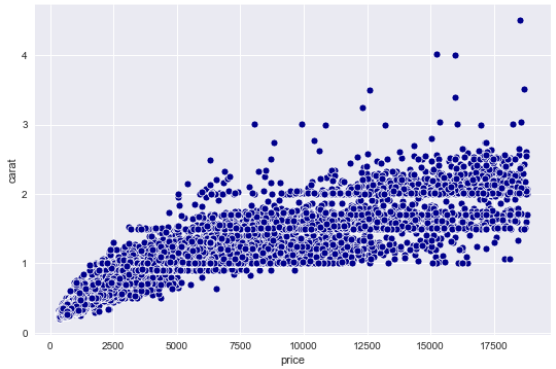
Outliers in variable Carat



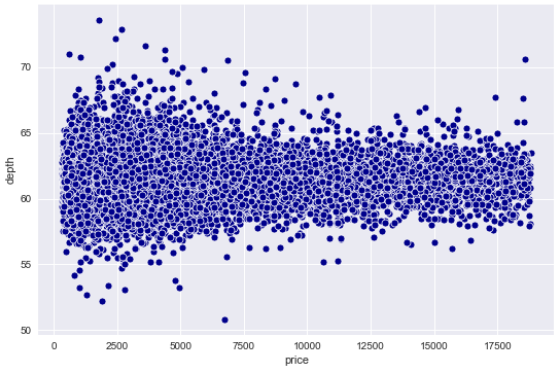
Outliers in other variables



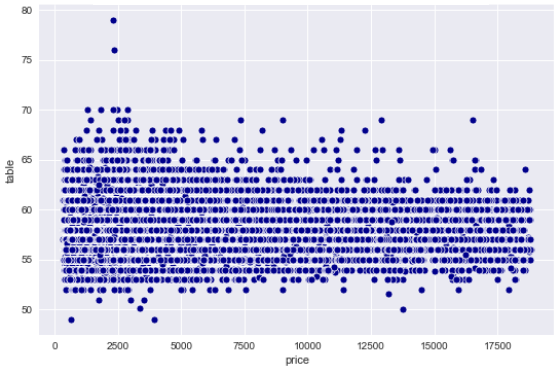
Price Vs Carat



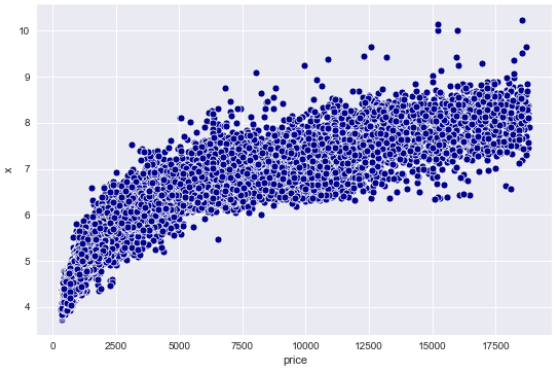
Price Vs Depth



Price Vs Table

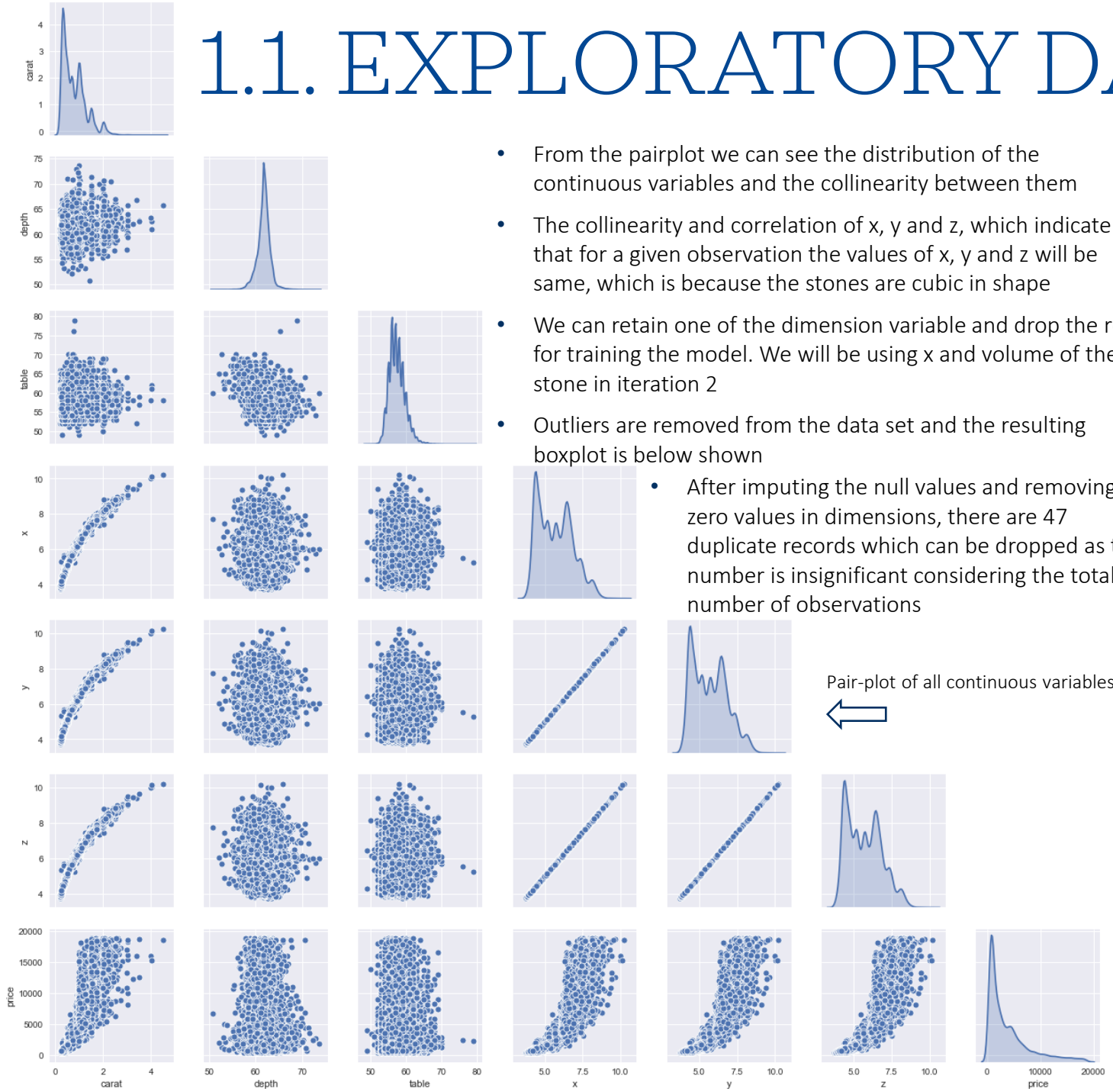


Price Vs X



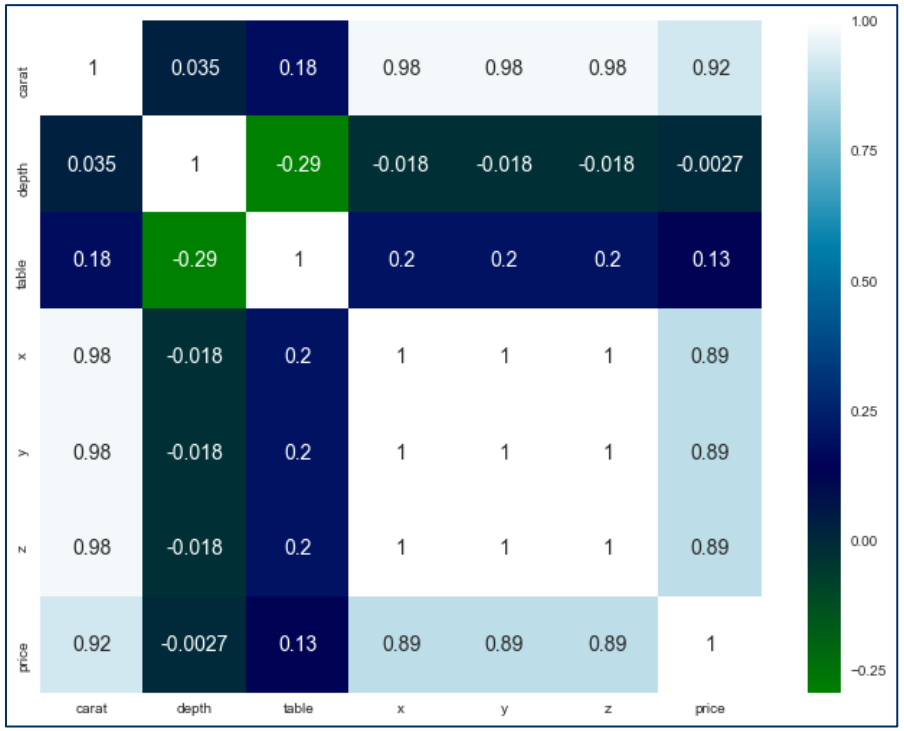
1.1. EXPLORATORY DATA ANALYSIS

- From the pairplot we can see the distribution of the continuous variables and the collinearity between them
- The collinearity and correlation of x, y and z, which indicate that for a given observation the values of x, y and z will be same, which is because the stones are cubic in shape
- We can retain one of the dimension variable and drop the rest for training the model. We will be using x and volume of the stone in iteration 2
- Outliers are removed from the data set and the resulting boxplot is below shown
- After imputing the null values and removing zero values in dimensions, there are 47 duplicate records which can be dropped as the number is insignificant considering the total number of observations

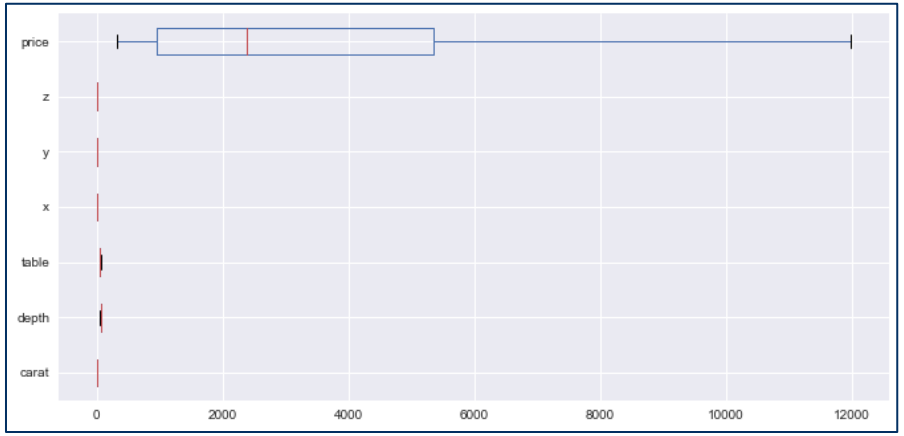


Pair-plot of all continuous variables

Heatmap of correlation between continuous variables



Boxplot of continuous variables after removing outliers



1.2. DATA PRE-PROCESSING

- **Null values:** There are 697 null value observations in the variable 'depth' which are imputed with the mean of the variable
- **Zero values:** The dimension variables, x, y & z shows a min value of zero, which are dropped as shape of the stone can't be zero and there are only 3 such observations
- **Ordinal Encoding:** All the three categoric ordinal variables such as cut, color and clarity, represents the respective quality of the stone from lower to higher order
    - Ordinal encoding is applied in this case as the observation with higher quality category must take precedence over the lower ones in training the model
    - The variables are encoded from zero to higher values in the increasing order of the respective quality attribute
- **Scaling the data:** The different factors in the given dataset is in different scales, which can potentially be standardized.
    - We are going use two methods such as LinearRegression function from sklearn and OLS function from statsmodels in this business case. Both of which use 'ordinary least square' method which may not be influenced by scaling
    - Scaling would be effective if the algorithm use gradient descent method to converge faster
    - Scaling using Z score has been applied in this business case to evaluate the impact of scaling and no significant improvement is found with scaling

Zero values of dimension variables x, y & z

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
carat	26967	NaN	NaN	NaN	0.798375	0.477745	0.2	0.4	0.7	1.05	4.5
cut	26967	5	Ideal	10816	NaN	NaN	NaN	NaN	NaN	NaN	NaN
