**PREDICTIVE TRAVEL PACKAGE PREDICTION CASE STUDY REPORT**

**SUBMITTED BY: RAGHAV GUPTA**

**DATE: 25-04-2021**

Table of Contents

[Problem : Predicting Travel Package Prediction Case Study 5](#_Toc72762464)

[Problem Statement: 5](#_Toc72762465)

[Data Description: 5](#_Toc72762466)

[Domain: 5](#_Toc72762467)

[Context: 5](#_Toc72762468)

[Attribute Information: 5](#_Toc72762469)

[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. 6](#_Toc72762470)

[Basic EDA summary:- 6](#_Toc72762471)

[Univariate Analysis 6](#_Toc72762472)

[Checking IQR, Coeffiecient of Variation, IQR, lower range and upper range of numerical cols with summary 6](#_Toc72762473)

[Histogram 7](#_Toc72762474)

[Box Plots 8](#_Toc72762475)

[Correcting Spelling error in Column names 8](#_Toc72762476)

[Checking the unique values for categorical variables 8](#_Toc72762477)

[Bivariate and Multivariate analysis 9](#_Toc72762478)

[Swarm Plots 9](#_Toc72762479)

[Count Plots 10](#_Toc72762480)

[Correlation matrix 10](#_Toc72762481)

[Heat Map 11](#_Toc72762482)

[Pairplots 11](#_Toc72762483)

[VIF Checking for Multicollinearity 12](#_Toc72762484)

[Outlier treatment (flooring and capping) 13](#_Toc72762485)

[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) 13](#_Toc72762486)

[Scaling 13](#_Toc72762487)

[Encoding the data 13](#_Toc72762488)

[Proportion of 1s and 0s in Target Variable 14](#_Toc72762489)

[Splitting data into training and test set 14](#_Toc72762490)

[Checking the dimensions of the training and test data 14](#_Toc72762491)

[Building Logistic Regression Model 15](#_Toc72762492)

[Model Evaluation 15](#_Toc72762493)

[AUC and ROC for the training data & test data 15](#_Toc72762494)

[Confusion Matrix for the training data and testing data 15](#_Toc72762495)

[Training Data and Test Data Classification Report Comparison 16](#_Toc72762496)

[Applying GridSearchCV for Logistic Regression 16](#_Toc72762497)

[Model Evaluation 16](#_Toc72762498)

[AUC and ROC for the training data & test data 16](#_Toc72762499)

[Confusion Matrix for the training data and testing data 17](#_Toc72762500)

[Training Data and Test Data Classification Report Comparison 17](#_Toc72762501)

[Getting the equation 18](#_Toc72762502)

[Logistics Regression Conclusion 19](#_Toc72762503)

[LDA Model 21](#_Toc72762504)

[Model Evaluation 21](#_Toc72762505)

[Training Data and Test Data Confusion Matrix Comparison 21](#_Toc72762506)

[Training Data and Test Data Classification Report Comparison 21](#_Toc72762507)

[AUC and ROC for the training data & test data 22](#_Toc72762508)

[LDA Conclusion 22](#_Toc72762509)

[AUC and ROC for the training data & test data 23](#_Toc72762510)

[Running other classification models 25](#_Toc72762511)

[ROC for Train Dataset for all the models :- 25](#_Toc72762512)

[ROC for Test Dataset for all the models :- 26](#_Toc72762513)

[Inference: Basis on these predictions, what are the insights and recommendations. 28](#_Toc72762514)

## Problem : Predicting Travel Package Prediction Case Study

## ****Problem Statement:****

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

## Data Description:

Holiday\_Package.csv is a dataset that contains the details of 872 employees of company

## Domain:

Tour and Travel

## Context:

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.

## Attribute Information:

* 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

## 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.

### Basic EDA summary:-

* Data contains 872 rows and 7 columns.
* Except Holiday\_Package and Foreign column, all other columns are numeric.
* Holiday\_Package and Foreign column are binary/boolean in nature (yes/no)
* There is no duplicate data present in the dataset
* There is no null/missing data in any of these columns
* There are few outliers in the dataset (details given in Univariate Analysis below)

