

Tour and travel agency

Data Analysis



August 29, 2021

Siddhi Padekar

Table of Contents

|  |  |
| --- | --- |
| Summary | 2 |
| Introduction | 2 |
| Data Dictionary | 2 |
| Sample of the dataset | 3 |
| 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. | 3 |
| 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). | 11 |
| Encode the data (having string values) for Modelling. Data Split: Split the data into train and test (70:30). Apply Linear regression. Performance Metrics: Check the performance of Predictions on Train and Test sets using Rsquare, RMSE. | 13 |
| Inference: Basis on these predictions, what are the business insights and recommendations. | 16 |

Executive Summary

This Data set is of tour and travel agency, which deals in selling holiday packages. The dataset contains details of 872 employees of a company. Among these employees, some opted for the package and some did not. In this problem statement, we need to predict whether an employee will opt for the package or not. In addition, we need to find find out the important factors based on which the company will focus on particular employees to sell their packages.

Introduction

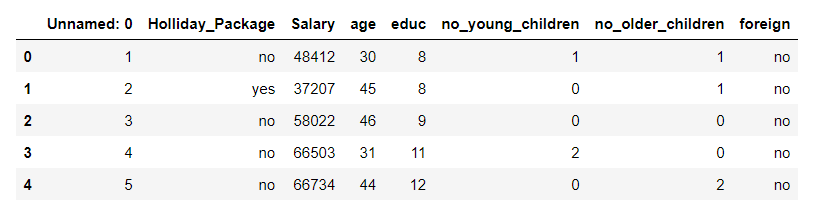
The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis. Explore the dataset, predict the holiday package on the bases of the details given in the dataset using logistic regression and LDA (linear discriminant analysis). Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score for each model Final Model.

Data Dictionary:

Below is the brief description of data set

|  |  |
| --- | --- |
| **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 |

Sample of the dataset:



Above is sample details of data set showing first 5 rows.

Holiday Package is the target variable and all other are predictor variable

Exploratory Data Analysis

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

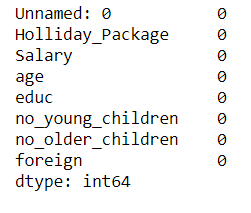
foreign 872 non-null object

dtypes: int64(6), object(2)

Shape of Data (872, 8)

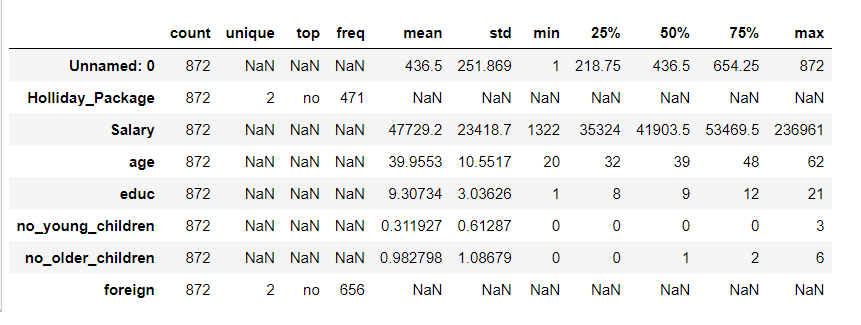
1. The data set contains 872 row, 8 columns .
2. In the given data set there are 6 Integer type features ,  2 Object type features.

Check for missing values in the dataset:

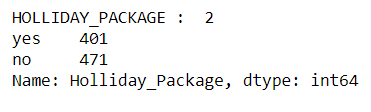
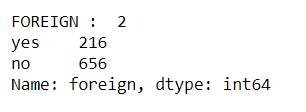


* There are No Nul values present in data set

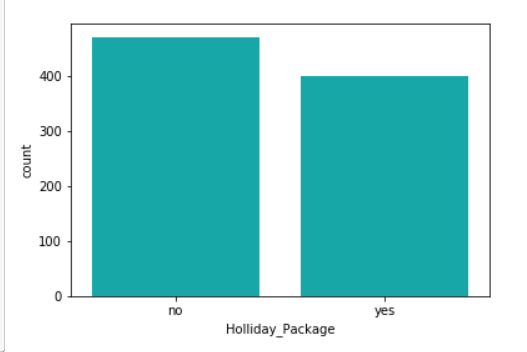
Mean, Standard deviation , Min, Max

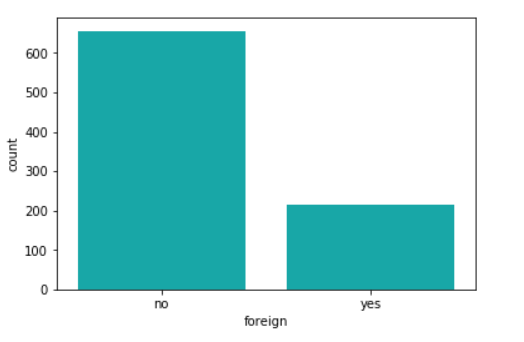


Categorising the object variable (Holiday Package , Foreign)