color	26967	7	G	5661	NaN	NaN	NaN	NaN	NaN	NaN	NaN
clarity	26967	8	SI1	6571	NaN	NaN	NaN	NaN	NaN	NaN	NaN
depth	26270	NaN	NaN	NaN	61.7451	1.41286	50.8	61	61.8	62.5	73.6
table	26967	NaN	NaN	NaN	57.4561	2.23207	49	56	57	59	79
x	26967	NaN	NaN	NaN	5.72985	1.12852	0	4.71	5.69	6.55	10.23
y	26967	NaN	NaN	NaN	5.73357	1.16606	0	4.71	5.71	6.54	58.9
z	26967	NaN	NaN	NaN	3.53806	0.720624	0	2.9	3.52	4.04	31.8
price	26967	NaN	NaN	NaN	3939.52	4024.86	326	945	2375	5360	18818

Null values in the variable depth

```
1  df.isnull().sum()
```

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

Dataset after scaling using Z score

	count	mean	std	min	25%	50%	75%	max
carat	18841.0	-1.394305e-16	1.000027	-1.279710	-0.848277	-0.201127	0.553881	2.657119
cut	18841.0	-9.571335e-16	1.000027	-2.620303	-0.821169	0.078398	0.977964	0.977964
color	18841.0	-3.684166e-16	1.000027	-1.995467	-0.820537	-0.233072	0.941859	1.529324
clarity	18841.0	1.387116e-16	1.000027	-1.860740	-0.638135	-0.026832	0.584470	2.418378
x	18841.0	-4.254922e-16	1.000027	-1.771287	-0.902872	-0.034457	0.727622	3.173362
volume	18841.0	-2.678065e-16	1.000027	-1.261872	-0.843421	-0.209022	0.561104	4.745797
price	18841.0	-2.006722e-16	1.000027	-0.980376	-0.802275	-0.392271	0.463993	2.372026

1.3. LINEAR REGRESSION MODELLING

- **Ordinal Encoding:** All the three categoric ordinal variables such as cut, color and clarity, represents the respective quality of the stone from lower to higher order
  - Ordinal encoding is applied in this case as the observation with higher quality category must take precedence over the lower ones
  - The variables are encoded from zero to higher values in the increasing order of quality attribute
- **Split data:** The data was split in 70:30 ratio for train and test data sets. The response variable 'price' as y and predictor variables as X in train and test data sets were created
- **Modelling:** For this exercise, two linear regression methods from sklearn and statsmodels packages has been applied.
  - OLS method from statsmodel has been applied to get a R like summary of the model
  - Multiple iterations of modelling has been applied to improve the coefficient of determination (R^2) and to improve upon basic assumptions of linear regressions
  - VIF values of the variables were used to validate the multicollinearity and few variables were eliminated for further iterations
- **Model performance:** Model performance measures such as R^2 and Root Mean of Squared Errors (RMSE) is used to validate the linear regression model
  - The predicted labels and the test labels are also plotted on a scatterplot to validate the predictions
  - The assumptions of linear regression such as normality of residuals and homoscedasticity of residuals were also validated using probability plot and residual plot respectively
- **Iteration 2 was selected as final model based on the performance measures**

