Problem 2 – Logistic Regression and LDA

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 Dictionary:

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.

The data is read, and the top five records of the dataset are observed to get an overview.

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

Exploratory Data Analysis

The column properties are observed

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 872 entries, 1 to 872
Data columns (total 7 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Holliday_Package    872 non-null    object
 1   Salary              872 non-null    int64
 2   age                 872 non-null    int64
 3   educ                872 non-null    int64
 4   no_young_children   872 non-null    int64
 5   no_older_children   872 non-null    int64
 6   foreign             872 non-null    object
dtypes: int64(5), object(2)
memory usage: 54.5+ KB
```

The descriptive statistics is also taken

	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

Inferences:

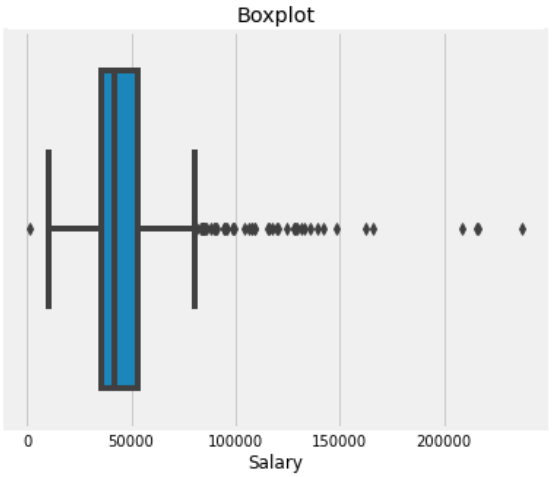
1. There are 872 rows (observations) and 7 columns (features).
2. There are no null values.
3. There are no duplicate rows.
4. The data types of the columns are as follows:
    - There are 5 numeric data types ('Salary', 'age', 'educ', 'no_young_children', 'no_older_children').
        - Here, 'age','educ', 'no_young_children', 'no_older_children' are numeric discrete variables and the 'Salary' is continuous.
    - There are 2 object data types – 'Holliday_Package', 'foreign'. These need to be encoded into categorical values later.
5. The object data types 'Holliday_Package' and 'foreign' each have values 'No' and 'Yes'.