### Univariate Analysis

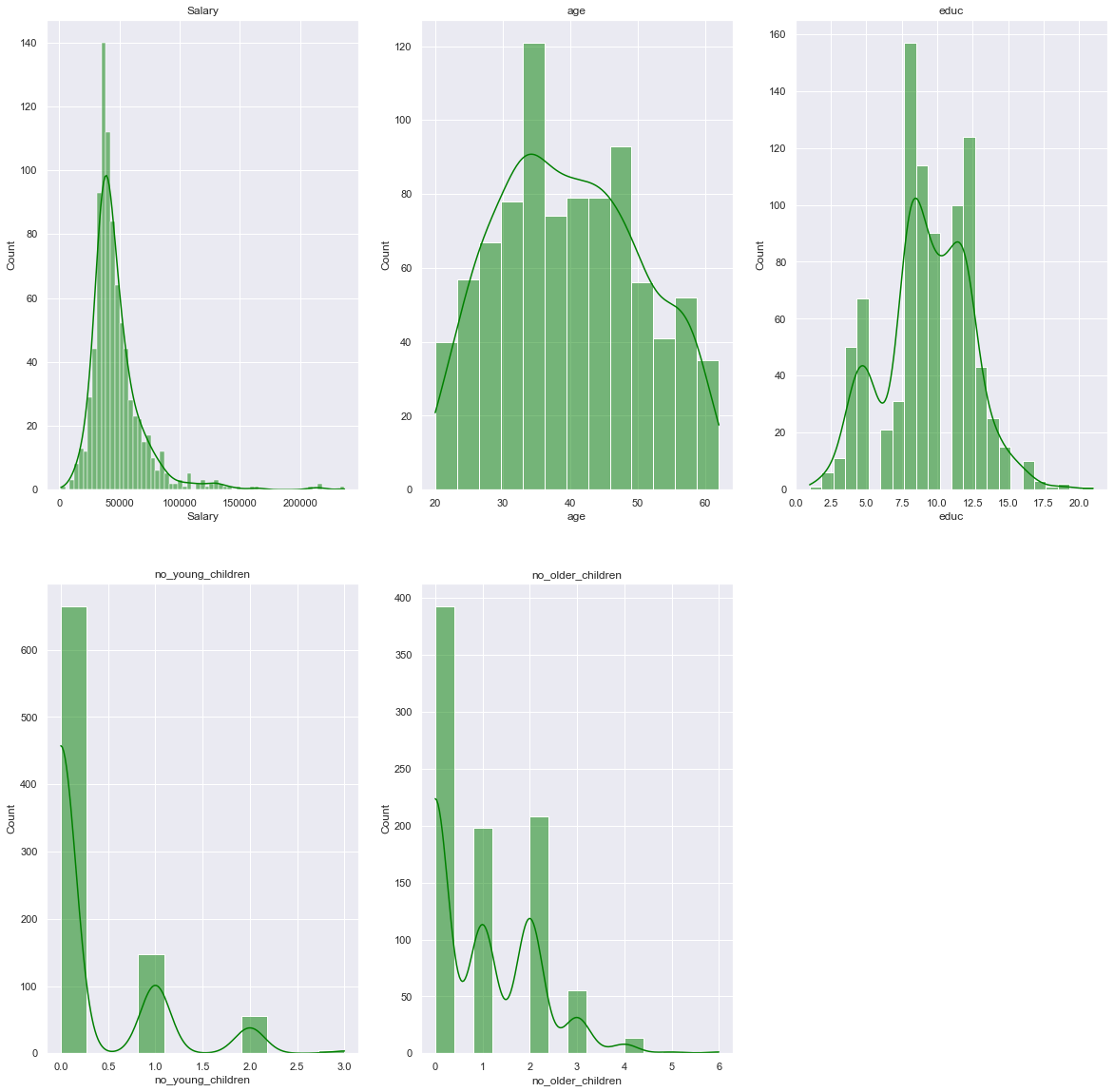
### Checking IQR, Coeffiecient of Variation, IQR, lower range and upper range of numerical cols with summary

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Salary | **age** | **educ** | **no\_young\_children** | **no\_older\_children** |
| count | **872** | 872 | 872 | 872 | 872 |
| mean | **47729.1720** | 39.9553 | 9.3073 | 0.3119 | 0.9828 |
| std | **23418.6685** | 10.5517 | 3.0363 | 0.6129 | 1.0868 |
| min | **1322.0** | 20.0 | 1.0 | 0.0 | 0.0 |
| 25% | **35324.0** | 32.0 | 8.0 | 0.0 | 0.0 |
| 50% | **41903.5** | 39.0 | 9.0 | 0.0 | 1.0 |
| 75% | **53469.5** | 48.0 | 12.0 | 0.0 | 2.0 |
| max | 236961 | 62.0 | 21.0 | 3.0 | 6.0 |
| CV | 0.49 | 0.26 | 0.33 | 1.96 | 1.11 |
| Skew | 3.10 | 0.15 | -0.05 | 1.95 | 0.95 |
| IQR | 18145.5 | 16.0 | 4.0 | 0.0 | 2.0 |
| UR | 80687.75 | 72.00 | 18.00 | 0.00 | 5.00 |
| LR | 8105.75 | 8 | 2 | 0 | -3 |

From summary, we can see that :-

* max salary(236K) is very high as compared to mean(47K) and median(42K). Hence it contains outlier
* Mean and median of age are approximately similar 39-40. It doesnt contains outlier.
* Education middle 50% of data lies in between 8 to 12 range with few outliers.
* Most employees have no of young children as 0.
* Most of the employees have 1 child who is older than 7 years
* All the columns are positively skewed except education

### Histogram



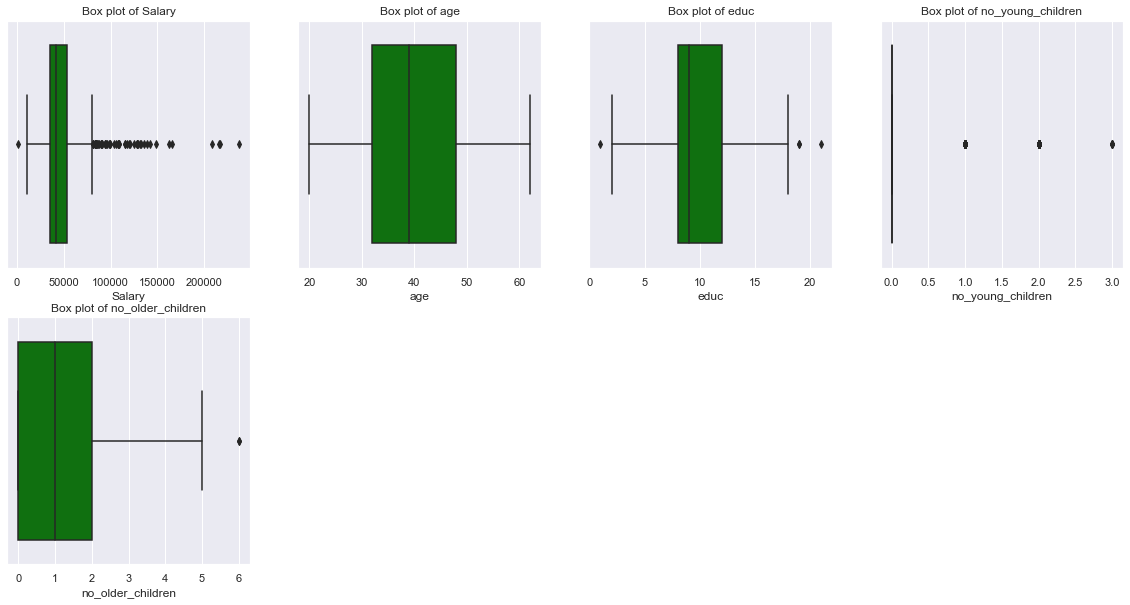
***From histograms we can see that***

* Salary range is 0-100000 for most of the employees. However few employees are getting more salary causing skewness
* Age appears to be normally distributed
* Around 650 employees out of 872 have their young children as 0.
* Around 380 out of 872 employees have no of older children as 0
* Education middle 50% of data lies in between 8 to 12 range with few outliers.