* Holiday Package is categorised into 2 heads –Yes, No (Most employees has not opted for Holiday package)
* Foreign is categorised into 2 heads –Yes, No ( There are less number of foreigners)





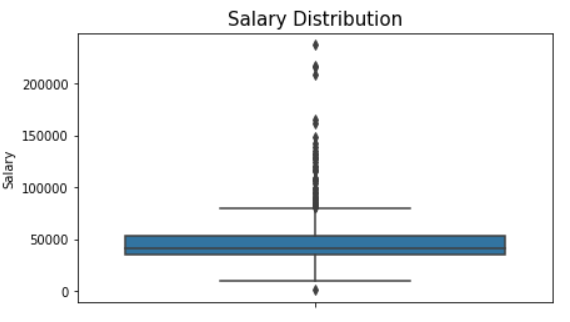
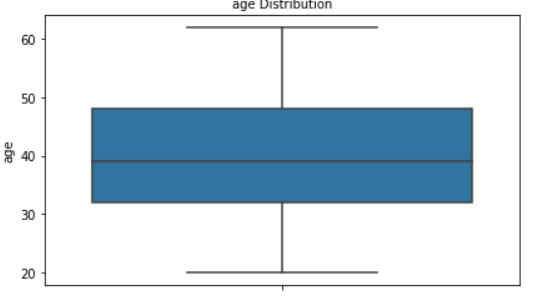
Check for Duplicate values and Removing Duplicate values in the dataset:

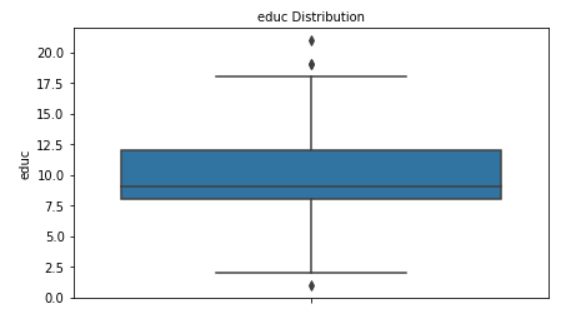
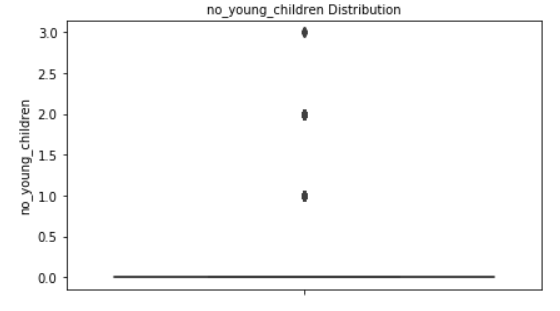
Number of duplicate rows = 0

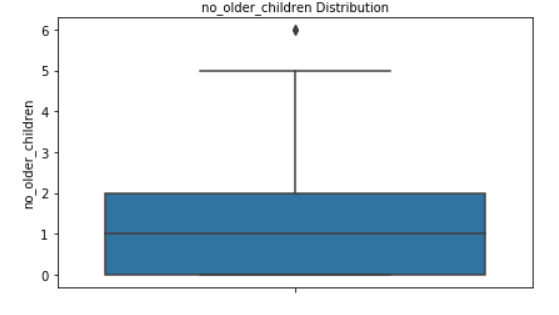
Note- we have dropped first column from the dataset ("Unnamed: 0") as this is only serial numbers

Outliers

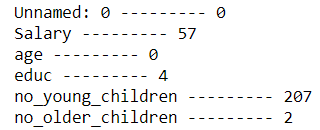
An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. A value that "lies outside" (is much smaller or larger than) most of the other values in a set of data.