6. The 75th percentile for Salary indicates that around 75% of the observations have salary around 53469.5, and 236961 is the maximum.
7. Most of the employees (75% of the data) seem to nearing their fifties, and 62 is the maximum age

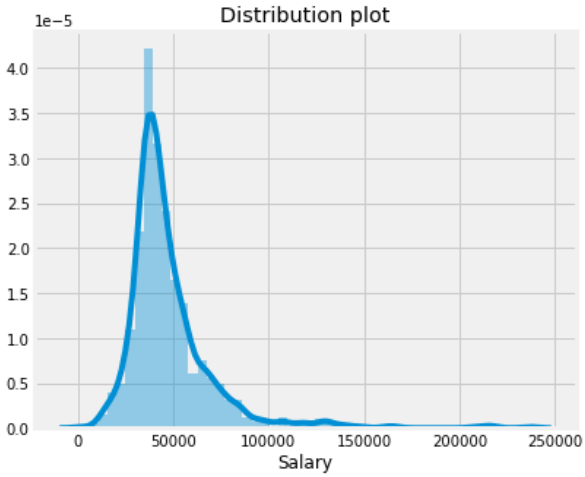
UNIVARIATE ANALYSIS FOR NUMERIC VARIABLES

SALARY

- The boxplot shows that there some outliers.
- The distribution plot shows that the data is right-skewed.

SALARY

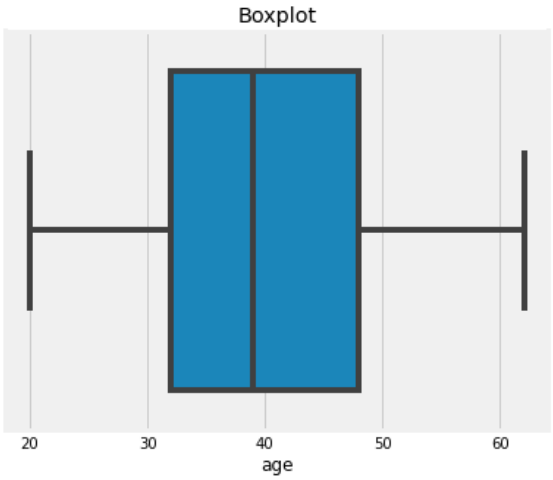


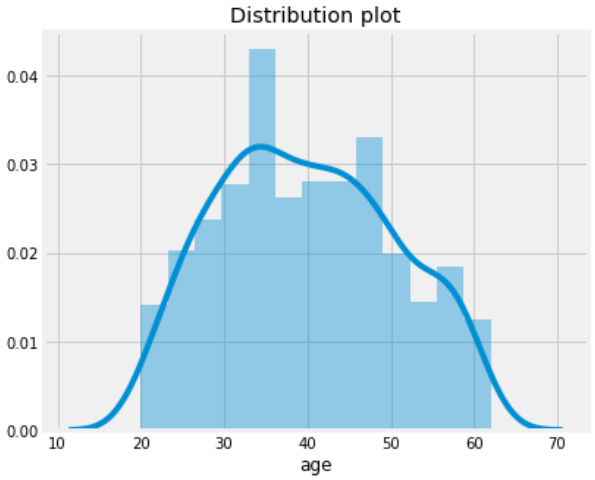


AGE

- The boxplot shows that there are no outliers.
- The distribution plot shows that the data is almost normal.

AGE

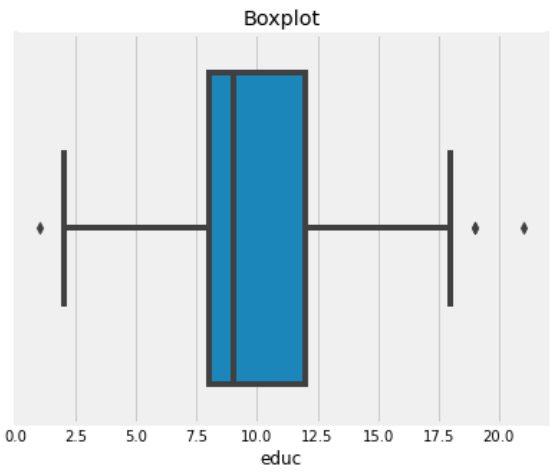


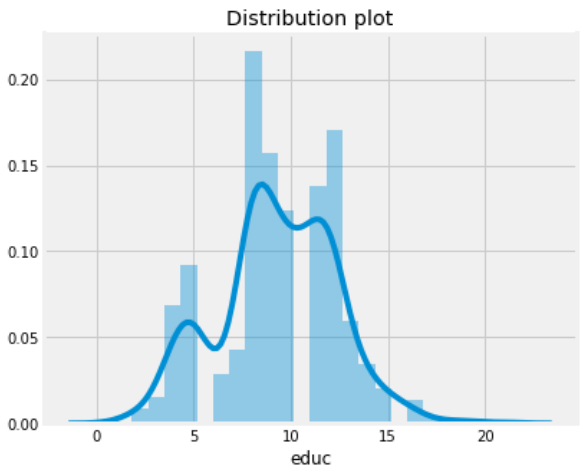


EDUCATION

- The boxplot shows that there are very few outliers.
- The distribution plot shows that the data is not exactly normal.

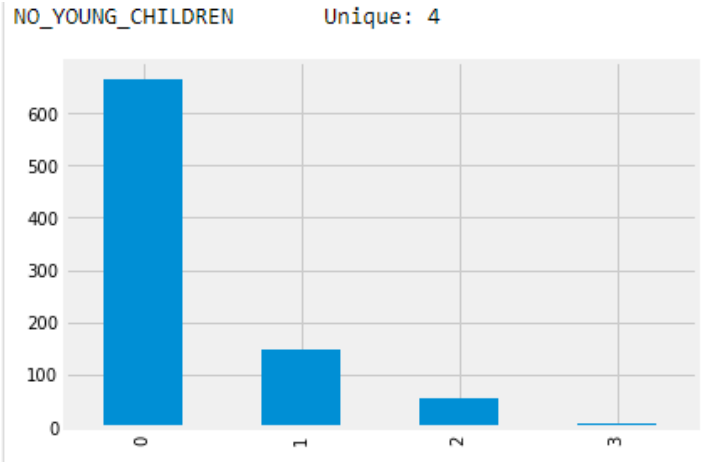
EDUC





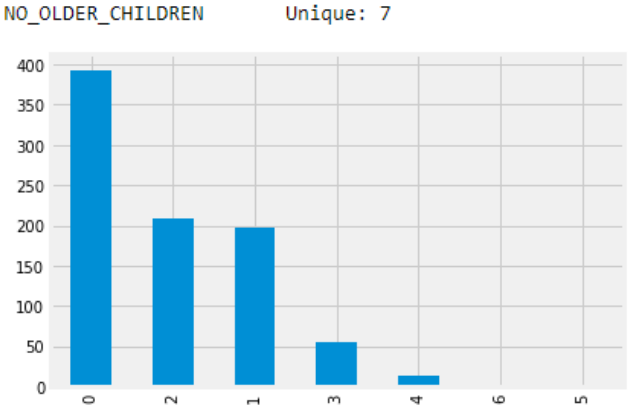
NO_YOUNG_CHILDREN

- The below bar graph shows that a majority of the observations (665) have no young children



NO_OLDER_CHILDREN

- The below bar graph shows that around 393 observations have no older children and very few have more than 3.

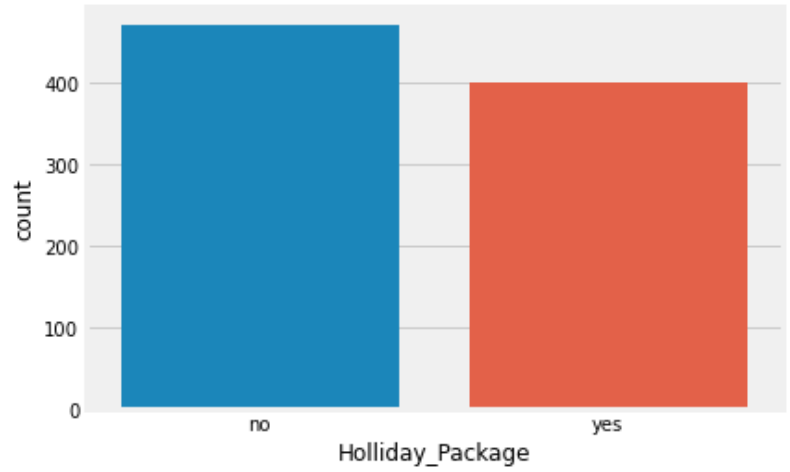


UNIVARIATE ANALYSIS FOR CATEGORICAL VARIABLES

HOLLIDAY_PACKAGE

- The below countplot shows that a large number of employees (471) have opted 'No' for the holiday package
- The number of employees that opted for 'Yes' is not that significant. The number of people choosing 'Yes' and 'No' are almost similar, so this dataset appears to be balanced.

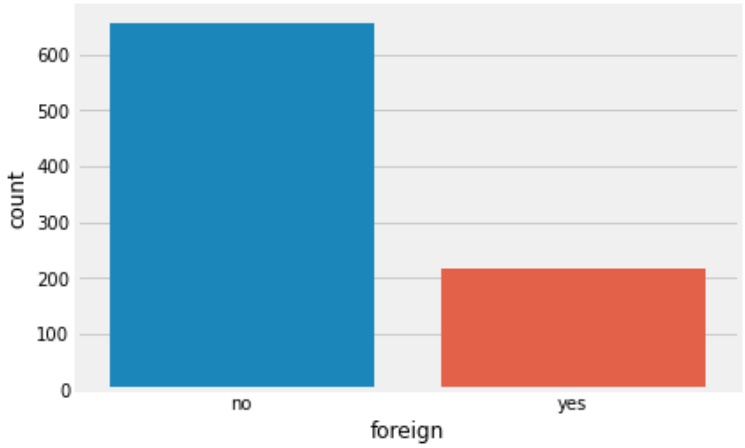
```
HOLLIDAY_PACKAGE          Unique no of values: 2