### Box Plots

As evident from above box plot, there are many outliers in salary column.

education, no of young childern and old children columns have very few outliers which we can ignore



### Correcting Spelling error in Column names

There were few typo errors in the column names. These were all corrected.

Column names before renaming:

['Holliday\_Package', 'Salary', 'age', 'educ', 'no\_young\_children' 'no\_older\_children', 'foreign']

Column names after renaming:

['HolidayPackage', 'Salary', 'Age', 'Educ', 'No\_young\_children',

'No\_older\_children', 'Foreign']

### Checking the unique values for categorical variables

HolidayPackage : 2

yes 401

no 471

Name: HolidayPackage, dtype: int64

Foreign : 2

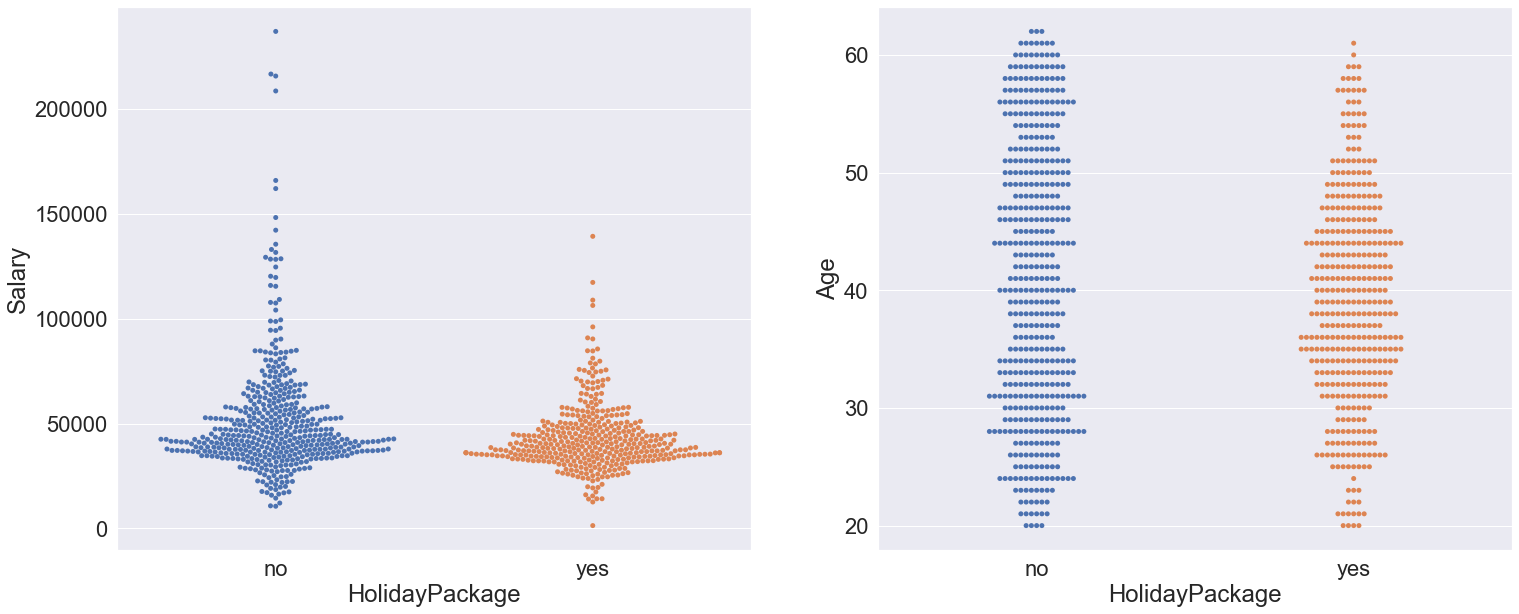
yes 216

no 656

Name: Foreign, dtype: int64

### Bivariate and Multivariate analysis

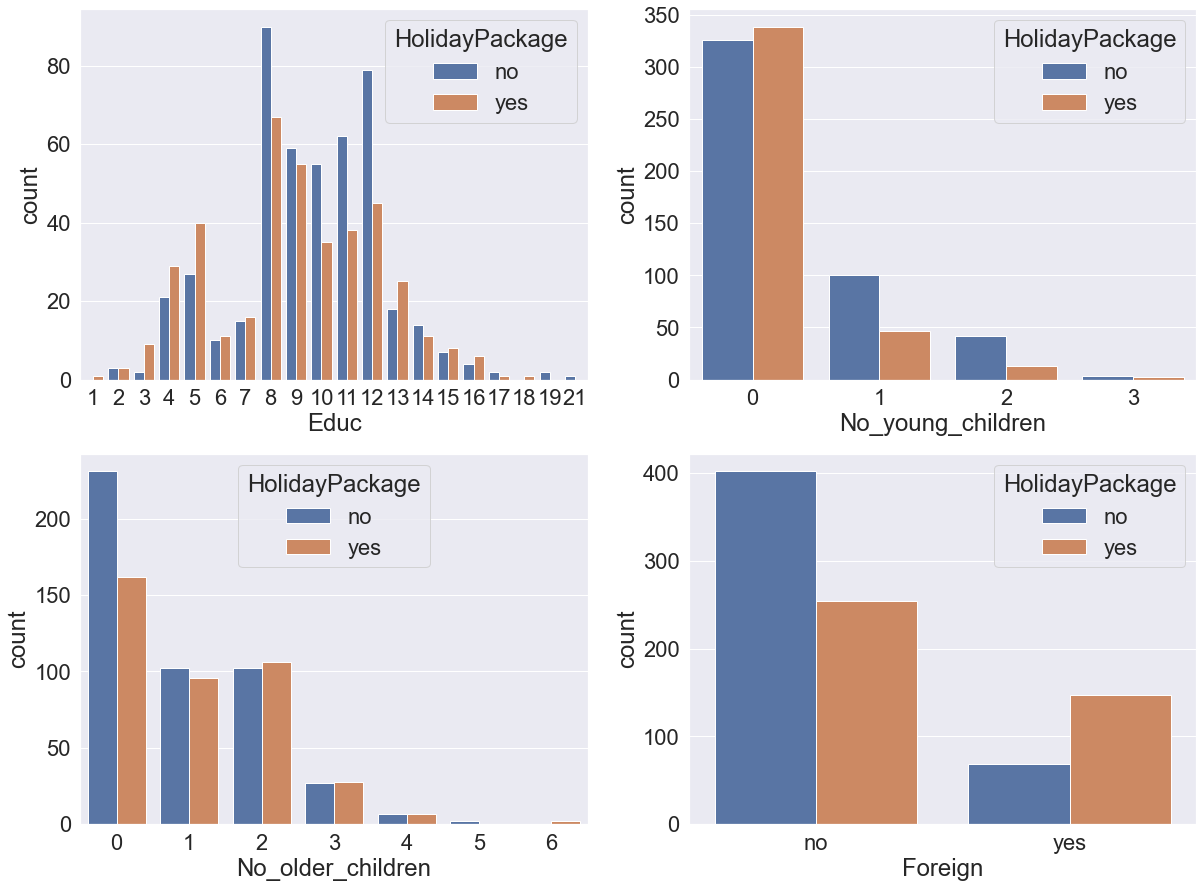
### Swarm Plots



We can see that :-

* As Salary increases to the max value, employees count increases for the not opting for the holiday package.
* As Age increases beyond 50 level, less employees opt for the holiday package

### Count Plots



We can see observe from above graph that:-

* More Employees opt for Tours if their education level is 3,4,5,6,7,13,14,15,16
* Employees don’t opt for tours if they have young child
* Older children count doesn’t appears to have much impact on tour opted by employees or not
* Foreigner employees tends to opt more for the tour

