PREDICTIVE MODELLING

BUSINESS REPORT

SHOBANADEVI R

TABLE OF CONTENTS

Introduction of Problem Case:

To help the company in predicting the price for the stone on the basis of the details given in the dataset so it can distinguish between higher profitable stones and lower profitable stones so as to have better profit share. Also, provide them with the best 5 attributes that are most important.

DATA SECTION:

DESCRIPTIVE DATA ANALYSIS

- The Dataset has 26967 Rows and 11 Columns.
- The datatypes are

```
: Unnamed: 0      int64
  carat         float64
  cut            object
  color          object
  clarity        object
  depth         float64
  table         float64
  x             float64
  y             float64
  z             float64
  price           int64
  dtype: object
```

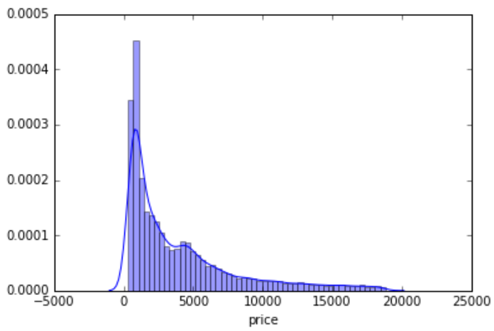
- Exploratory Data Analysis - Describing about mean and quantile ranges

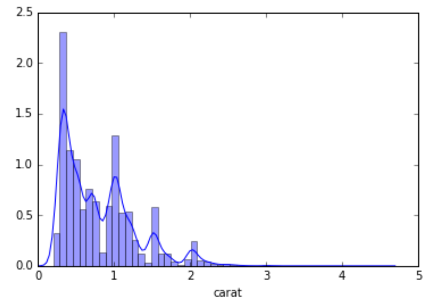
	carat	cut	color	clarity	depth	table	x	y	z	price
count	26967.000000	26967	26967	26967	26270.000000	26967.000000	26967.000000	26967.000000	26967.000000	26967.000000
unique	NaN	5	7	8	NaN	NaN	NaN	NaN	NaN	NaN
top	NaN	Ideal	G	SI1	NaN	NaN	NaN	NaN	NaN	NaN
freq	NaN	10816	5661	6571	NaN	NaN	NaN	NaN	NaN	NaN
mean	0.798375	NaN	NaN	NaN	61.745147	57.456080	5.729854	5.733569	3.538057	3939.518115
std	0.477745	NaN	NaN	NaN	1.412860	2.232068	1.128516	1.166058	0.720624	4024.864666
min	0.200000	NaN	NaN	NaN	50.800000	49.000000	0.000000	0.000000	0.000000	326.000000
25%	0.400000	NaN	NaN	NaN	61.000000	56.000000	4.710000	4.710000	2.900000	945.000000
50%	0.700000	NaN	NaN	NaN	61.800000	57.000000	5.690000	5.710000	3.520000	2375.000000
75%	1.050000	NaN	NaN	NaN	62.500000	59.000000	6.550000	6.540000	4.040000	5360.000000
max	4.500000	NaN	NaN	NaN	73.600000	79.000000	10.230000	58.900000	31.800000	18818.000000

- Null Value Check- It contains 697 null values. Imputing it with the values of MEAN.

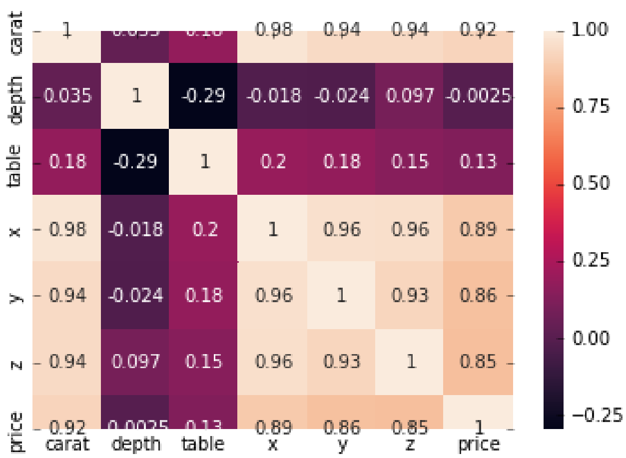
```
|: Unnamed: 0      0
   carat           0
   cut             0
   color           0
   clarity         0
   depth         697
   table           0
   x               0
   y               0
   z               0
   price           0
   dtype: int64
```

- Univariate Analysis :The Price range is mostly lies between 1000-5000
- In the dataset the weight of the carat has continuous decreasing.