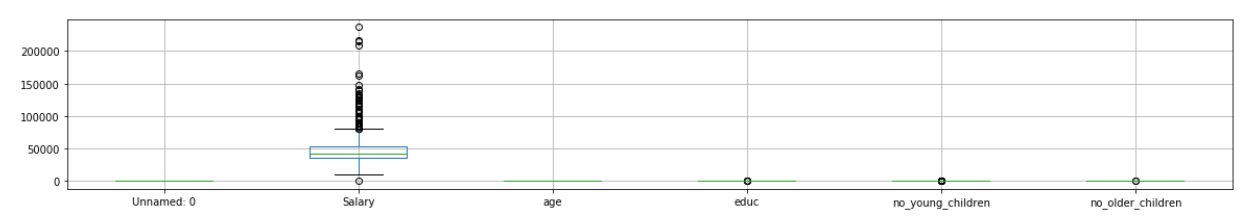
 



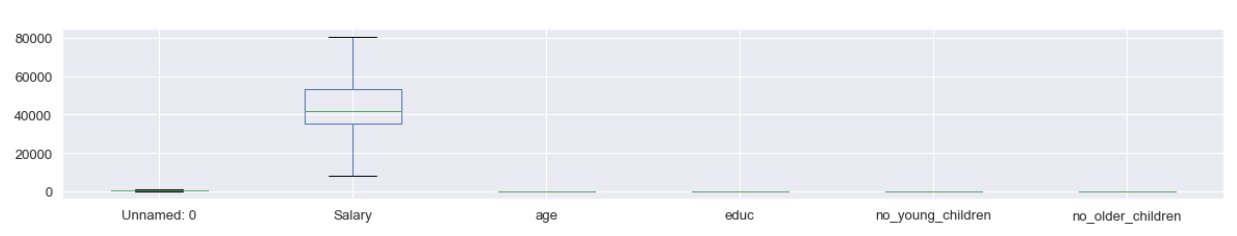
There are significant number of outliers in all the attributes