Ordinal encoding of categoric variables

```
cut_dict = {'Fair': 0,
            'Good': 1,
            'Very Good': 2,
            'Premium': 3,
            'Ideal':4
           }
df.cut = df.cut.map(cut_dict)
df.cut.astype(str).astype(int)

color_dict = {'D': 6,'E': 5,'F': 4,'G': 3,'H'
df.color = df.color.map(color_dict)
df.color.astype(str).astype(int)

clarity_dict = {'FL': 10,
                'IF': 9,
                'VVS1': 8,
                'VVS2': 7,
                'VS1': 6,
                'VS2': 5,
                'SI1': 4,
                'SI2': 3,
                'I1': 2,
                'I2': 1,
                'I3': 0}
df.clarity = df.clarity.map(clarity_dict)
df.clarity.astype(str).astype(int)
```

Train and test datasets

```
1 X_train.shape
(18841, 9)

1 X_test.shape
(8076, 9)

1 y_train.shape
(18841, 1)

1 y_test.shape
(8076, 1)
```

VIF values of variables

	column	VIF
0	carat	107.749383
1	cut	8.843211
2	color	5.535496
3	clarity	12.451131
4	depth	574.800463
5	table	577.205320
6	x	inf
7	y	inf
8	z	inf

Performance measures from three iterations

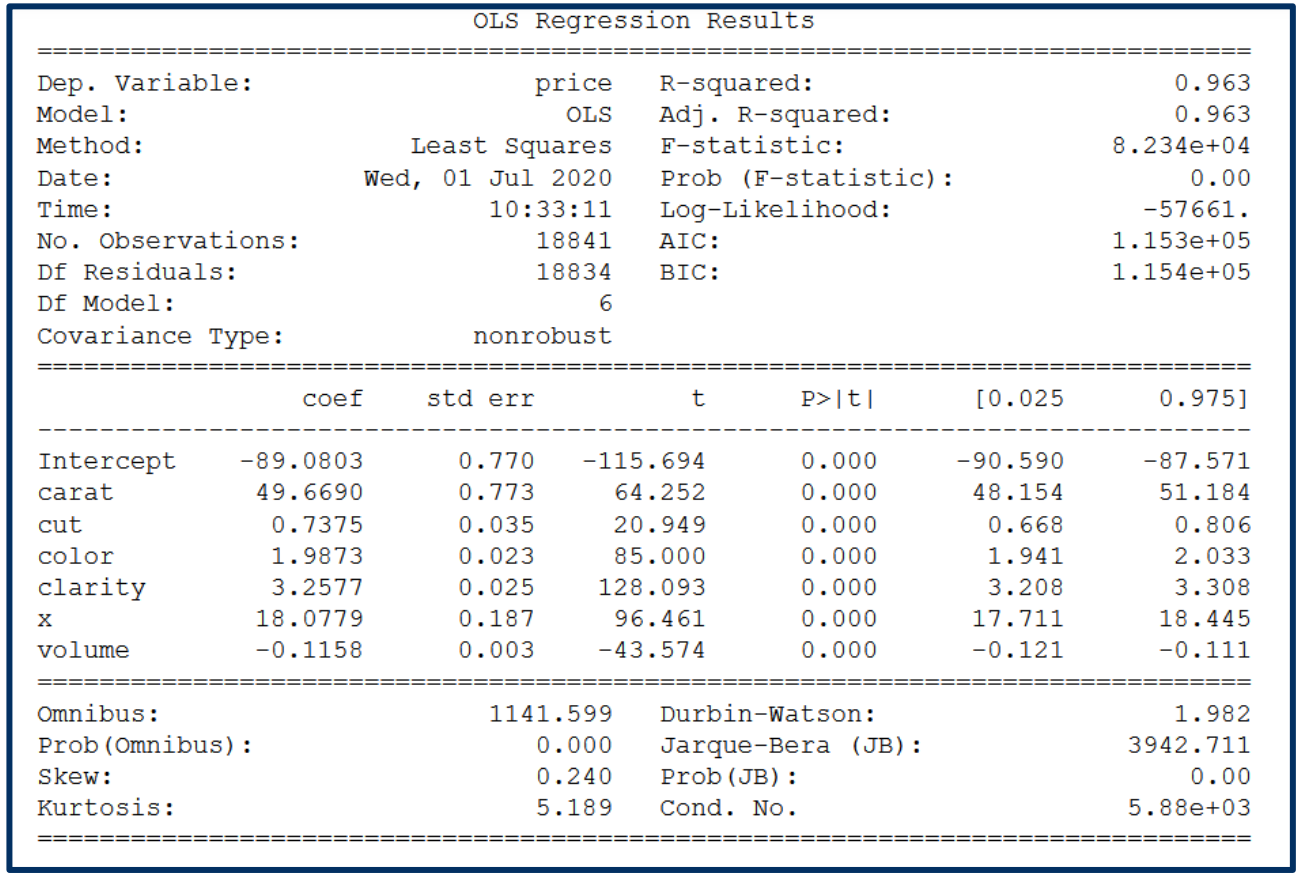
	R^2 Train	R^2 Test	RMSE Train	RMSE Test
Iteration 1	0.932	0.927	905.05	931.92
Iteration 2*	0.963	0.938	5.16	816.75
Iteration 3~	0.932	0.921	0.2606	0.2704

* In iteration 2, the response variable was transformed using square root to train the model.
~ In iteration 3, the dataset was scaled using Z score and OLS method applied.

1.3. LINEAR REGRESSION MODELLING

Summary of the final model

- In iteration 2, the square root of the response variable 'price' was used to normalise the response variable for training the model.
- To validate against test dataset the predictions using the trained model was squared.
- VIF values of x, y and z came out to be infinity as the values are same for each observation
- Volume feature was also introduced from the x, y and z factors. As x, y and z values are same and multicollinear to each other, only 'x' was used in the final model
- The overall P value is less than alpha (0.05), so rejecting the null hypothesis and accepting the alternate hypothesis, that at least one of the predictor variable is influencing the response variable
- That is, at least one regression co-efficient is not zero. Here all regression co-efficient are not zero
- The P value of all the predictor variables are less than alpha, which means that all the variables are statistically significant in deciding the response variable
- The model can explain 96.3%of the variance of price in train as both the R- squared and adjusted R-squared values are 0.963
- In test, the model can explain 93.8% of variation as the R-squared came out to be 0.938, making the model slightly overfit