### Correlation matrix

Salary Age Educ No\_young\_children No\_older\_children

Salary 1.00 0.07 0.33 -0.03 0.11

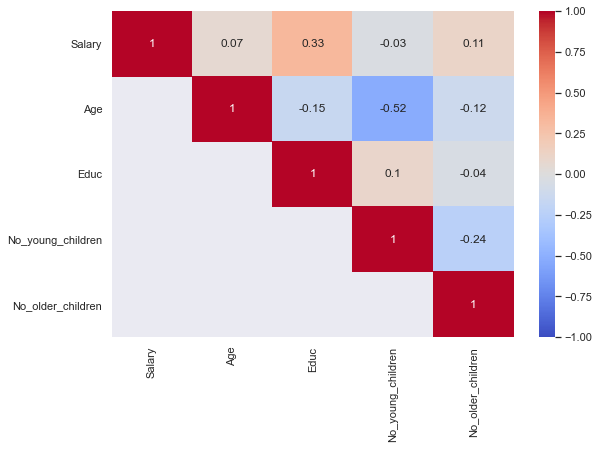
Age 0.07 1.00 -0.15 -0.52 -0.12

Educ 0.33 -0.15 1.00 0.10 -0.04

No\_young\_children -0.03 -0.52 0.10 1.00 -0.24

No\_older\_children 0.11 -0.12 -0.04 -0.24 1.00

### Heat Map

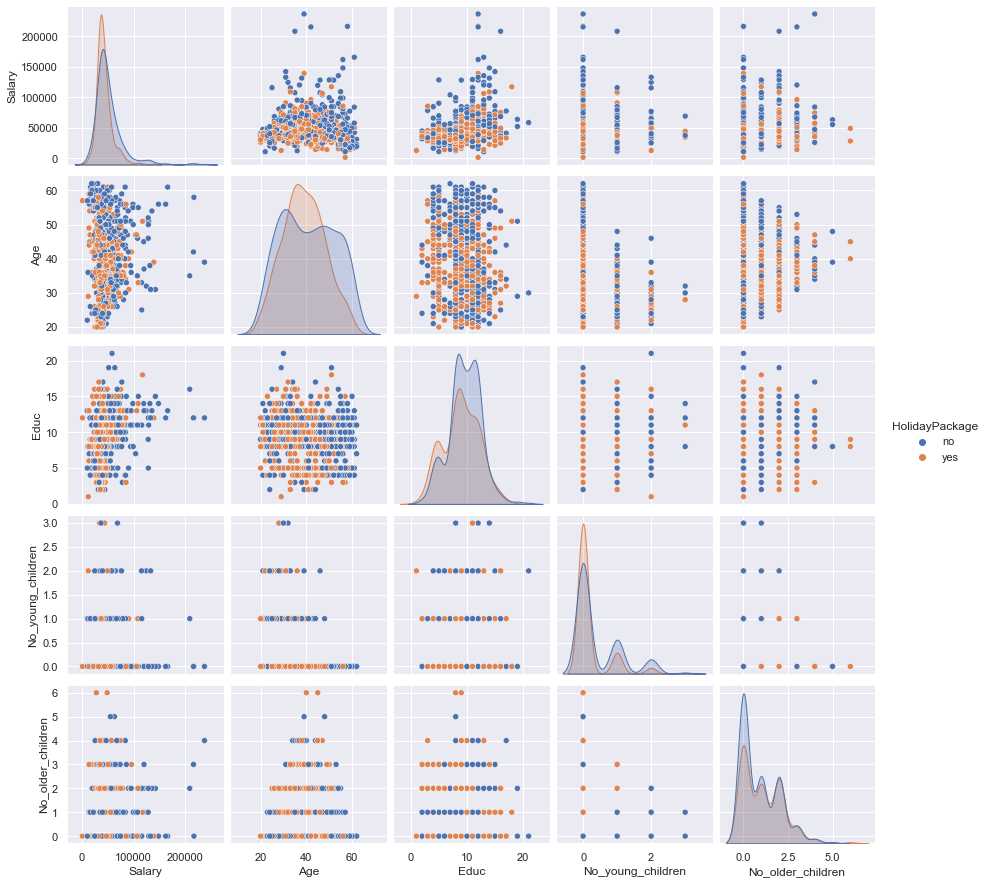
****

We can see in heatmap & correlation matrix that

* Salary has correlation with educ.
* Age is negatively correlated with No\_young\_children

### Pairplots

To check the correlation of 2 columns in more detail we can draw **pairplots/Scatter Diagram**



As depicted in heat map of correlation matrix, we can see that no of young children negatively correlated with age.

Salary is slightly correlated with Educ

### VIF Checking for Multicollinearity

VIF output for different column attributes are as shown below :-

Variables VIF

0 Salary 6.027872

1 Age 6.832751

2 Educ 8.890845

3 No\_young\_children 1.403995

4 No\_older\_children 1.817912

We can see that VIF is greator than 5 for salary, age and education. However its value is less than 10. So dataset has some multicollinearity but not very high multicollinearity

### Outlier treatment (flooring and capping)

We are only doing outlier treatment for Salary attribute as other columns have very less outliers and that are near lower and upper ranges

col: Salary ,lower range : 8105.75 ,upper range: 80687.75 ,No of outliers: 57 ,outlier %: 6.54

Salary column contains 6.54% of its total rows as outliers.

After outlier treatment, Box plot of salary column now shows no outliers present



## 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)

### Scaling

Scaling is not required for the logistic regression model. Hence its not performed here