no     471
yes    401
Name: Holliday_Package, dtype: int64
```



FOREIGN

The below countplot shows that a large number of employees (656) are not foreigners.

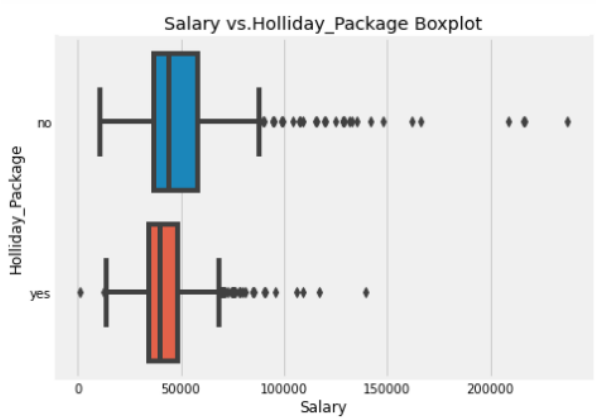
```
FOREIGN          Unique no of values: 2

no     656
yes    216
Name: foreign, dtype: int64
```

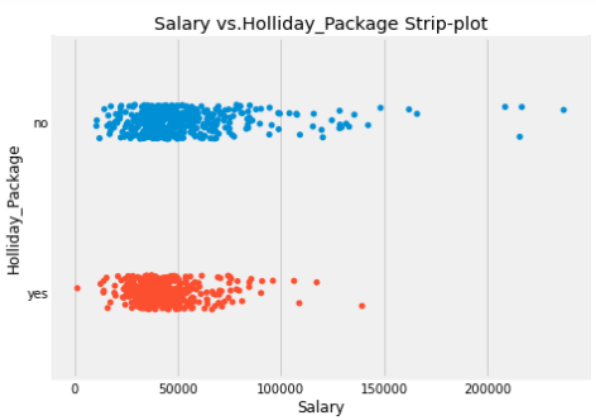


BIVARIATE ANALYSIS

SALARY Vs. HOLLIDAY_PACKAGE

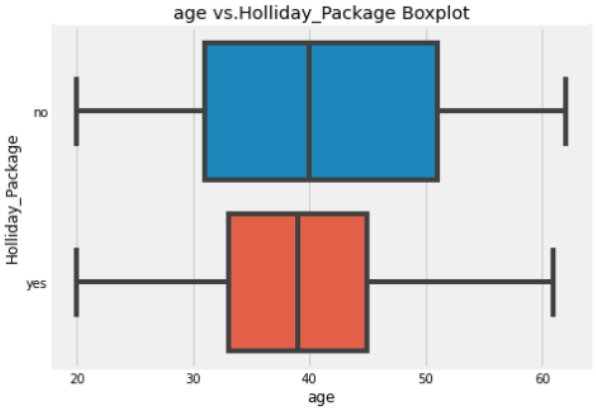
- From the boxplot, it appears that employees who have higher salaries opt 'No' for holiday package.
- The strip plot also shows the same. The employees who have salaries less than 100000 opt for 'Yes'.

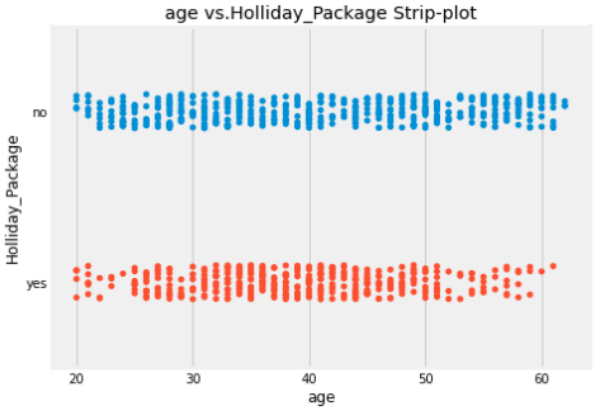




AGE Vs. HOLLIDAY_PACKAGE

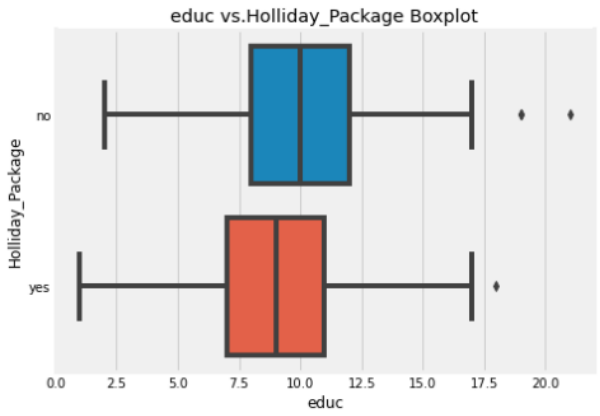
- From the boxplot, it appears that most of the employees who are above 40 years opt 'No' for holiday package.
- The strip plot shows that there are very few observations where people in their 20's above 50 years opt for 'Yes'.

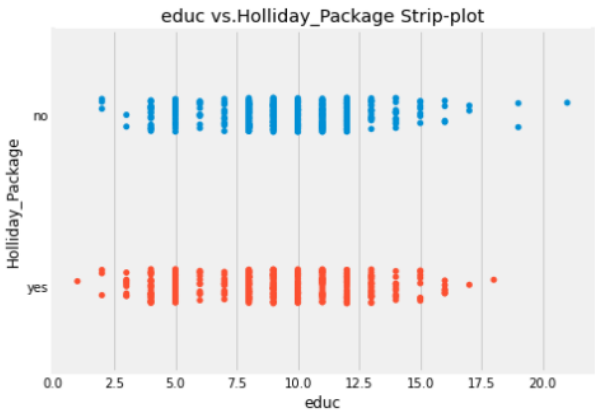




EDUCATION Vs. HOLLIDAY_PACKAGE

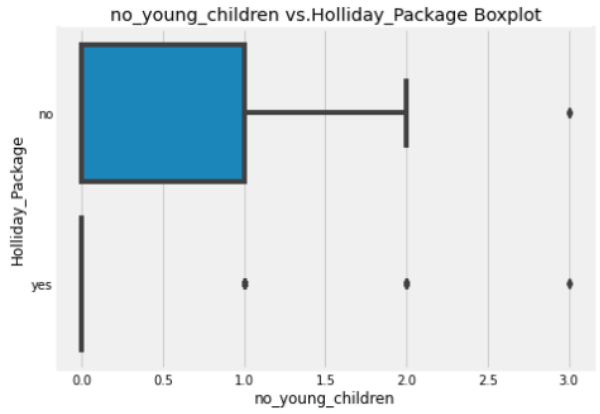
- From the boxplot, it appears that employees who have more than 11 years of education opt for 'No' for Holiday Package.
- There are a large number of data points where employees who have number of years of education between 7 and 13 opt for 'Yes'.

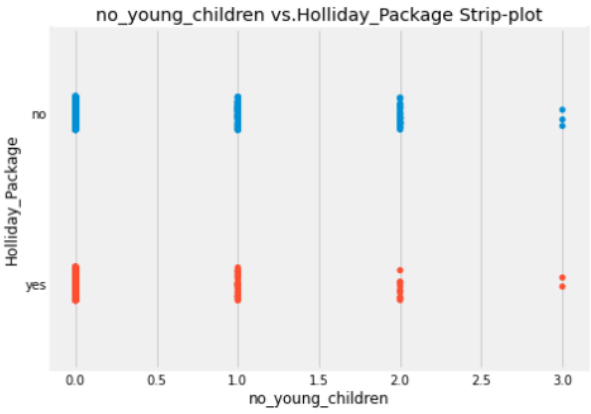




NO_YOUNG_CHILDREN Vs. HOLLIDAY_PACKAGE

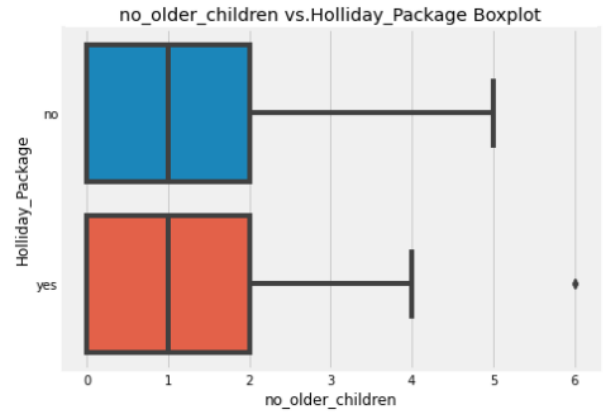
It appears that employees who don't have young children or who have 1 young child opt mostly opt for 'Yes' for the holiday package.

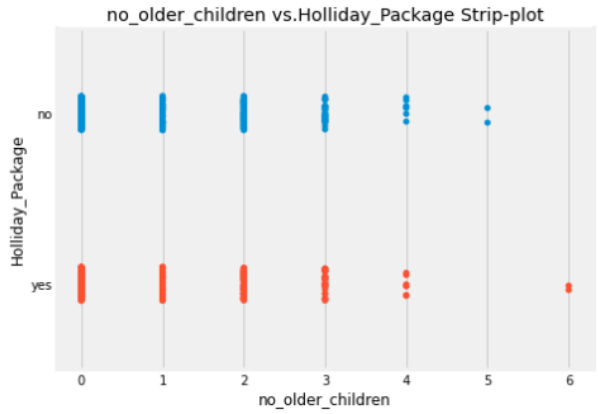




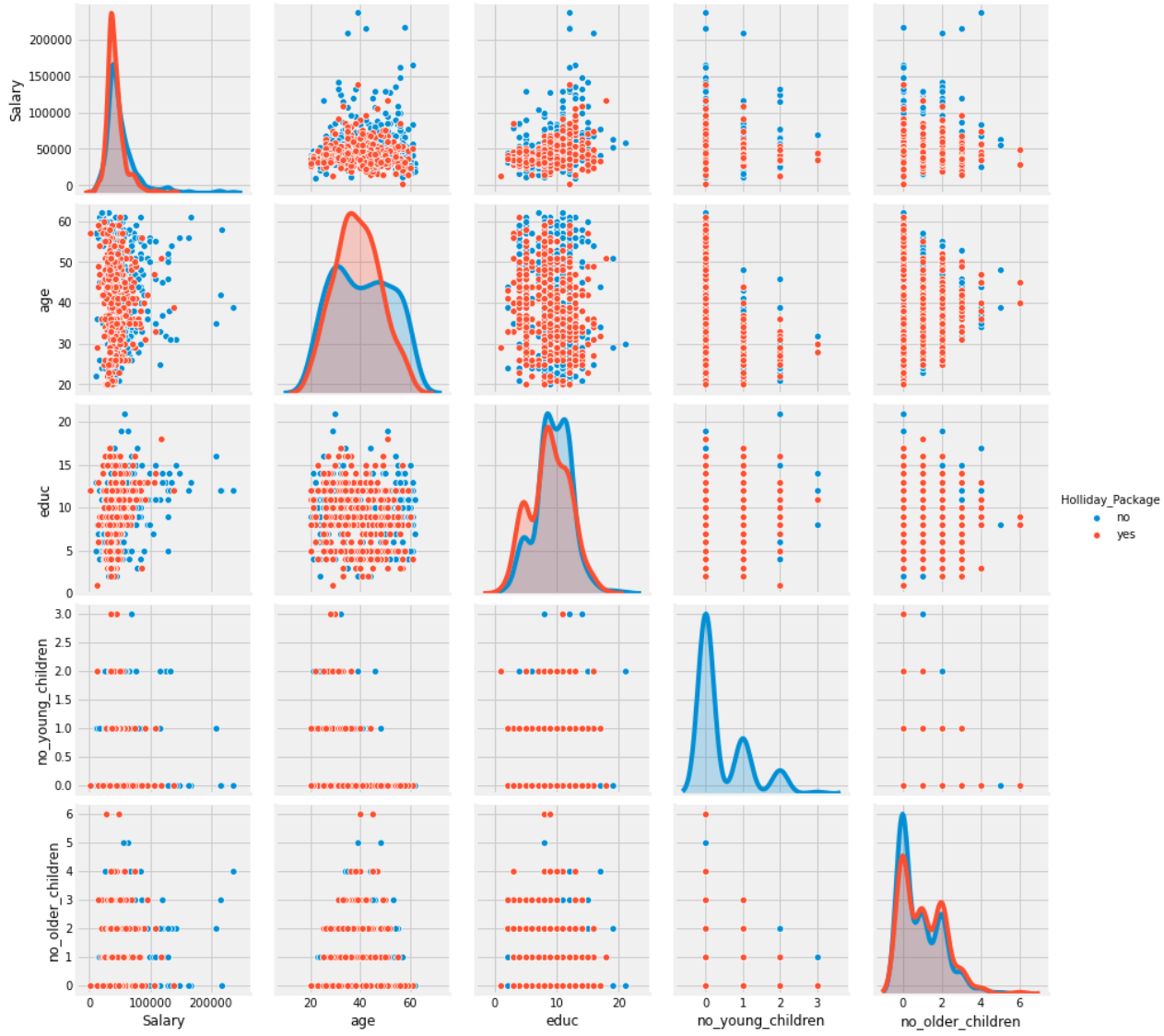
NO_OLDER_CHILDREN Vs. HOLLIDAY_PACKAGE

It appears that very few employees with more than 4 older children, who opt 'Yes' for the holiday package.