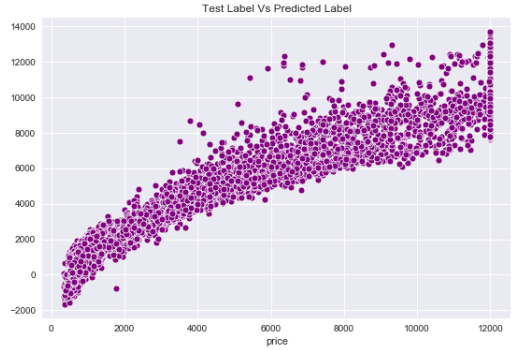
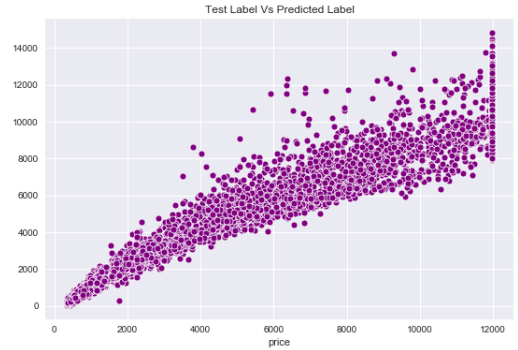
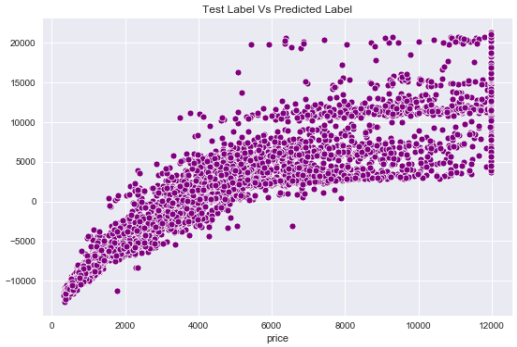
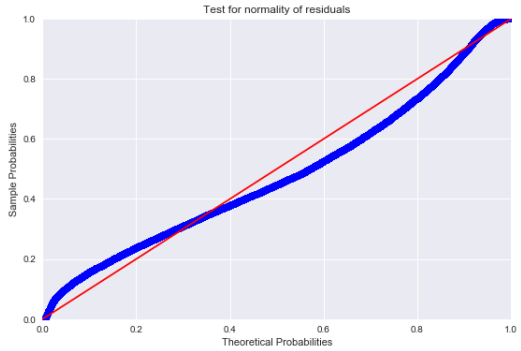
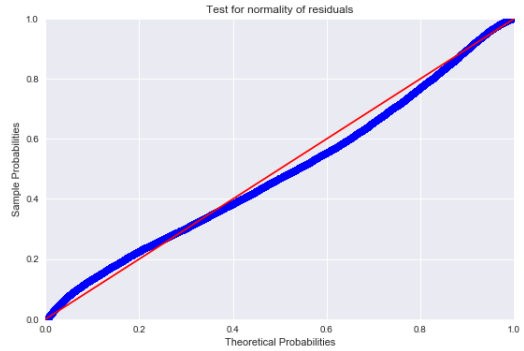
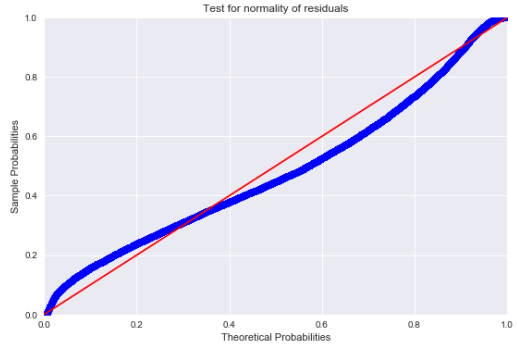
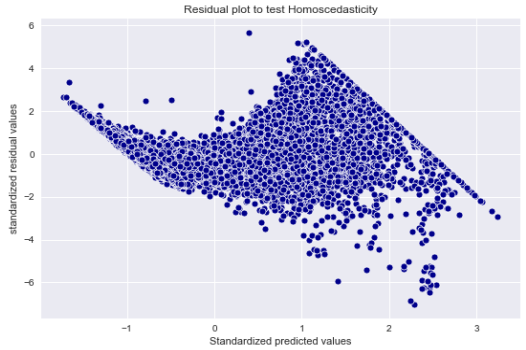
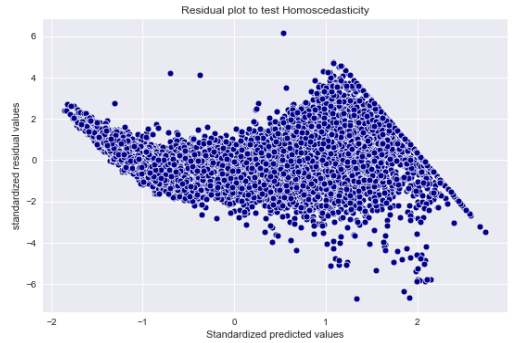
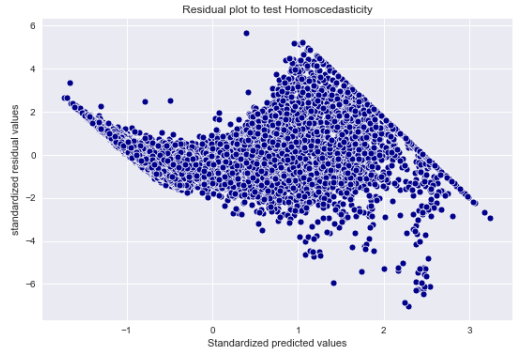
OLS Regression Results

Dep. Variable:	price	R-squared:	0.963
Model:	OLS	Adj. R-squared:	0.963
Method:	Least Squares	F-statistic:	8.234e+04
Date:	Wed, 01 Jul 2020	Prob (F-statistic):	0.00
Time:	10:33:11	Log-Likelihood:	-57661.
No. Observations:	18841	AIC:	1.153e+05
Df Residuals:	18834	BIC:	1.154e+05
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-89.0803	0.770	-115.694	0.000	-90.590	-87.571
carat	49.6690	0.773	64.252	0.000	48.154	51.184
cut	0.7375	0.035	20.949	0.000	0.668	0.806
color	1.9873	0.023	85.000	0.000	1.941	2.033
clarity	3.2577	0.025	128.093	0.000	3.208	3.308
x	18.0779	0.187	96.461	0.000	17.711	18.445
volume	-0.1158	0.003	-43.574	0.000	-0.121	-0.111

Omnibus:	1141.599	Durbin-Watson:	1.982
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3942.711
Skew:	0.240	Prob(JB):	0.00
Kurtosis:	5.189	Cond. No.	5.88e+03

1.3. LINEAR REGRESSION MODELLING

	Iteration 1	Final Model	Iteration 3
Test label Vs Predicted label			
Test of normality of residuals			
Test of homoscedasticity of residuals			

1.4. INSIGHTS AND RECOMMENDATIONS

- The regression coefficient of the predictor variable ‘carat’ is found to be highest (8892.06), which means that for 1 unit increase in carat, the price of a Zirconia stone will increase by 8892 units
- ‘Clarity’ is found to be the next most important influencer in deciding the price with the coefficient of 434.16
- ‘Color’ and ‘cut’ are the next influencers, with coefficient values of 267.77 and 104.56 respectively
- The factors ‘depth’ and ‘table’ are found to be insignificant in deciding the price of Zirconia stones, also VIF score is also high
- Additional factor, volume was introduced in iteration 2 and 3 from the dimension variables, x, y and z
- In the final model, only ‘x’ was used as dimension and it was found to be the second most significant influencer in deciding the price
- In all the three iterations, the test for homoscedasticity was done by plotting the residuals and it was found that in all the three cases it follows a funnel shape pattern. It indicates that the residuals are heteroscedastic in nature, which violates one of the assumptions
- It means scope for further improvement and would recommend further features to be added to improve the prediction of the response variable ‘price’
- The recommendation for the company to identify higher profitable stones is to classify them based on the carat, dimension, clarity, color and cut in the respective order of significance

Regression coefficients from iteration 1

```
The coefficient for carat is 8892.062993696502
The coefficient for cut is 104.55511505651353
The coefficient for color is 267.7662819773506
The coefficient for clarity is 434.15906554192566
The coefficient for depth is -35.00943288069546
The coefficient for table is -11.740996670103042
The coefficient for x is -135.76097535248093
The coefficient for y is -135.76097535248087
The coefficient for z is -135.76097535248087
```

VIF values of variables

	column	VIF
0	carat	107.749383
1	cut	8.843211
2	color	5.535496
3	clarity	12.451131
4	depth	574.800463
5	table	577.205320
6	x	inf
7	y	inf
8	z	inf

Regression coefficients after feature engineering in final model

	coef
Intercept	-89.0803
carat	49.6690
cut	0.7375
color	1.9873
clarity	3.2577
x	18.0779
volume	-0.1158

Best 5 factors in deciding the price

1. Carat
2. Dimension (x/y/z)
3. Clarity
4. Color
5. Cut

PROBLEM 2 : LOGISTIC REGRESSION & 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.