Before removing Outliers

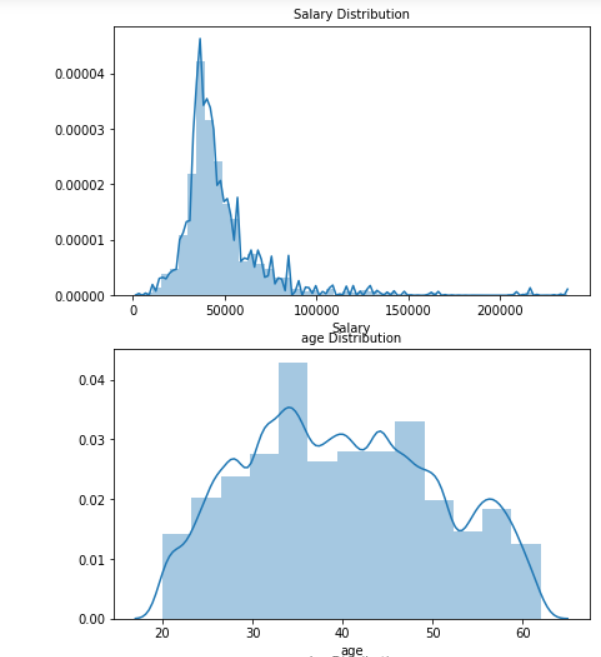


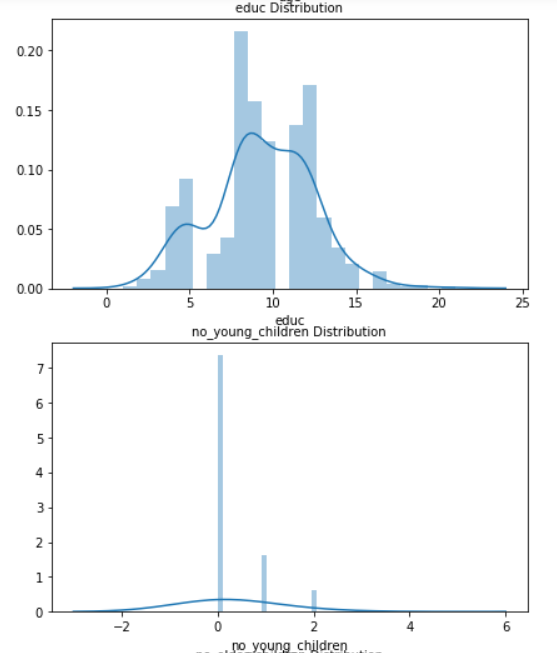
After removing outliers

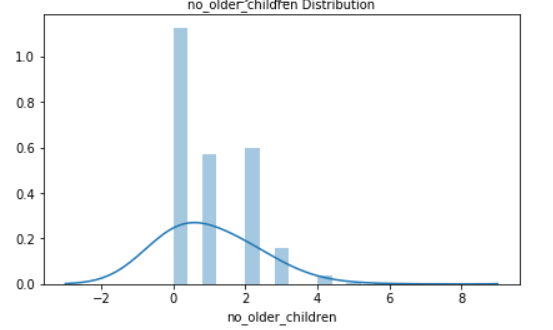


Univariate Analysis

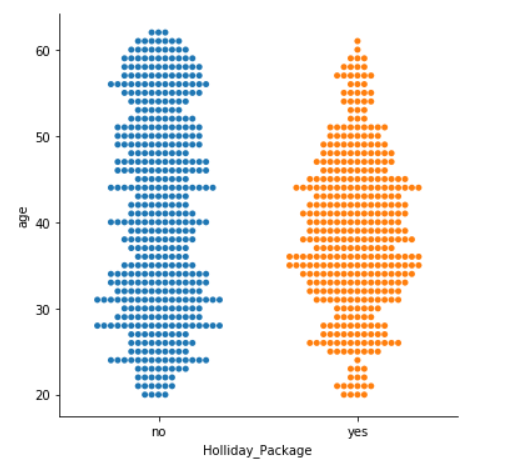
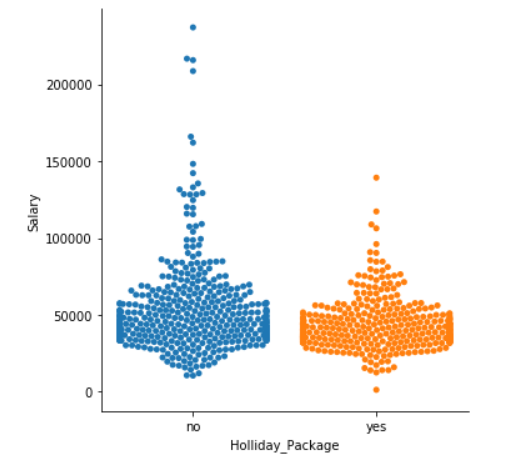
Histograms

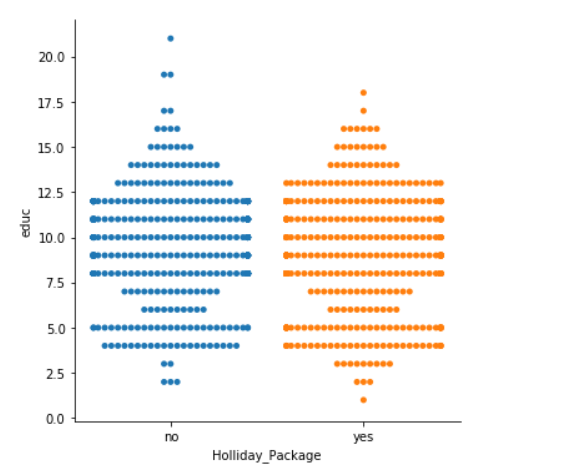
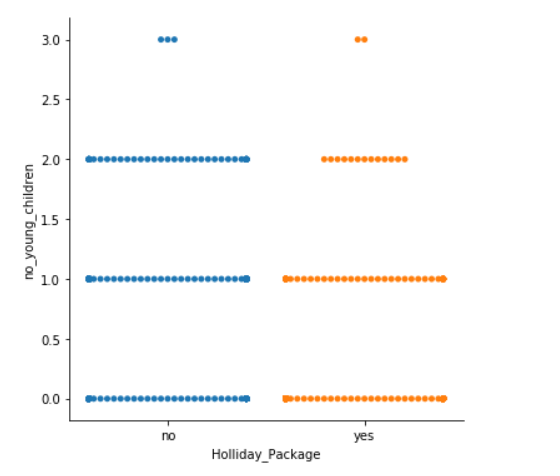