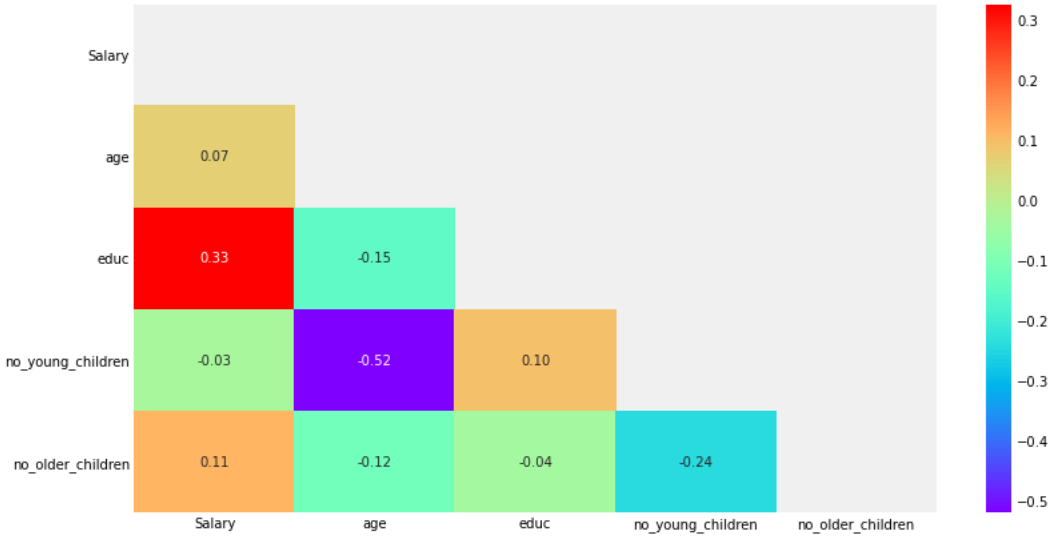




The below pairplot displays no significant correlations between the variables



This is also confirmed in the heatmap as there doesn't appear to be any strong multi-collinearity.



Inferences

- The Univariate analysis indicates that most of the employees do not have children (younger/older)
- Most of the employees are not foreigners.
- The Bivariate analysis shows no strong multi-collinearlity between the independent variables.
    - There is a negative correlation (-0.52) between 'age' and 'no_young_children', however, this is kept for now, as in the heatmap, correlations above a threshold of 0.6 or 0.7 were observed.

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).

Encoding string values

The object data type variables 'Holliday_Package' and 'foreign' are encoded into categorical values.

	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
1	0	48412	30	8	1	1	0
2	1	37207	45	8	0	1	0
3	0	58022	46	9	0	0	0
4	0	66503	31	11	2	0	0
5	0	66734	44	12	0	2	0

Test & Train data split

The independent variables (all the variables other than 'Holliday_Package') are taken as 'X', and the dependent variable 'Holliday_Package' is taken as 'y'.

Then, the dataframes 'X' and 'y' are split into the training data and testing data. Around 30% of the data from 'X' and 'y' is taken for testing and the remaining 70% is for training.

Logistic Regression

It is a supervised classification machine learning model which is similar to linear regression, however the linear equation is passed through a sigmoid curve so that the output is classified as either 0 or 1. In this case, 0 means 'No' for Holiday package and 1 means 'Yes'.

RandomizedSearchCV is applied to the logistic regression model in order to select the best optimized parameters.