### Encoding the data

Since only two categorical inputs are there in the dataset, we have converted them to 0 and 1.

‘yes’ value changed to 1

‘no’ value change to 0.

HolidayPackage value count after label encoding :-

0 471

1 401

Foreign variable value count after dummy encoding :-

Foreign

0 656

1 216

### Proportion of 1s and 0s in Target Variable

No 54.0

Yes 46.0

This target variable is balanced

### Splitting data into training and test set

After extracting the Target column, data set is split into 70:30 ratio i.e. 70% of input observations into train data set for building the model and 30% observations into test data set for testing and validating the model.

### Checking the dimensions of the training and test data

After splitting, dimensions of training and test data sets are as shown below :-

X\_train (610, 6) 🡪 input train

X\_test (262, 6) 🡪 input test

y\_train (610,1)🡪 actual output train

y\_test (262,1) 🡪 actual output test

Total Observations (872, 7)

We can see that data has been splitted into train(70%) and test(30%) successfully

### Building Logistic Regression Model

Initially we built the logistic regression model using sklearn as per the following hyper parameters :-

* solver='newton-cg',
* max\_iter=10000,
* penalty='none',
* verbose=True,
* n\_jobs=2

### Model Evaluation

Accuracy score for train and test data set is as shown below :-

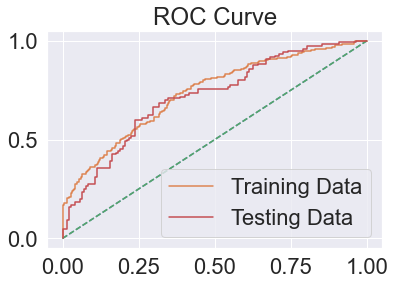
Model score for training dataset 0.6672131147540984

Model score for training dataset 0.648854961832061

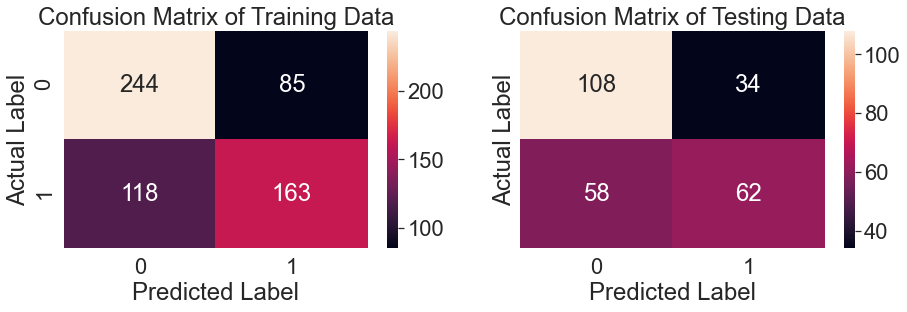
### AUC and ROC for the training data & test data

AUC for Train dataset: 0.733

AUC for test dataset: 0.733



### Confusion Matrix for the training data and testing data



### Training Data and Test Data Classification Report Comparison

Classification Report of the training data:

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

Classification Report of the test data:

precision recall f1-score support

0 0.65 0.76 0.70 142

1 0.65 0.52 0.57 120

accuracy 0.65 262

macro avg 0.65 0.64 0.64 262

weighted avg 0.65 0.65 0.64 262

### Applying GridSearchCV for Logistic Regression

Showing best parameters for the grid search

{'penalty': 'l1', 'solver': 'liblinear', 'tol': 0.0001}

### Model Evaluation

Model score for training dataset 0.6639344262295082

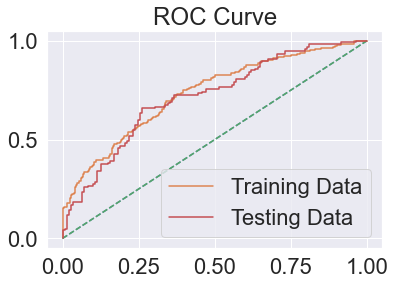
Model score for training dataset 0.6564885496183206

### AUC and ROC for the training data & test data

AUC for Train dataset: 0.733

AUC for test dataset: 0.733

In [747]:



### Confusion Matrix for the training data and testing data

#### 

### Training Data and Test Data Classification Report Comparison

Classification Report of the training data:

precision recall f1-score support

0 0.67 0.74 0.71 329

1 0.66 0.57 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

Classification Report of the test data:

precision recall f1-score support

0 0.65 0.77 0.71 142

1 0.66 0.52 0.58 120

accuracy 0.66 262

macro avg 0.66 0.65 0.64 262

weighted avg 0.66 0.66 0.65 262

### Getting the equation

Using statsmodel, we can find the equation of log odds and we can find which coefficient has the more weightage in deciding the target response variable

|  |  |  |  |
| --- | --- | --- | --- |
| Logit Regression Results | | | |
| **Dep. Variable:** | HolidayPackage | **No. Observations:** | 872 |
| **Model:** | Logit | **Df Residuals:** | 865 |
| **Method:** | MLE | **Df Model:** | 6 |
| **Date:** | Sat, 24 Apr 2021 | **Pseudo R-squ.:** | 0.1244 |
| **Time:** | 19:02:34 | **Log-Likelihood:** | -526.78 |
| **converged:** | True | **LL-Null:** | -601.61 |
| **Covariance Type:** | nonrobust | **LLR p-value:** | 9.138e-30 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **z** | **P>|z|** | **[0.025** | **0.975]** |
| **Intercept** | 2.5432 | 0.559 | 4.550 | 0.000 | 1.448 | 3.639 |
| **Salary** | -2.088e-05 | 5.26e-06 | -3.970 | 0.000 | -3.12e-05 | -1.06e-05 |
| **Age** | -0.0496 | 0.009 | -5.491 | 0.000 | -0.067 | -0.032 |
| **Educ** | 0.0342 | 0.029 | 1.172 | 0.241 | -0.023 | 0.091 |
| **No\_young\_children** | -1.3287 | 0.180 | -7.386 | 0.000 | -1.681 | -0.976 |
| **No\_older\_children** | -0.0251 | 0.074 | -0.341 | 0.733 | -0.169 | 0.119 |
| **Foreign** | 1.3037 | 0.200 | 6.519 | 0.000 | 0.912 | 1.696 |

We can see that the p value of No\_older\_children is the highest (.733) , followed by Educ(.241) and it is greator than 0.05.

Hence it confirms that No\_older\_children and Educ attribute has no impact on dependent variable HolidayPackage

Removing these 2 columns, we ran the model again to get the following report :-

|  |  |  |  |
| --- | --- | --- | --- |
| Logit Regression Results | | | |
| **Dep. Variable:** | HolidayPackage | **No. Observations:** | 872 |
| **Model:** | Logit | **Df Residuals:** | 867 |
| **Method:** | MLE | **Df Model:** | 4 |
| **Date:** | Sat, 24 Apr 2021 | **Pseudo R-squ.:** | 0.1231 |
| **Time:** | 18:57:14 | **Log-Likelihood:** | -527.58 |
| **converged:** | True | **LL-Null:** | -601.61 |
| **Covariance Type:** | nonrobust | **LLR p-value:** | 5.267e-31 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **z** | **P>|z|** | **[0.025** | **0.975]** |
| **Intercept** | 2.8128 | 0.448 | 6.282 | 0.000 | 1.935 | 3.690 |
| **Salary** | -1.932e-05 | 4.94e-06 | -3.911 | 0.000 | -2.9e-05 | -9.64e-06 |
| **Age** | -0.0504 | 0.008 | -5.962 | 0.000 | -0.067 | -0.034 |
| **No\_young\_children** | -1.3023 | 0.169 | -7.707 | 0.000 | -1.633 | -0.971 |
| **Foreign** | 1.2092 | 0.183 | 6.592 | 0.000 | 0.850 | 1.569 |

Now all p values are less than 0.05. Hence all these attributes and their coeffficients have importance in deciding the target variable HolidayPackage.

**Also we can see that coef value is highest for No\_young\_children followed by foreign, Age and salary**

**Salary coefficient value is very low i.e -00001932. So its impact is almost 0 on dependent variable**

### Logistics Regression Conclusion

**Train Data:**

AUC: 73%

Accuracy: 66%

Precision: 66%

f1-Score: 61%

Recall: 57%

**Test Data:**

AUC: 73%

Accuracy: 66%

Precision: 68%

f1-Score: 58%

Recall: 52%

**Train and Test dataset have similar statistics, hence model is giving similar result for test and train data set..**

With accuracy of 66% and recall rate of 52%, model is only able to predict 52% of total tours which were actually claimed as claimed.

Precision is 68% of test data which means, out of total employees predicted by model as opt for tour, 68% employees actually opted for the tour

F1-score is the harmonic mean of precision and recall, it takes into the effect of both the scores and this value is low if any of these 2 value is low.

Since we are building a model to predict if whether employee will opt for tour or not, for practical purposes, we will be more interested in correctly classifying 1 (employees opting for tour) than 0(employees not opting for tour).