- Multivariate Analysis:The multivariate analysis explains the correlation and covariance between the variables, Here Carat, X,Y,Z are correlated to each one.
- Moreover the covariance leads to multicollinearity and thus give poor performance of a model.Check for variance inflation factor and need to decide whether to keep the variable or not.



METHOD SECTION:

Here to predict the Price for the stone with its feature variables we are using **‘*LINEAR REGRESSION’.***

NULL CORRECTION - Imputing the null values with their respected means will make the dataset more powerful for modelling.The variables doesn't have any values equal to Zero.

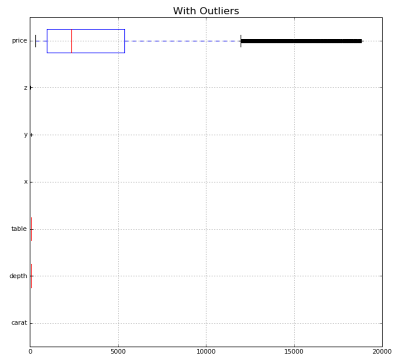
```
carat      0
cut        0
color      0
clarity    0
depth      0
table      0
x          0
y          0
z          0
price      0
dtype: int64
```

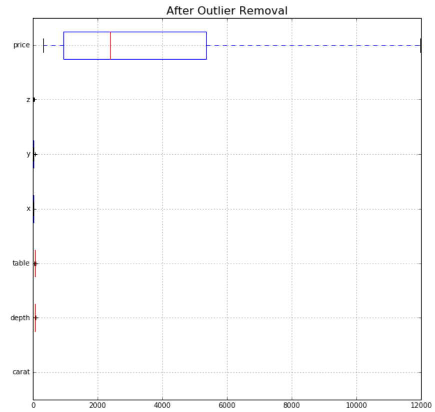
carat	cut	color	clarity	depth	table	x	y	z	price

Duplicates: Removing the duplicates from the dataset.

```
Before (26967, 10)
After (26933, 10)
```

OUTLIER TREATMENT: The price column has more outliers treating them with the quantile ranges will not change any original form of data.





SCALING : Yes Scaling is necessary,It is always a good practice to scale all the dimensions using z scores or some other methods to address the problem of different scales.

ENCODING:

Changing the categorical variables to code will make the model to predict easily.

	carat	depth	table	x	y	z	price	cut_Good	cut_Ideal	cut_Premium	...	color_H	color_I	color_J	clarity_IF	clarity_SI1	clarity_SI2	clarity_VS1
0	0.30	62.1	58.0	4.27	4.29	2.66	499	0	1	0	...	0	0	0	0	1	0	0
1	0.33	60.8	58.0	4.42	4.46	2.70	984	0	0	1	...	0	0	0	1	0	0	0
2	0.90	62.2	60.0	6.04	6.12	3.78	6289	0	0	0	...	0	0	0	0	0	0	0
3	0.42	61.6	56.0	4.82	4.80	2.96	1082	0	1	0	...	0	0	0	0	0	0	1
4	0.31	60.4	59.0	4.35	4.43	2.65	779	0	1	0	...	0	0	0	0	0	0	0

DATA SPLIT:

To make samples and test on it is the main thing in splitting the dataset

The ratio between Train and Test data is 70:30

(X Train - (18876, 23), Y Train - (18876, 1),X Test - (8091, 23),Y Test - (8091, 1))

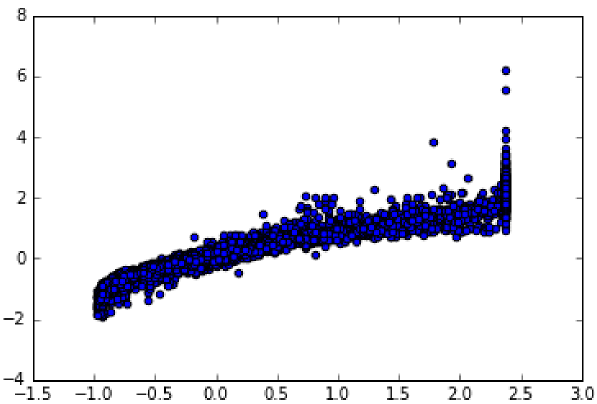
ANALYSIS SECTION:

LINEAR REGRESSION - PERFORMANCE METRICS:

- Coefficients - For the Linear equation formulating the coefficients.

```
The coefficient for carat is 0.9336627212882266
The coefficient for depth is 0.003021346021744358
The coefficient for table is -0.00829496184525669
The coefficient for x is 0.11934473649390785
The coefficient for y is 0.01202207724455986
The coefficient for z is 0.0004631368240287378
The coefficient for cut_Good is 0.041798452965412944
The coefficient for cut_Ideal is 0.1005523164608686
The coefficient for cut_Premium is 0.08137137884115554
The coefficient for cut_Very Good is 0.07780208702937445
The coefficient for color_E is -0.019706937329263312
The coefficient for color_F is -0.0268595154610537
The coefficient for color_G is -0.04872843563516146
The coefficient for color_H is -0.08529581454479944
The coefficient for color_I is -0.11309912030742296
The coefficient for color_J is -0.12518843128267673
The coefficient for clarity_IF is 0.22478956152654692
The coefficient for clarity_SI1 is 0.3544068925151113
The coefficient for clarity_SI2 is 0.21709642982462216
The coefficient for clarity_VS1 is 0.38345231618306336
The coefficient for clarity_VS2 is 0.40876865197025514
The coefficient for clarity_VVS1 is 0.3050502208331473
The coefficient for clarity_VVS2 is 0.3492340057451036
```

- The intercept for our model is -3.924745826117373e-17
- Model Score (Coefficient of determinant) - 0.9294
- Mean Squared Errors - 0.2655
- Linear Graph



RESULTS:

- Summary

OLS Regression Results

Dep. Variable:	price	R-squared:	0.929
Model:	OLS	Adj. R-squared:	0.929
Method:	Least Squares	F-statistic:	1.526e+04
Date:	Sun, 14 Jun 2020	Prob (F-statistic):	0.00
Time:	22:24:41	Log-Likelihood:	-2.2217e+05
No. Observations:	26933	AIC:	4.444e+05
Df Residuals:	26909	BIC:	4.446e+05
Df Model:	23		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-7155.6218	473.621	-15.108	0.000	-8083.945	-6227.299
carat	6760.3971	58.214	116.131	0.000	6646.295	6874.499
depth	7.2708	5.163	1.408	0.159	-2.848	17.390
table	-13.5055	3.388	-3.986	0.000	-20.146	-6.865
x	371.6309	34.022	10.923	0.000	304.946	438.316
y	47.4002	18.064	2.624	0.009	11.994	82.807
z	-0.7016	29.478	-0.024	0.981	-58.479	57.076
cut_Good	500.0484	39.268	12.734	0.000	423.081	577.016
cut_Ideal	709.1934	39.171	18.105	0.000	632.416	785.971
cut_Premium	648.4706	37.665	17.217	0.000	574.646	722.295
cut_Very_Good	639.1148	37.699	16.953	0.000	565.223	713.007
color_E	-194.4283	20.803	-9.346	0.000	-235.203	-153.654
color_F	-248.1206	21.080	-11.770	0.000	-289.439	-206.802
color_G	-406.7053	20.589	-19.754	0.000	-447.061	-366.350
color_H	-819.0134	21.956	-37.303	0.000	-862.048	-775.979
color_I	-1284.6431	24.464	-52.511	0.000	-1332.594	-1236.692
color_J	-1913.5410	30.056	-63.665	0.000	-1972.453	-1854.629
clarity_IF	4480.0325	59.364	75.468	0.000	4363.677	4596.388
clarity_SI1	2990.4743	50.761	58.912	0.000	2890.979	3089.969
clarity_SI2	2141.0847	51.008	41.975	0.000	2041.106	2241.064
clarity_VS1	3820.1041	51.804	73.742	0.000	3718.566	3921.642
clarity_VS2	3536.7631	51.010	69.334	0.000	3436.780	3636.746
clarity_VVS1	4272.6020	54.802	77.965	0.000	4165.188	4380.016
clarity_VVS2	4238.9003	53.323	79.494	0.000	4134.384	4343.417