```
RandomizedSearchCV(cv=5, estimator=LogisticRegression(), n_jobs=5,
                   param_distributions={'max_iter': [10000],
                                        'penalty': ['none', 'l1', 'l2'],
                                        'solver': ['newton-cg', 'lbfgs',
                                                   'liblinear'],
                                        'tol': [1e-09, 1e-06, 0.0001, 0.01]},
                   scoring='accuracy')
```

The best parameters are as follows:
'tol': 1e-06, 'solver': 'newton-cg', 'penalty': 'l2', 'max_iter': 10000

Linear Discriminant Analysis

It is a supervised machine learning model which uses a linear combination of independent variables to predict the class of the output variable. It is also a classification algorithm like Logistic Regression.

RandomizedSearchCV is applied to the model in order to select the best optimized parameters.

```
RandomizedSearchCV(cv=5, estimator=LinearDiscriminantAnalysis(), n_jobs=5,
                   param_distributions={'solver': ['svd', 'lsqr', 'eigen'],
                                        'tol': [1e-09, 1e-06, 0.0001, 0.01]},
                   scoring='accuracy')
```

The best parameters are as follows:

'tol': 1e-06, 'solver': 'eigen'

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.

The performance metrics are determined for each model for the training and test data sets and below are the inferences for each model.

PERFORMANCE OF PREDICTIONS ON TRAIN SET

LOGISTIC REGRESSION

I. The Classification report displays the below details:

- The **Precision** metric tells us how many of the classified data points are true positives. For Holiday package 1 ('Yes'), precision is 67%.
- The **Recall** metric tells us how many of the classified data points are true positives out of the true positives and false negatives. For Holiday package 1 ('Yes'), Recall is 55%.
- The **Accuracy Score** of the Logistic Regression model is 67.05 %. This is also shown in the below classification report for the field "accuracy".
- The **F1 Score** for the Holiday package 1 – which takes the harmonic mean of Precision and Recall (i.e. it checks for Type 1 and Type 2 errors) – is somewhat decent (60%).