If an employee not opting for tour is incorrectly predicted to be "opted for tour" by the model, then the impact on cost for the travel company would be bare minimum. But if an employee opted for tour is incorrectly predicted to be not opted by the model, then the cost impact would be very high for the tour and travel company. It’s a loss of potential lead for the company. Hence recall rate (actual data point identified as True by model) is very important in this scenario.

#### As Recall rate of test dataset is very poor around 52% thus this doesnt looks good enough for classification

Logistic regression equation is as shown below :-

Log (odd) = (2.81) + (-0.0) \* Salary + (-0.05) \* Age + (-1.3) \* No\_young\_children + (1.21) \* Foreign

We can see that salary coefficient is very small , this it can be removed. So our equation would become :-

**Log (odd) = (2.81) + (-0.05) \* Age + (-1.3) \* No\_young\_children + (1.21) \* Foreign**

Most important attribute here is No of young children followed by Foreign and age

### LDA Model

We have built the model using sklearn LinearDiscriminantAnalysis method

### Model Evaluation

### Training Data and Test Data Confusion Matrix Comparison

#### 

### Training Data and Test Data Classification Report Comparison

Classification Report of the training data:

precision recall f1-score support

0 0.67 0.74 0.70 329

1 0.65 0.57 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

Classification Report of the test data:

precision recall f1-score support

0 0.65 0.76 0.70 142

1 0.65 0.52 0.57 120

accuracy 0.65 262

macro avg 0.65 0.64 0.64 262

weighted avg 0.65 0.65 0.64 262

### AUC and ROC for the training data & test data

AUC for Train dataset: 0.731

AUC for test dataset: 0.714

#### 

### LDA Conclusion

**Train Data:**

AUC: 73%

Accuracy: 66%

Precision: 65%

f1-Score: 61%

Recall: 57%

**Test Data:**

AUC: 71%

Accuracy: 65%

Precision: 65%

f1-Score: 57%

Recall: 52%

**Train and Test dataset have similar statistics, hence model is giving similar result for test and train data set..**

With accuracy of 65% and recall rate of 52%, model is only able to predict 52% of total tours which were actually claimed as claimed.

Precision is 65% of test data which means, out of total employees predicted by model as opt for tour, 65% employees actually opted for the tour

F1-score is the harmonic mean of precision and recall, it takes into the effect of both the scores and this value is low if any of these 2 value is low.

Since we are building a model to predict if whether employee will opt for tour or not, for practical purposes, we will be more interested in correctly classifying 1 (employees opting for tour) than 0(employees not opting for tour).

If a employee not opting for tour is incorrectly predicted to be "opted for tour" by the model, then the impact on cost for the travel company would be bare minimum. But if am employee opted for tour is incorrectly predicted to be not opted by the model, then the cost impact would be very high for the tour and travel company. Its a loss of potential lead for the company. Hence recall rate (actual data point identified as True by model) is very important in this scenario.

#### As Recall rate of test dataset is very poor around 52% thus this doesn’t looks good enough for classification

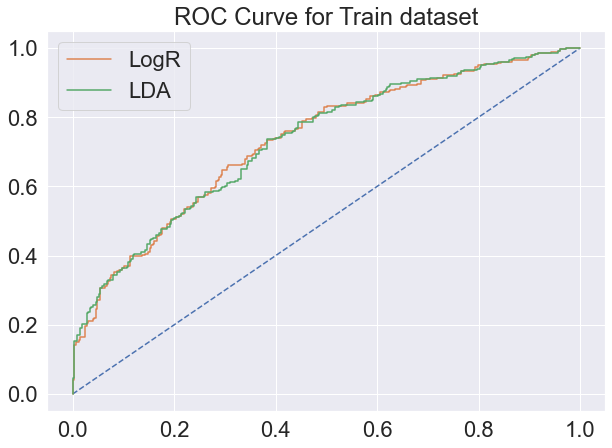
**Comparing Logistic Regression vs LDA**

|  | **Logistic Regression** | **LDA** |
| --- | --- | --- |
| **Train Accuracy** | 0.67 | 0.66 |
| **Test Accuracy** | 0.66 | 0.65 |
| **Train AUC** | 0.73 | 0.73 |
| **Test AUC** | 0.72 | 0.71 |
| **Train Recall** | 0.57 | 0.57 |
| **Test Recall** | 0.52 | 0.52 |
| **Train precision** | 0.66 | 0.65 |
| **Test precision** | 0.67 | 0.65 |
| **Train f1** | 0.61 | 0.61 |
| **Test f1** | 0.58 | 0.57 |

### AUC and ROC for the training data & test data

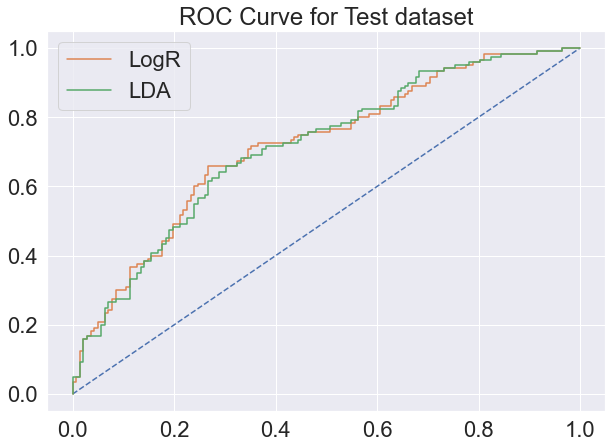
AUC for LogR is: 0.73

AUC for LDA is: 0.73

****

AUC for LogR is: 0.72

AUC for LDA is: 0.71



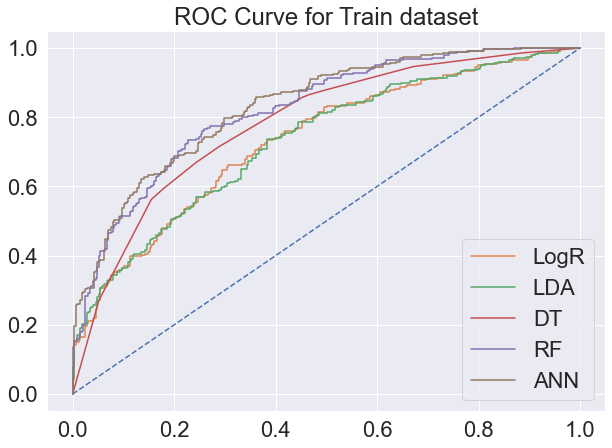
On comparison of both these models, it looks like that no model is over-fitting/under fitting.