Omnibus:	4895.704	Durbin-Watson:	2.002
Prob(Omnibus):	0.000	Jarque-Bera (JB):	80955.986
Skew:	0.395	Prob(JB):	0.00
Kurtosis:	11.457	Cond. No.	7.15e+03

- Linear Equation -
  Price = (-7155.62) * Intercept + (6760.4) * carat + (7.27) * depth + (-13.51) * table + (371.63) * x + (47.4) * y + (-0.7) * z + (500.05) * cut_Good + (709.19) * cut_Ideal + (648.47) * cut_Premium + (639.11) * cut_Very_Good + (-194.43) * color_E + (-248.12) * color_F + (-406.71) * color_G + (-819.01) *

color_H + (-1284.64) * color_I + (-1913.54) * color_J + (4480.03) * clarity_IF + (2990.47) * clarity_SI1 + (2141.08) * clarity_SI2 + (3820.1) * clarity_VS1 + (3536.76) * clarity_VS2 + (4272.6) * clarity_VVS1 + (4238.9) * clarity_VVS2

- Five best Feature Importance Variables -
  1. Clarity
  2. Carat
  3. Cut
  4. Length
  5. Width

BUSINESS INSIGHTS AND RECOMMENDATIONS :

1. Based on its Cut of Zirconia the company can keep the good price of the stone, the more the cut is accurate and ideal the more the price will be.
2. As an increase in the weight carat of the stone will also have a higher price.
3. The highest Profitable stone which will has cut as ‘Ideal’, clarity as ‘IF’, color as ‘E’ .
4. Lowest Profitable stones will have cut as ‘Good’, clarity as ‘SI2’, color as ‘J’.
5. Here length and width also plays a major role in determining the price of the stone. The Longer the length, the more valuable the stone will be to the customers.

Introduction of Problem Case 2:

To help the company in predicting whether an employee will opt for the package or not on the basis of the information given in the data set. Also, find out the important factors on the basis of which the company will focus on particular employees to sell their packages.

DATA SECTION:

Descriptive Statistics

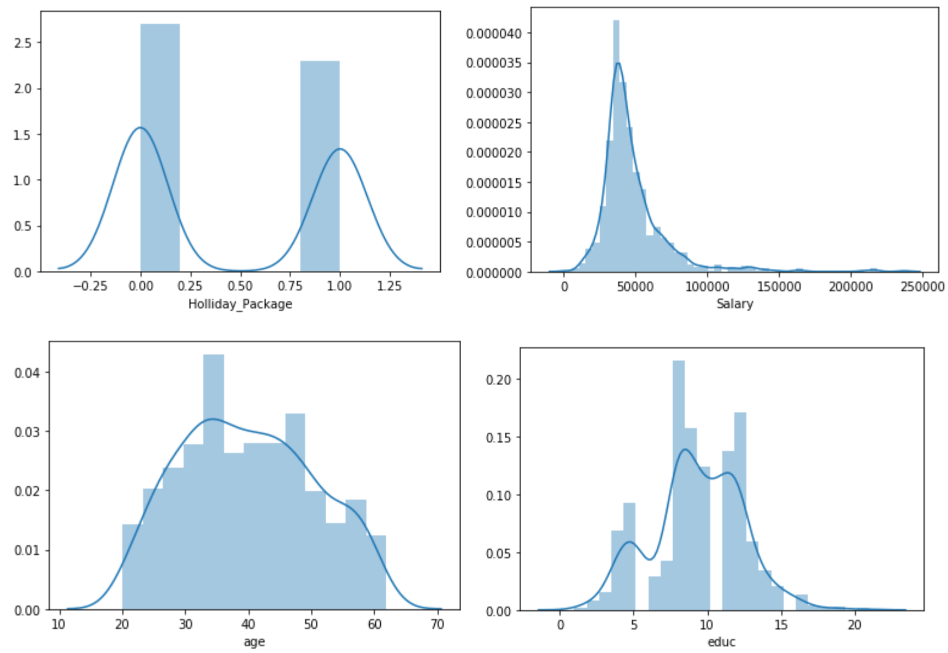
- The Dataset has 872 Rows and 8 Columns.
- Understanding quantile ranges of the dataset