```
LOGISTIC REGRESSION

Classification Report

Accuracy:        67.05 %

              precision    recall  f1-score   support

           0       0.67      0.77      0.72       332
           1       0.67      0.55      0.60       278

    accuracy                           0.67       610
   macro avg       0.67      0.66      0.66       610
weighted avg       0.67      0.67      0.67       610
```

II. The **Confusion Matrix** displays the number of correct and incorrect predictions made in the model in a tabular format. Below is the data understood from the confusion matrix.

TN (True Negatives) = 257

TP (True Positives) = 152

FN (False Negatives) = 126

FP (False Positives) = 75

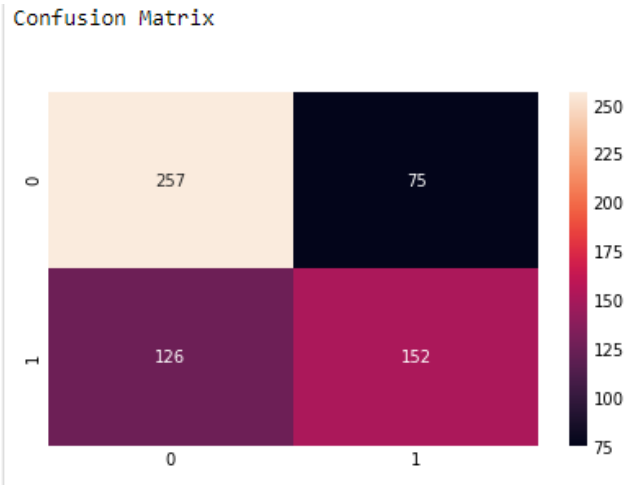
The Accuracy can also be calculated from the Confusion Matrix.

Accuracy = (TP + TN)/ (TP + TN + FP + FN)

= (152 + 257) / (152 + 257 + 75 + 126)

= 0.6704918

i.e. **67.05 %**



LINEAR DISCRIMINANT ANALYSIS

I. The Classification report displays the below details:

- The **Precision** for Holiday Package 1 is 66%.
- The **Recall** for Holiday Package 1 is 54%.
- The **Accuracy Score** of the Random Forest model is 66.56 %.
- The **F1 Score** for the Holiday Package 1 is a bit less (60%).

```
LINEAR DISCRIMINANT ANALYSIS

Classification Report

Accuracy:        66.56 %

              precision    recall  f1-score   support

           0       0.67      0.77      0.71       332
           1       0.66      0.54      0.60       278