Both these models have similar evaluation score but Logistic Regression model has accuracy, f1 score and AUC slightly better than LDA. **So we will go for Logistic Regression model in this case.**

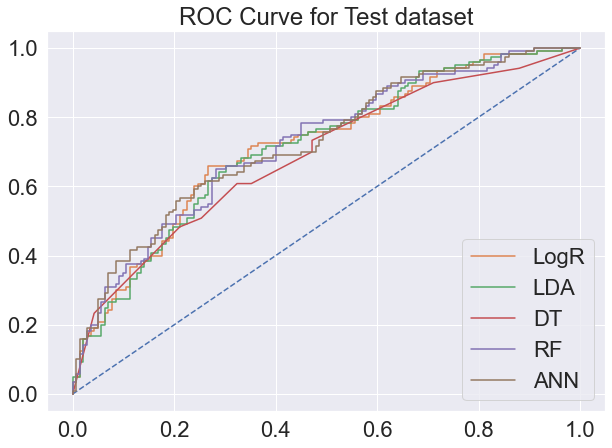
### Running other classification models

We have run the decision tree classification, Random Forest classification, Artificial Nuero network MLP classification for predicting the target variable response.

### ROC for Train Dataset for all the models :-

****

### ROC for Test Dataset for all the models :-

****

Summary of performance metrics of all the models is as shown below :-

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | LR | LDA | DT | RF | ANN |
| Train Accuracy | 0.67 | 0.66 | 0.72 | 0.74 | 0.74 |
| Test Accuracy | 0.66 | 0.65 | 0.64 | 0.67 | 0.68 |
| Train AUC | 0.73 | 0.73 | 0.78 | 0.82 | 0.83 |
| Test AUC | 0.72 | 0.71 | 0.69 | 0.72 | 0.73 |
| Train Recall | 0.57 | 0.57 | 0.59 | 0.66 | 0.67 |
| Test Recall | 0.52 | 0.52 | 0.51 | 0.6 | 0.62 |
| Train precision | 0.66 | 0.65 | 0.74 | 0.75 | 0.75 |
| Test precision | 0.67 | 0.65 | 0.63 | 0.65 | 0.65 |
| Train f1 | 0.61 | 0.61 | 0.66 | 0.7 | 0.71 |
| Test f1 | 0.58 | 0.57 | 0.56 | 0.62 | 0.64 |

We are building a machine learning model to predict if an employee of a particular company will opt for the tour package or not, for practical purposes, we will be more interested in correctly classifying 1 (employees opting for tour) than 0(employees not opting for tour).

If an employee not opting for tour is incorrectly predicted to be "opted for tour" by the model, then the impact on cost for the travel company would be bare minimum. But if an employee opted for tour is incorrectly predicted to be not opted by the model, then the cost impact would be very high for the tour and travel company. It’s a loss of potential customer/sale for the company. Hence recall rate (actual data point identified as True by model) is very important in this scenario.

F1 score which is dependent on recall and precision is also an important factor in scenario.

So we are more interested in Recall and F1 score for comparing three models.

**In all the models, Train and Test dataset have shown similar statistics. But overall recall percentage of test data set is low (62% being the best) and low accuracy score of around (66%)**

Out of all the models, Logistic regression , LDA and CART are giving recall values as 51-52% for test dataset. However ANN and RF are giving better recall rate of 60-62%.

Similarly F1 score for ANN and RF is also high as compare to other three models.

**Hence ANN/RF performance is slightly better than the other 3 models with ANN giving the best results**

**So overall percentages being low, our model is not good enough for classification. More sample data or better attributes which are deciding the target variables are required for improving the accuracy scores.**

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

**Business Recommendations:-**

We have run five different models (Logistic Regression/ Linear Discriminant Analysis/ CART/ Random Forest/ Artificial Neuro Network) for predicting whether an employee is opting for holiday package or not.

Based on the reports and analysis done it was found that all models were not good enough for classification as accuracy coming out is 66%.

So our recommendation to the business is as shown below:-

* In order to further improve the predictive model results for finding the employees which will opt for tour in future more accurately, more data sample is required for analysis.
* Current model is useful to predict when tours are not getting claimed with more than 66% accuracy.
* Most important attribute here is No of young children of an employee followed by Foreigner column and lastly the age.
* As seen in EDA, 68% of foreign employees are opting for the tour packages. So the travel company should make dedicated tours for these foreigners keeping in mind which places/areas that these foreigners would like to travel. If these customers are satisfied then when they will travel back to their country they will refer more of their friends/family members for tours. This way company can retain and increase its customer base.
* Currently 24% of employees have 1 or more young child. It was found during EDA that out of these employees 70% are not opting for the tours. So the travel company should make dedicated tour for the employees who have young child and provide them with some extra child care benefits (like play area for child, child food, medical facilities etc.) so as to lure these employees.
* Old age employees (age greater than 50) opt less for the tours. So company can provide dedicated tour plans for old aged senior employees.
* As per the analysis, Salary of the employees is not an important attribute in deciding that whether employee will opt for tour or not. So the travel company should not focus on salary of the employee. May be salary can decide which tour (economical or lavish) that particular employee will be interested in but he/she will opt for some tour irrespective of his/her salary.