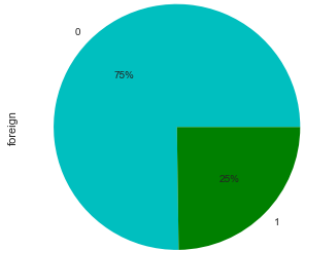
2.1. DESCRIPTIVE STATISTICS

- The dataset contain 872 observations and 7 variables. The dependent variable is Holiday_Package, indicating whether an employee has opted for a holiday package or not
- Salary, age, education, number of young children and number of older children are the continuous independent variables. The categoric variable foreign indicate whether an employee is a foreign national or not
- The average salary of an employee is around 47729.2, maximum is 236961 and minimum 1322. The range of salary indicates presence of outliers
- The average age of an employee is around 40 years and the youngest being 20 years and the oldest employee is of 62 years
- On an average the employees had approximately 10 years of formal education. The highest being 21 years and the lowest at 1 year
- The proportion of foreign nationals in the company is 25%
- The proportion of the target variable 'Holiday_Package' indicate that 46% of the employees opted for a package and the rest 54% didn't
- There are no null values in the dataset
- After removing the outliers from the continuous variables, it is noticed that all the observations for no_of_young_children are in zero values. Thus the variable has been removed before further analysis

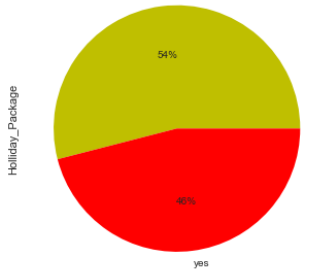
Descriptive statistical summary of data

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Holliday_Package	872	2	no	471	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Salary	872	NaN	NaN	NaN	47729.2	23418.7	1322	35324	41903.5	53469.5	236961
age	872	NaN	NaN	NaN	39.9553	10.5517	20	32	39	48	62
educ	872	NaN	NaN	NaN	9.30734	3.03626	1	8	9	12	21
no_young_children	872	NaN	NaN	NaN	0.311927	0.61287	0	0	0	0	3
no_older_children	872	NaN	NaN	NaN	0.982798	1.08679	0	0	1	2	6
foreign	872	2	no	656	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Proportion of foreign



Proportion of target variable

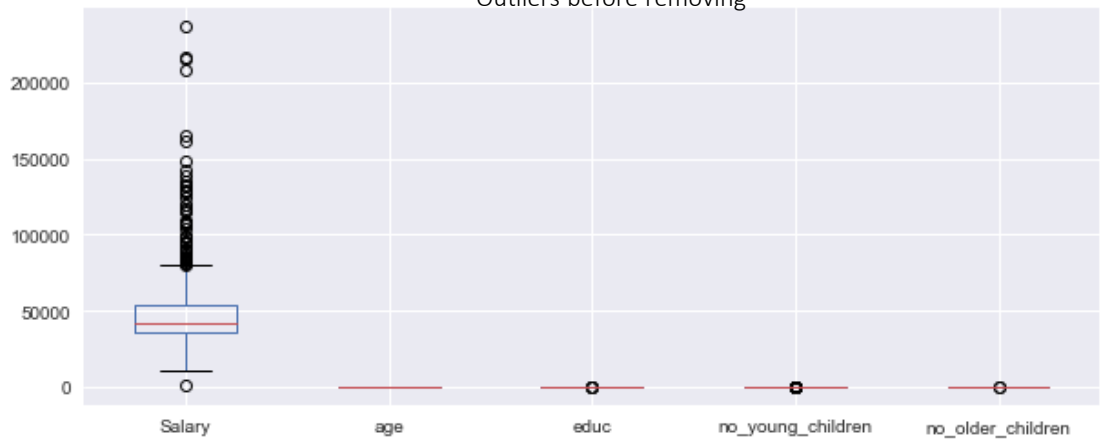


Checking null values

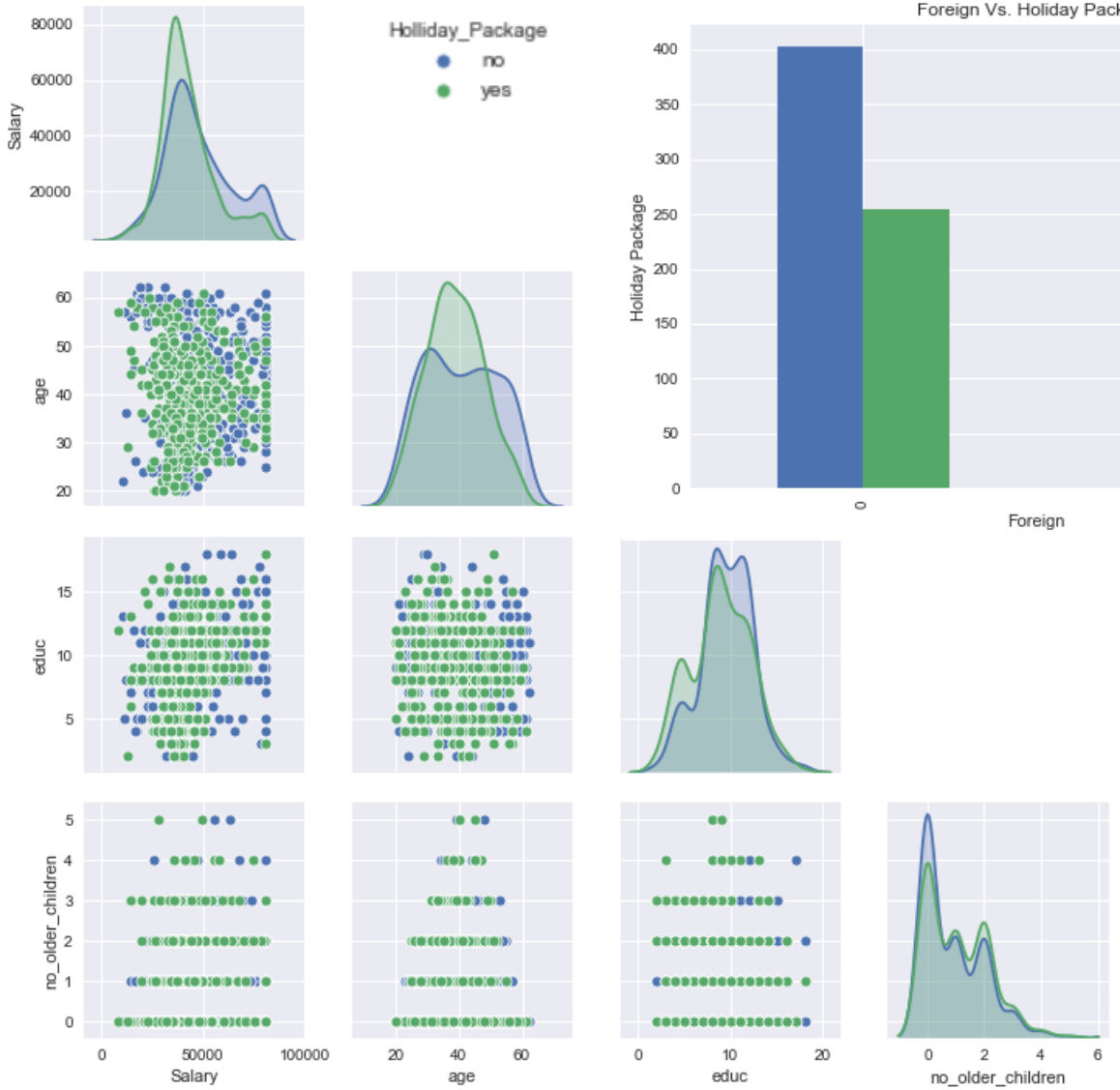
```
1  df.isnull().sum()
```

```
Holliday_Package      0
Salary                0
age                   0
educ                  0
no_older_children     0
foreign               0
```