    accuracy                           0.67       610
   macro avg       0.66      0.66      0.66       610
weighted avg       0.67      0.67      0.66       610
```

II. The **Confusion Matrix** displays the below data.

TN (True Negatives) = 255

TP (True Positives) = 151

FN (False Negatives) = 127

FP (False Positives) = 77

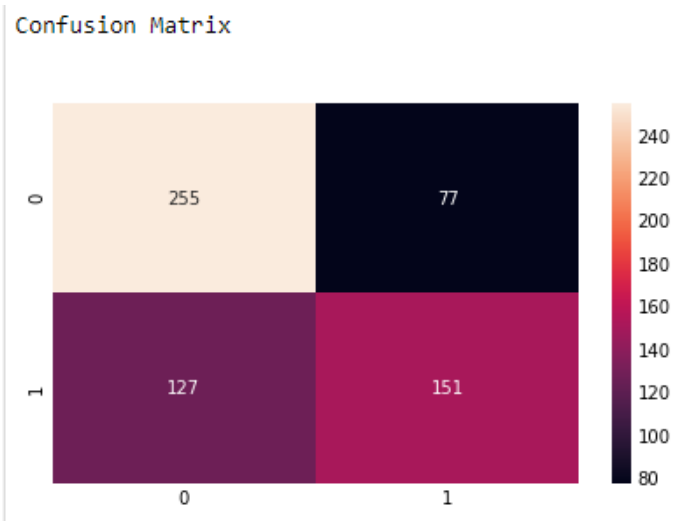
The Accuracy can also be calculated from the Confusion Matrix.

Accuracy = (TP + TN)/ (TP + TN + FP + FN)

= (151 + 255) / (151 + 255 + 77 + 127)

= 0.6655737

i.e. **66.56 %**



ROC AUC Score for the train data set:

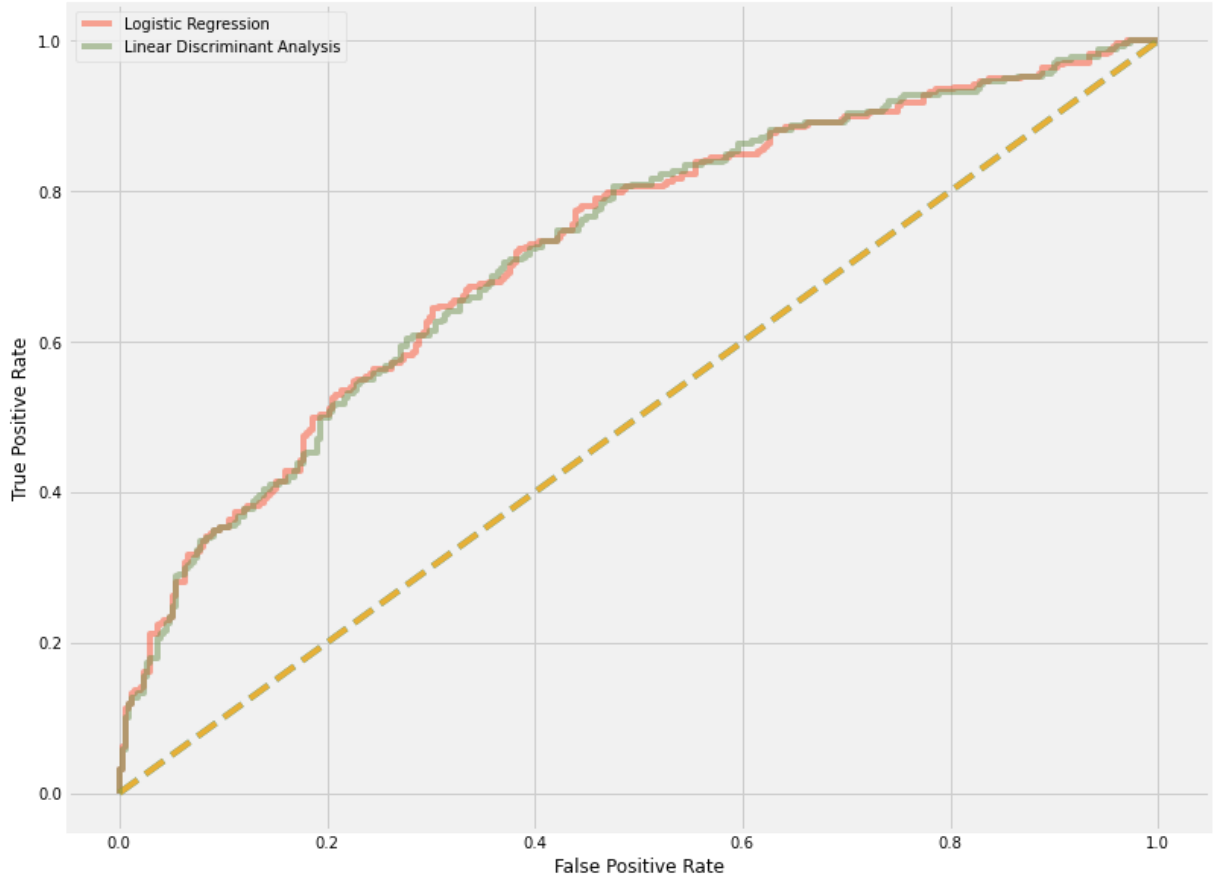
ROC stands for Receiver Operating Characteristic Curve and AUC stands for Area Under Curve.

True Positive Rate (TPR) i.e. Recall - TP / (TP + FN)

False Positive Rate (FPR) is - FP / (FP + TN)

The ROC curve plots the TPR (True Positive Rate) versus the FPR (False Positive Rate) at different classification threshold values. Generally, steeper the curve, stronger the model.

From the below graph, it is appears that both Logistic Regression and Linear Discriminant Analysis models have done equally well on the training set.



The ROC AUC scores for the models are as follows:

AUC Score Logistic Regression	**0.7221873103926497**
AUC Score Linear Discriminant Analysis	**0.7210713357025224**

PERFORMANCE OF PREDICTIONS ON TEST SET

LOGISTIC REGRESSION

I. The Classification report displays the below details:

- The **Precision** for Holiday Package 1 is 68%.
- The **Recall** for Holiday Package 1 is 53%.
- The **Accuracy Score** of the Random Forest model is 66.41 %.
- The **F1 Score** for the Holiday Package 1 is a bit less (60%).

LOGISTIC REGRESSION

Classification Report

Accuracy: 66.41 %

	precision	recall	f1-score	support
0	0.65	0.78	0.71	139
1	0.68	0.53	0.60	123
accuracy			0.66	262
macro avg	0.67	0.66	0.65	262
weighted avg	0.67	0.66	0.66	262

II. Below is the data understood from the **Confusion Matrix**.

TN (True Negatives) = 109

TP (True Positives) = 65

FN (False Negatives) = 58

FP (False Positives) = 30

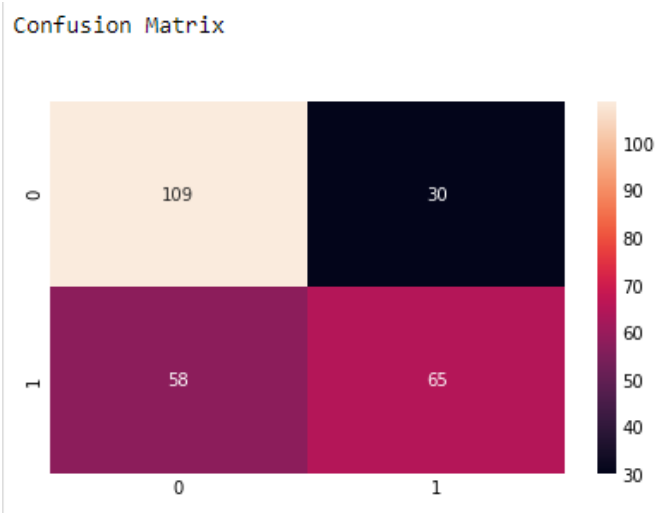
The Accuracy can also be calculated from the Confusion Matrix.

Accuracy = (TP + TN)/ (TP + TN + FP + FN)

= (65 + 109) / (65 + 109 + 30 + 58)

= 0.6641221

i.e. **66.41 %**



LINEAR DISCRIMINANT ANALYSIS

I. The Classification report displays the below details:

- The **Precision** for Holiday Package 1 is 68%.
- The **Recall** for Holiday Package 1 is 51%.
- The **Accuracy Score** of the Random Forest model is 66.03 %.
- The **F1 Score** for the Holiday Package 1 is a bit less (59%).

LINEAR DISCRIMINANT ANALYSIS

Classification Report

Accuracy: 66.03 %

	precision	recall	f1-score	support
0	0.65	0.79	0.71	139
1	0.68	0.51	0.59	123
accuracy			0.66	262
macro avg	0.67	0.65	0.65	262
weighted avg	0.66	0.66	0.65	262

II. The **Confusion Matrix** displays the below data.

TN (True Negatives) = 110

TP (True Positives) = 63

FN (False Negatives) = 60

FP (False Positives) = 29

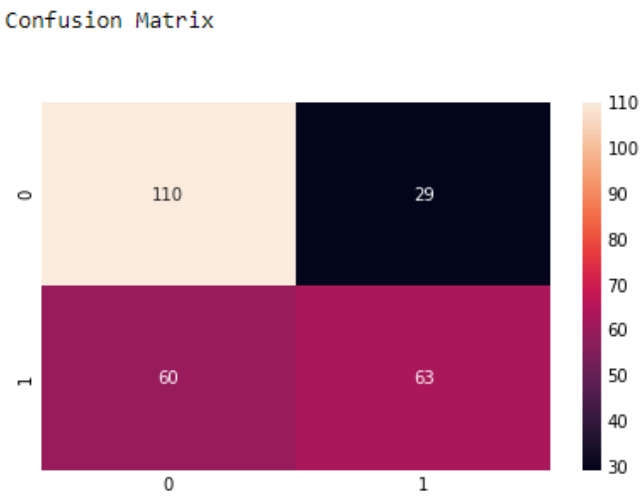
The Accuracy can also be calculated from the Confusion Matrix.

Accuracy = (TP + TN)/ (TP + TN + FP + FN)

= (63 + 110) / (63 + 110 + 29 + 60)

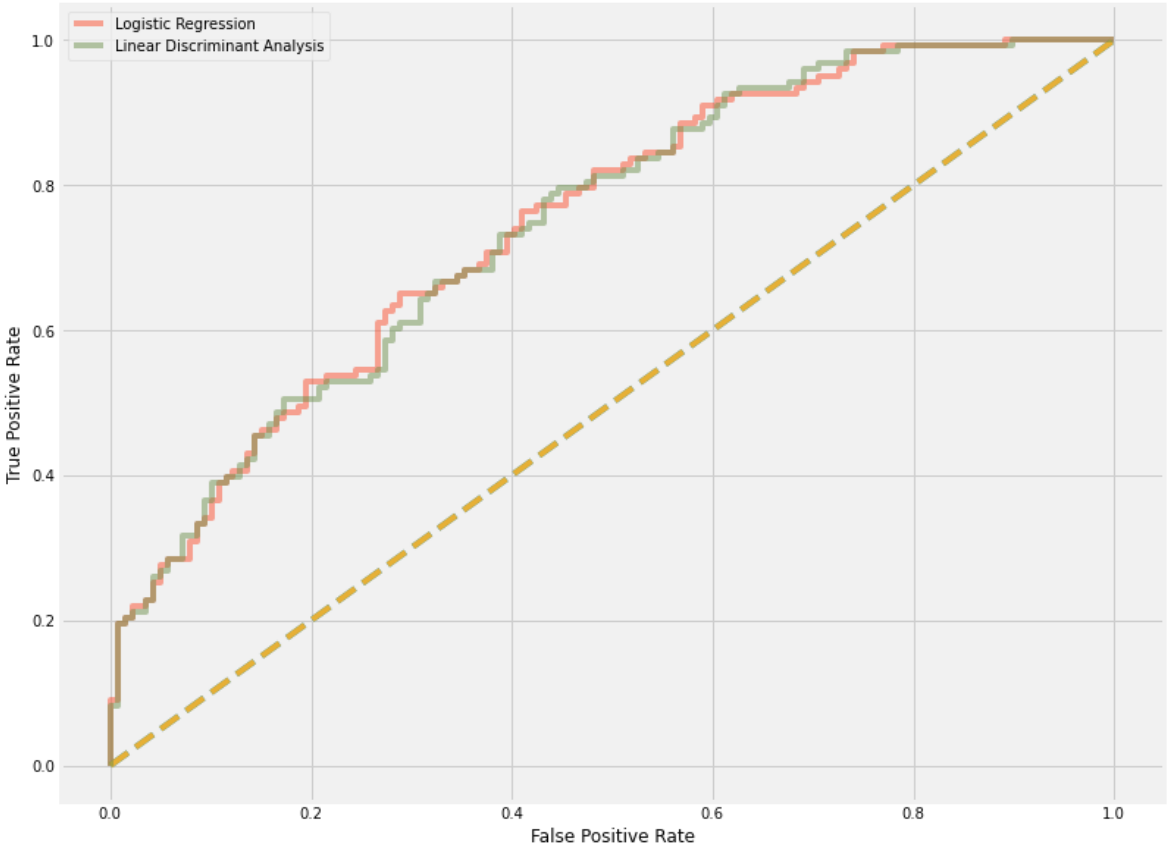
= 0.6603053

i.e. **66.03 %**



ROC AUC Score for the train data set:

From the below graph, it is again evident that the performance on the test set is almost similar for both models.



The ROC AUC scores for the models are as follows:

AUC Score Logistic Regression **0.7477335205006727**

AUC Score Linear Discriminant Analysis **0.7459203369012107**

MODEL COMPARISON

The performance metrics for each model are summarized in the below table.

Metrics	Logistic Regression	Linear Discriminant Analysis
Accuracy_Train	67.05	66.56
Accuracy_Test	66.41	66.03
F1_score_Train	60.2	59.68
F1_score_Test	59.63	58.6
AUC_Score_Train	72.22	72.11
AUC_Score_Test	74.77	74.59

Inferences:

1. Overall, the Logistic Regression model performs slightly better than the Linear Discriminant Analysis model (LDA). Below are the parameters for the optimized model.

   'tol': 1e-06, 'solver': 'newton-cg', 'penalty': 'l2', 'max_iter': 10000

2. The accuracy score of Logistic Regression for the train and test sets are 67.05% and 66.41% respectively.

   The Area under the curve for the train and test sets is 72.22% and 74.77% respectively.

   There is no significant difference in the performance metrics between the train and test sets, hence it is inferred that Logistic Regression is the best model for this case study.

3. The F1 Score the train and test sets for Logistic Regression are almost equal.

4. The Confusion Matrix for the test sets for both models show that Logistic Regression classified 174 true data points, i.e. true negatives and true positives were predicted correctly.

5. Hence, Logistic Regression will be chosen for predicting whether employees will opt for the Holiday package or not.

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

Please explain and summarise the various steps performed in this project. There should be proper business interpretation and actionable insights present.

The objective of this case study is to select the best model which has higher performance in predicting whether the employees will choose the holiday package or not.

Below are the steps that were followed before selecting the optimized model

1. The exploratory data analysis was done in order to find any relationship between the holiday package (Yes/No) and the other features.
    a. Univariate Analysis was done (boxplot/distribution plot/countplot)
    b. Bivariate Analysis was done (boxplot/strip-plot/pairplot/heatmap)
2. No features were dropped as there are no strong correlations.
3. The object type variables ('Holliday_Package', 'foreign') were encoded into categorical values.
4. The data is split into training and testing sets.
5. RandomizedSearchCV was used in order to find the best hyperparameters for an optimized model for both Logistic Regression and LDA.
6. After hyperparameter tuning, the models are fit to the training data using the best parameters for each.
7. Predictions are made for the training and testing sets.
8. The performance metrics (Accuracy Score, F1 Score, AUC_Score) are computed for both models for the train and test sets.
9. It was determined that Logistic Regression is the best model to use for achieving this project's objective, as it performs a bit better than LDA.

Insights:

- Around 75% of the data has employees having no young children.
- As for the number of older children, most of the employees have 2.
- A majority of the employees (656) are not foreigners.
- Those with higher salaries don't seem to opt for the Holiday package.
- The best age range to concentrate on for promoting holiday packages is 30-50 years.
- Employees with less than 4 older children appear to opt 'Yes' for the Holiday package.
- The Logistic Regression model can be used for predicting the output for new data.
- The equation for the logistic regression model is:
    - **(-0.0 * Salary) + (-0.05 * age) + (0.03 * educ) + (-1.11 * no_young_children) + (0.02 * no_older_children) + (1.15 * foreign)**
    - Based on the linear equation, it appears that 'foreign', 'no_young_children' , 'age' and 'education'are relevant attributes.

Recommendations:

- The travel agency can focus on employees having 3 older children or less for promoting the holiday package.
- A discussion can be held to see if any improvements can be made in the holiday package so that the employees having higher salaries are also more likely to opt for it.
- Holiday packages can be sold to employees in the age range 30-50.
- Employees having number of years of education between 2.5 to 12 can be focussed on.
- Some more samples can also be taken to confirm these findings.