	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

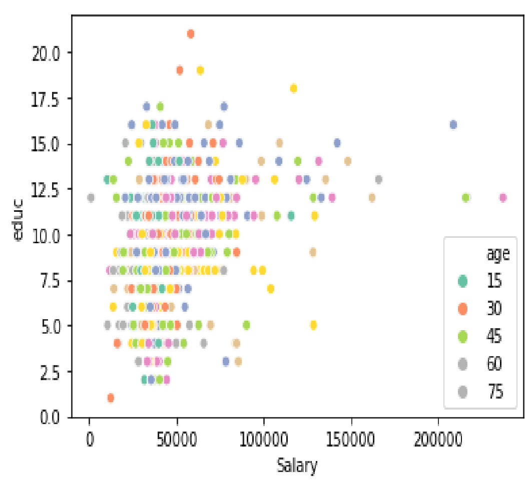
- Null Value Condition check - It clearly shows that there are no null values in the dataset.

```
Unnamed: 0              0
Holliday_Package        0
Salary                  0
age                     0
educ                    0
no_young_children       0
no_older_children       0
foreign                 0
dtype: int64
```

- **Univariate analysis** -
  - Show that equal distribution in Holliday_Package.
  - Most people are getting salaries around 10 Thousand-1 Lakh.
  - Age between people who are added to it is 30-50.
  - At max the people have 7+ years of education.



- **Multivariate Analysis**-Here explaining about the relationship between multi variables like salary, educ, age. These are correlated variables where salary is correlated with age, salary is correlated with educ.



- **Exploratory Data Analysis -** Defining the head values in the dataset with its data types and info of the dataset.

	Unnamed: 0	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
0	1	no	48412	30	8	1	1	no
1	2	yes	37207	45	8	0	1	no
2	3	no	58022	46	9	0	0	no
3	4	no	66503	31	11	2	0	no
4	5	no	66734	44	12	0	2	no

```
Unnamed: 0              int64
Holliday_Package       object
Salary                  int64
age                     int64
educ                    int64
no_young_children       int64
no_older_children       int64
foreign                object
dtype: object
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 872 entries, 0 to 871
Data columns (total 8 columns):
Unnamed: 0           872 non-null int64
Holliday_Package     872 non-null object
Salary               872 non-null int64
age                  872 non-null int64
educ                 872 non-null int64
no_young_children    872 non-null int64
no_older_children    872 non-null int64
foreign              872 non-null object
dtypes: int64(6), object(2)
memory usage: 47.8+ KB
```

METHOD SECTION:

ENCODING

- Changing the object variable into categorical.

```
feature: Holliday_Package
[no, yes]
Categories (2, object): [no, yes]
[0 1]

feature: foreign
[no, yes]
Categories (2, object): [no, yes]
[0 1]
```

- Check the distribution of the target variable here 0 represents 'NO' and 1 represents 'YES'.

```
0    471
1    401
Name: Holliday_Package, dtype: int64
```

DATA SPLIT

- Train Data and Test Data - we can see that the proportion of Ones and Zeroes in the training and test set is the same as the proportion of Ones and Zeroes that were present in the whole dataset..

```
0    0.534426
1    0.465574
Name: Holliday_Package, dtype: float64
```

```
0    0.553435
1    0.446565
Name: Holliday_Package, dtype: float64
```

- Here we are splitting the Data in the ratio of (70:30) with the random level of any value, scaling is not required for this dataset.

ANALYSIS SECTION:

LOGISTIC REGRESSION:

Confusion Matrix of Train Data:It clearly shows that it has 26 False Negatives.

```
array([[299,  34],
       [251,  26]], dtype=int64)
```

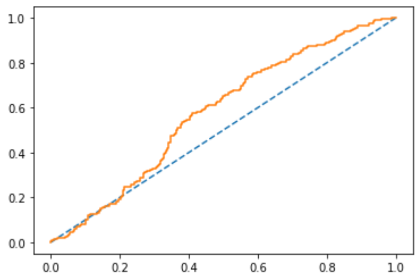
Confusion Matrix of Test Data: In test data it has 13 False Negatives.

```
array([[122,  16],
       [111,  13]], dtype=int64)
```