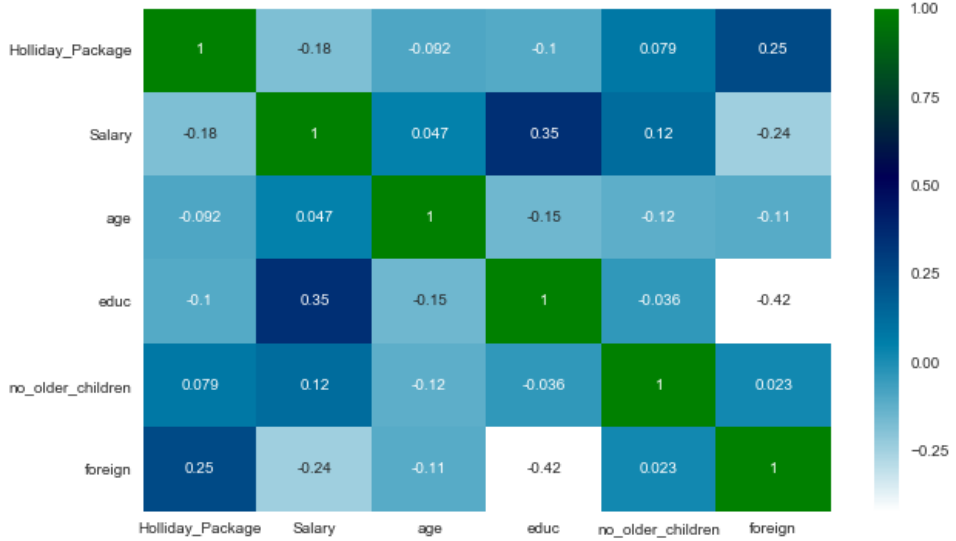
Outliers before removing



2.1. DESCRIPTIVE STATISTICS



Heatmap of Pearson's coefficient of correlation

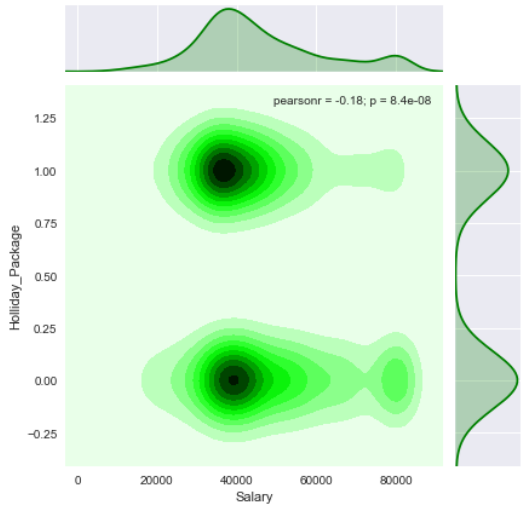


- There is no significant correlation found between the independent and target variable
- Only salary and age shows some level of differentiation in their distribution with respect to an employee's decision to holiday package or not
- Education and no of older children aren't good predictors for the target variable as their distributions overlaps each other
- Employees of foreign nationality are more likely to opt for a holiday package, whereas more local employees opted not to take a holiday package

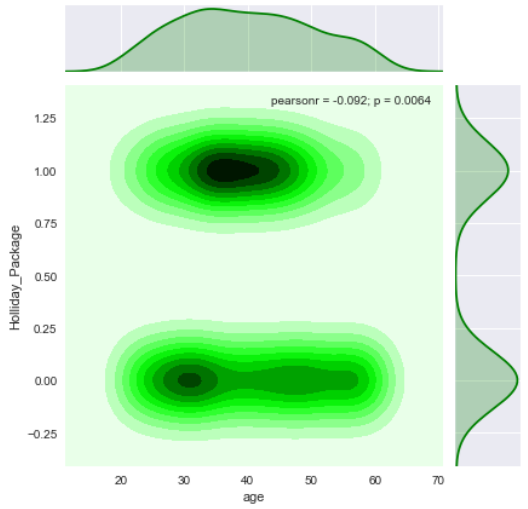
2.1. DESCRIPTIVE STATISTICS

- The pair plot and heatmap of the variables are as in the previous slide
- Jointplots of holiday package Vs other dependent variables such as salary, age, years of education and no of children above 7 years has been plotted
- From the plots, we can infer that there is no significant correlation between salary, age, years of education and no of older children with an employee opting for a holiday package or not
- We can infer that among those who opted for a holiday package, their density is higher around 30000 and 50000 range of salary. But we can't find any relation between those who opted for a holiday package or not
- Similarly age does not have any significant correlation with an employee's decision to opt for a holiday package. However there is a higher density between the age group of 30 to 50 years in opting for a package
- The years of formal education is also found to be poor predictor in deciding whether an employee would opt for a holiday package or not
- The no of older children an employee has, is also turned out to be a very poor predictor as the distribution on status of holiday package is completely overlapping
- Though non of the predictor variables is helpful in deciding the status of the target variable, the salary and age may be able to provide some level of differentiation

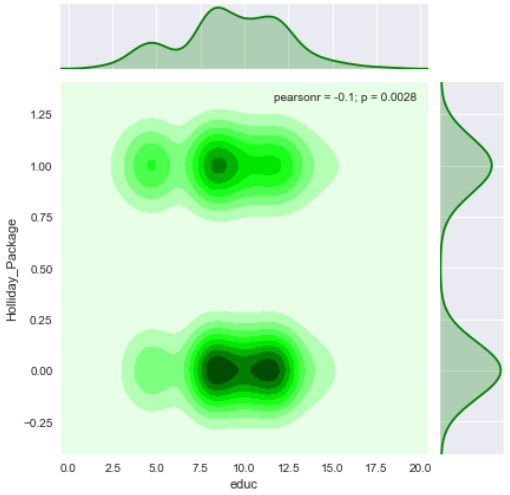
Salary Vs Holiday_Package



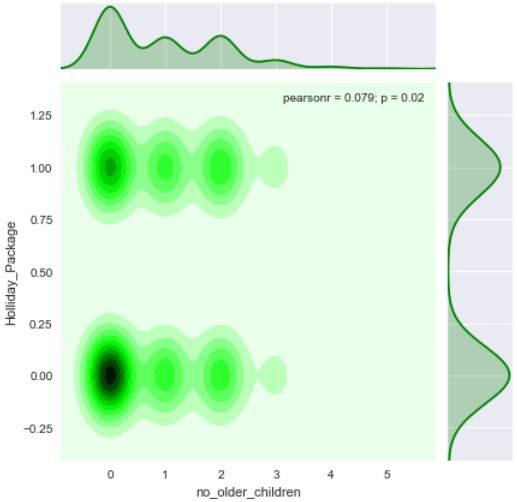
Age Vs Holiday_Packages



Education Vs Holiday_Packages



No of older children Vs Holiday_Packages



2.2. PREDICTIVE MODELLING

Encoding the categoric variables

```
df['Holliday_Package'] = pd.Categorical(df['Holliday_Package']).codes
df['foreign'] = pd.Categorical(df['foreign']).codes
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 872 entries, 0 to 871
Data columns (total 6 columns):
Holliday_Package     872 non-null int8
Salary               872 non-null float64
age                  872 non-null float64
educ                 872 non-null float64
no_older_children    872 non-null float64
foreign              872 non-null int8
dtypes: float64(4), int8(2)
memory usage: 29.1 KB
```

Splitting the data into train & test sets

```
X = df.drop('Holliday_Package', axis=1)

# Copy target into the y dataframe.
y = df['Holliday_Package']

from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30 , random_state=25)
```

Logistic regression modelling using GridSearchCV

```
from sklearn.model_selection import GridSearchCV
clf = LogisticRegression()
grid_values = {'penalty': [ 'l2', 'none']
               ,'C':np.logspace(-3,3,20)
               ,'solver' : ['newton-cg', 'lbfgs', 'sag' ,'saga' ]
              }
logit_model = GridSearchCV(clf, param_grid = grid_values,cv = 10, verbose=True,scoring = 'recall')
logit_model.fit(X_train, y_train)
```

Optimal hyper parameters for Logit

```
logit_model.best_params_
```

```
{'C': 0.004281332398719396, 'penalty': 'l2', 'solver': 'newton-cg'}
```

Modelling Linear Discriminant Analysis

```
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
clf = LinearDiscriminantAnalysis()
LDA_model=clf.fit(X_train,y_train)
LDA_model
```

```
LinearDiscriminantAnalysis(n_components=None, priors=None, shrinkage=None,
                           solver='svd', store_covariance=False, tol=0.0001)
```

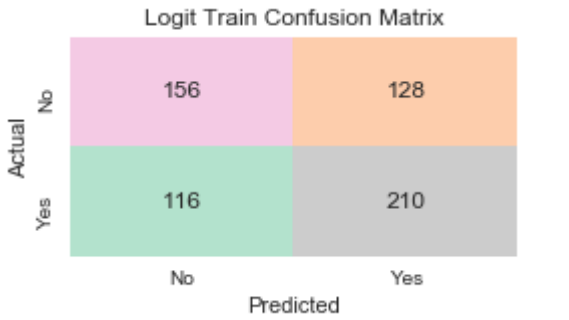
- The categoric variables Holiday_Package and foreign are encoded and converted to integers
- The target positive case ‘yes’ for Holiday_Package is 1 and the negative case ‘No’ is 0
- The data is split in 70:30 ratio into train and test datasets
- LogisticRegression() method from sklearn package is used to build the model
- To optimise the hyper-parameters, GridSearchCV is used
- LinearDiscriminantAnalysis() from sklearn package is used for building the LDA Model
- The classification report and confusion metrics for the models are as in next slide

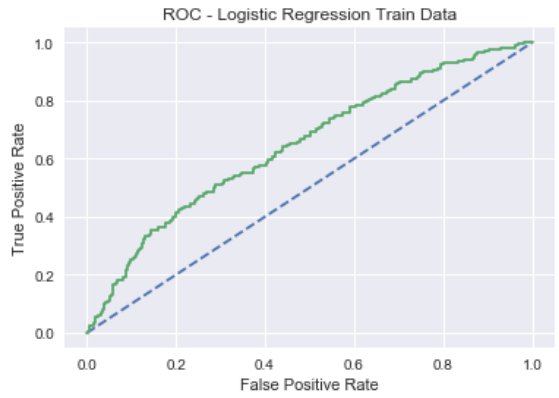
2.3. PERFORMANCE METRICS

Logistic Regression

Train

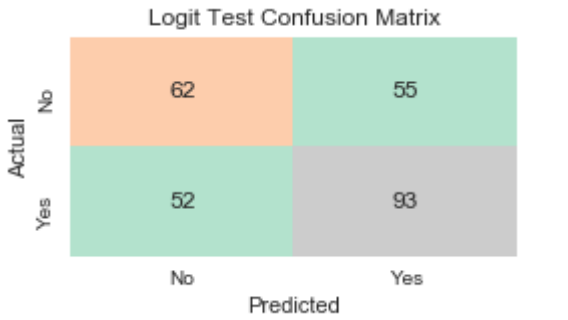
	precision	recall	f1-score	support
0	0.621302	0.644172	0.632530	326.0
1	0.573529	0.549296	0.561151	284.0
accuracy	0.600000	0.600000	0.600000	0.6
macro avg	0.597416	0.596734	0.596841	610.0
weighted avg	0.599060	0.600000	0.599298	610.0

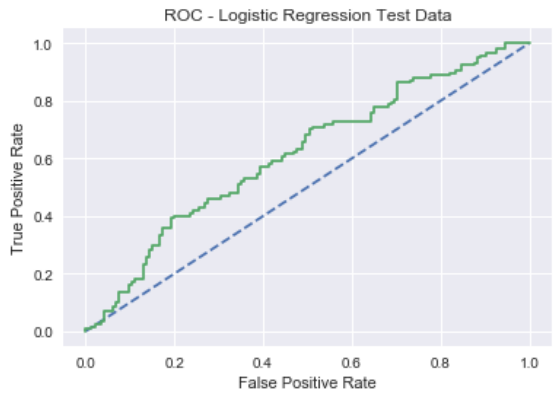




Test

	precision	recall	f1-score	support
0	0.628378	0.641379	0.634812	145.000000
1	0.543860	0.529915	0.536797	117.000000
accuracy	0.591603	0.591603	0.591603	0.591603
macro avg	0.586119	0.585647	0.585804	262.000000
weighted avg	0.590635	0.591603	0.591042	262.000000





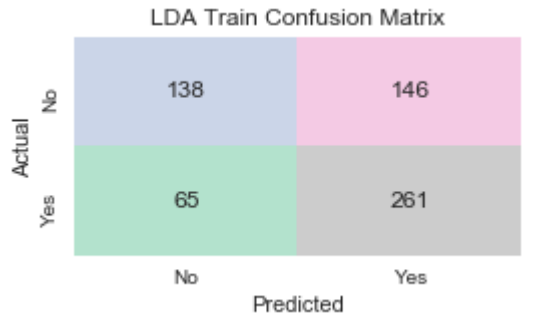
- The Logistic Regression Model was optimised to improve the recall rate for the positive target case
- While the accuracy and precision dropped a little, the recall rate improved significantly
- The model produced an accuracy of 60% in train and 59% in test
- The precision for the positive target case in train is 58% and in test it is 54%
- As recall is considered as the most significant performance measure for this business case, the recall rate for the positive target case in train is 55% and in test it is 53%
- The confusion metrices for both the train and test is given here
- While the recall rate was optimised, the Area Under Curve (AUC) value also reduced significantly
- AUC for train is 0.65 and for test it is 0.61
- Based on accuracy values, the model seems to be a right fit one

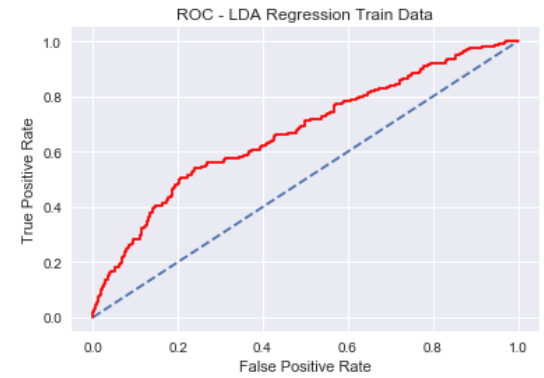
2.3. PERFORMANCE METRICS

Linear Discriminant Analysis

Train

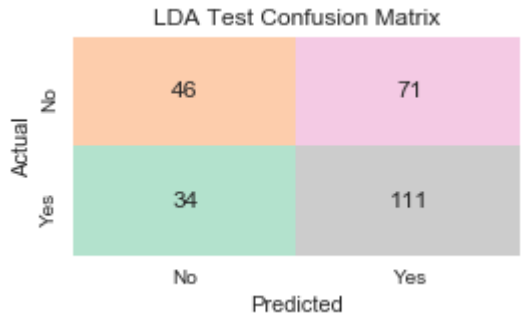
	precision	recall	f1-score	support
0	0.641278	0.800613	0.712142	326.000000
1	0.679803	0.485915	0.566735	284.000000
accuracy	0.654098	0.654098	0.654098	0.654098
macro avg	0.660540	0.643264	0.639438	610.000000
weighted avg	0.659214	0.654098	0.644444	610.000000