Accuracy - Training Data - 0.5327

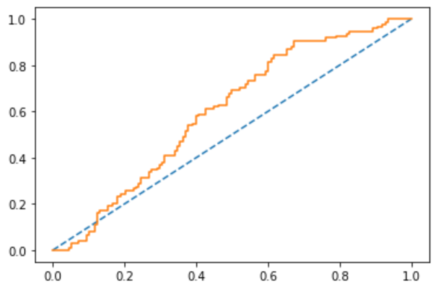
Accuracy - Test Data - 0.5152

ROC Curve- Training Data, ROC Score - AUC: 0.578



	precision	recall	f1-score	support
0	0.54	0.90	0.68	333
1	0.43	0.09	0.15	277
accuracy			0.53	610
macro avg	0.49	0.50	0.42	610
weighted avg	0.49	0.53	0.44	610

ROC Curve - Testing Data, ROC Score - AUC: 0.578



	precision	recall	f1-score	support
0	0.52	0.88	0.66	138
1	0.45	0.10	0.17	124
accuracy			0.52	262
macro avg	0.49	0.49	0.41	262
weighted avg	0.49	0.52	0.43	262

LINEAR DISCRIMINANT ANALYSIS:

Confusion Matrix of Train Data:It clearly shows that it has 158 False Negatives.

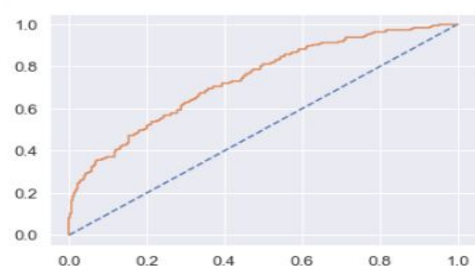
```
array([[254,  72],
       [126, 158]], dtype=int64)
```

Confusion Matrix of Test Data: In test data it has 65 False Negatives.

```
array([[103,  42],
       [ 52,  65]], dtype=int64)
```

Accuracy - Training Data - 0.6655
Accuracy - Test Data - 0.6564
ROC Curve- Training Data, ROC Score - AUC: 0.739

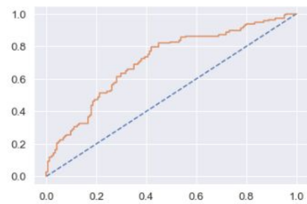
```
AUC: 0.738
[<matplotlib.lines.Line2D at 0xf46cc90>]
```



	precision	recall	f1-score	support
0	0.67	0.78	0.72	326
1	0.69	0.56	0.61	284
accuracy			0.68	610
macro avg	0.68	0.67	0.67	610
weighted avg	0.68	0.68	0.67	610

ROC Curve - Testing Data, ROC Score - AUC: 0.738

```
AUC: 0.738
[<matplotlib.lines.Line2D at 0xf4a6670>]
```



	precision	recall	f1-score	support
0	0.68	0.73	0.70	145
1	0.63	0.57	0.60	117
accuracy			0.66	262
macro avg	0.66	0.65	0.65	262
weighted avg	0.66	0.66	0.66	262

RESULTS:
Comparison between Logistic Regression and Linear Discriminant Analysis`

Logistic Regression
Accuracy - Training Data - 0.5327 Accuracy - Test Data - 0.5152
LDA - Accuracy - Training Data - 0.6655
Accuracy - Test Data - 0.6564

Comparing the ROC SCore of Test Data
Logistics Regression - 0.578
LDA - 0.738

As we have understood the more the area the curve has the better the model will be. Here '*LINEAR DISCRIMINANT ANALYSIS*' have ROC score of 0.738 will have more area in the curve and thus suits best model for predicting the Holiday Packages.

IMPORTANT FACTORS : For Eligibility

1. Salary.
2. Age.
3. No of young children.
4. No of older children.

BUSINESS INSIGHTS AND RECOMMENDATIONS:

1. The most important and first thing the company needs to look at its employee's salary and decide whether they will opt for a holiday package.
2. The less the age the more the employee will opt the holiday package.
3. Here Education or experience level will not sound more.
4. The number of young children will also come into account in choosing the holiday package as employees will decide to entertain with their kids.