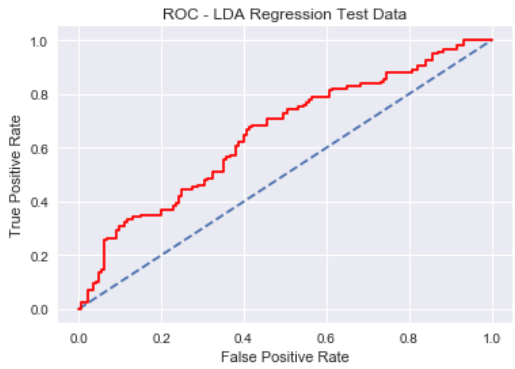




Test

	precision	recall	f1-score	support
0	0.609890	0.765517	0.678899	145.000000
1	0.575000	0.393162	0.467005	117.000000
accuracy	0.599237	0.599237	0.599237	0.599237
macro avg	0.592445	0.579340	0.572952	262.000000
weighted avg	0.594309	0.599237	0.584275	262.000000



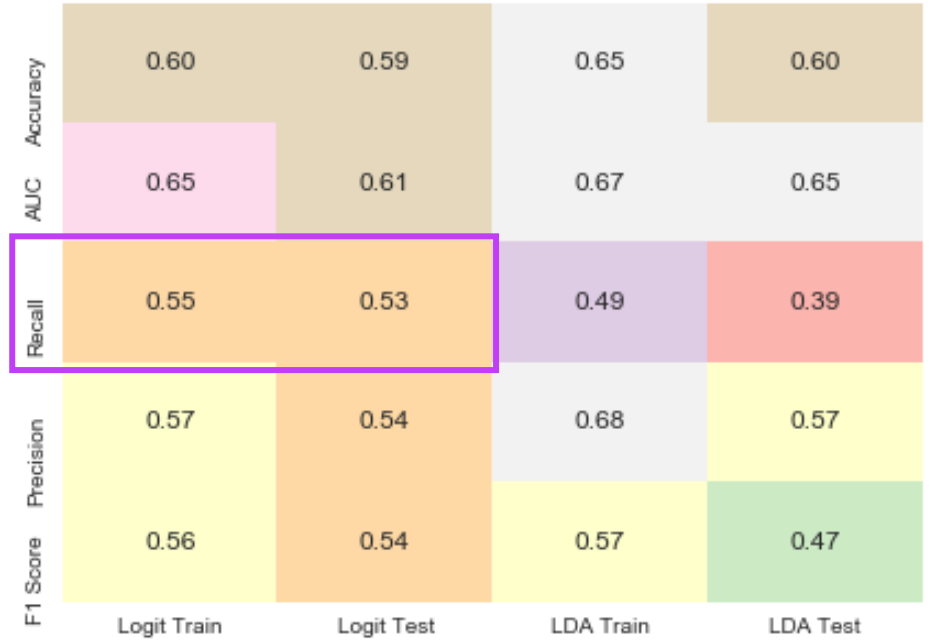


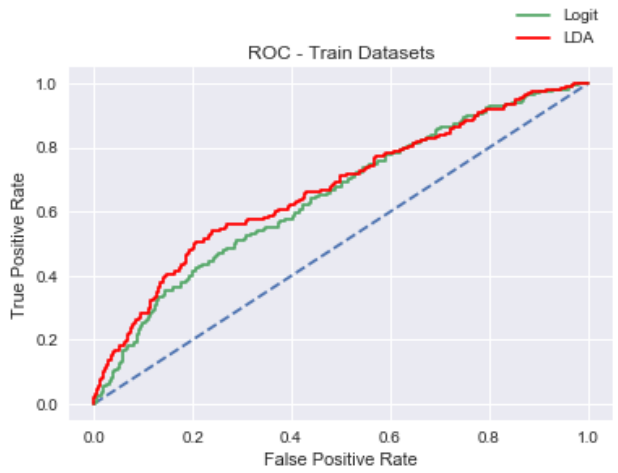
- The LDA model produced an accuracy of 65% in train and 60% in test
- The precision for the positive target case in train is 68% and in test it is 58%
- As recall is considered as the most significant performance measure for this business case, the recall rate for the positive target case in train is 49% and in test it is 39%
- The confusion metrices for both the train and test is given here
- AUC for train is 0.67 and for test it is 0.65
- The recall rate came out to be too low and the model is found to be overfitting in terms of accuracy
- The derived model is not reliable to predict the target cases

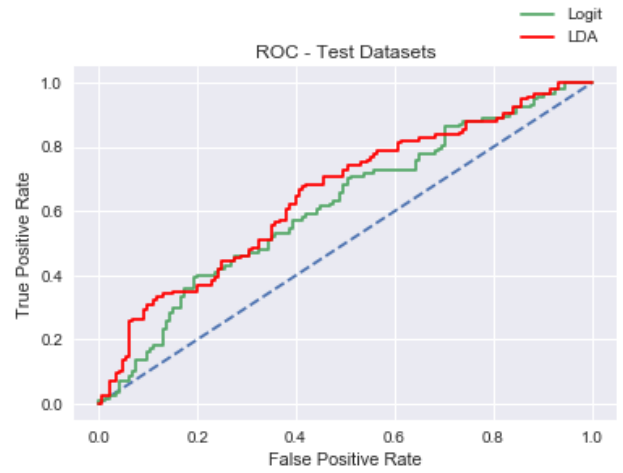
2.3. PERFORMANCE METRICS

Final model comparison

- The accuracy of the LDA model is higher than that of the Logit model, 60% and 59% respectively for the test dataset
- The precision for the test dataset is also higher in the LDA model than the Logit model, 57% and 54% respectively
- The AUC score on the test dataset from the LDA model is significantly higher than the Logit model, 0.65 and 0.61 respectively
- However considering the business case the ability to recall the positive target case is important than other performance measures
- The recall score is higher from the logit model at 53% for the test dataset, were as the recall score from the LDA model is only 39% which is unacceptable
- The overall F1 score is also higher in the Logit model at 54% for test data. Whereas it is only 47% for test data in the LDA Model
- The ROC curve shows a better coverage in the LDA model, but based on the recall and F1 score Logit model is opted as the final model for this business case







2.4. INSIGHTS & RECOMMENDATIONS

- Both the models has not resulted in a very optimistic model to predict the target variable, which is because of the fact that none of the given predictor variables could differentiate the positive and negative target cases
- The resulted predictions are more akin to ‘toss of a coin’ than a reliable one to predict the target response variable
- The recall rate 53% indicates that the model has been able to recollect only half of the actual cases of employees opting for a holiday package
- The travel company can provide tailored packages based on age groups to attract employees at different age levels to opt for respective packages
- Value add-ons could be considered based on number of children the employee has and the age of the children.
- Foreign nationals can be encouraged with targeted add-ons and specific packages based on their interest which will vary from that of locals
- Would recommend the travel company to identify new features which can positively correlate the pattern of opting for a holiday package or not, for training and building a better model
- Features such as marital status, gender, interest categories in travel(adventure, beach side, hill stations, theme parks) etc could be considered as additional features


THANK YOU !