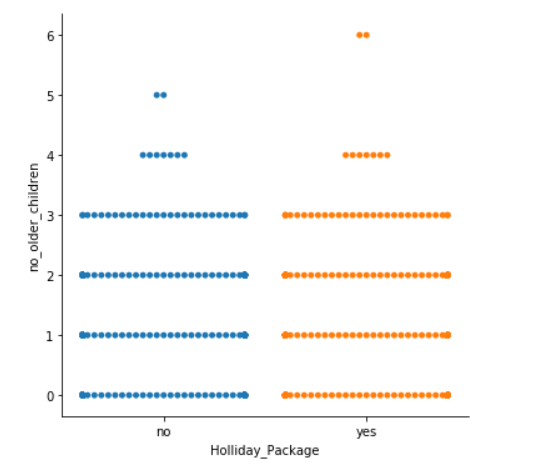


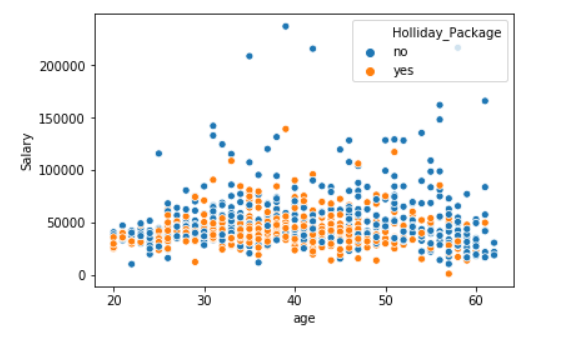
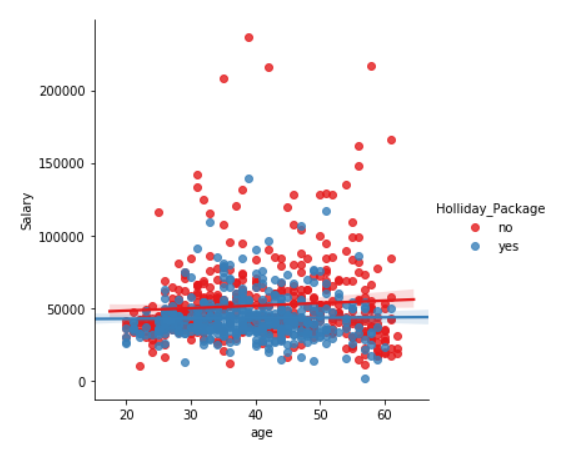


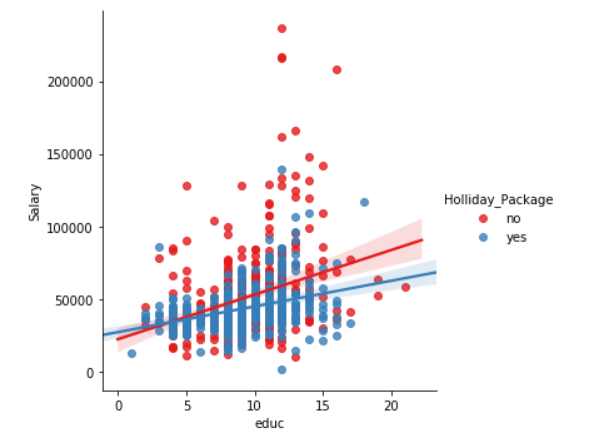
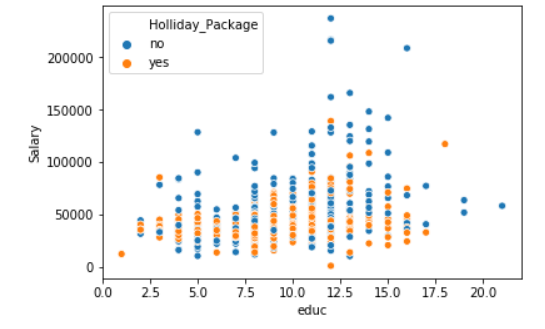
Bivariate Analysis

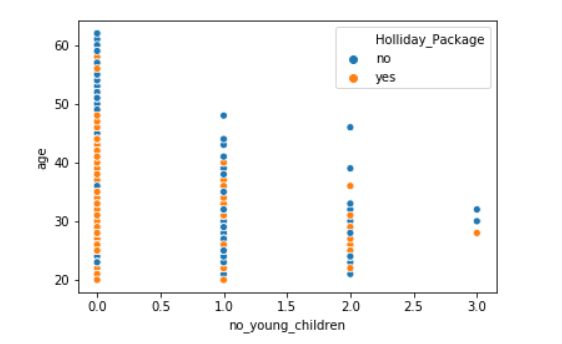
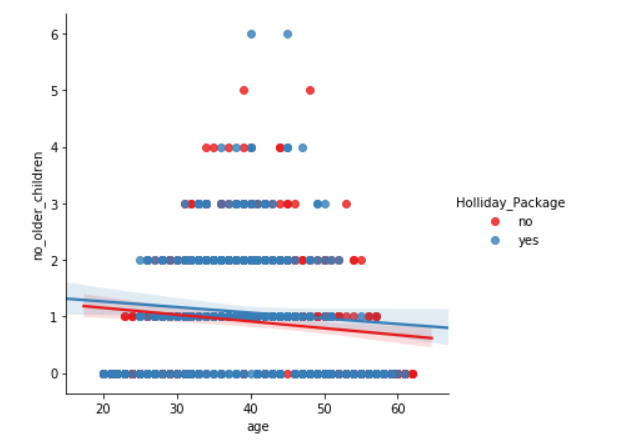
 

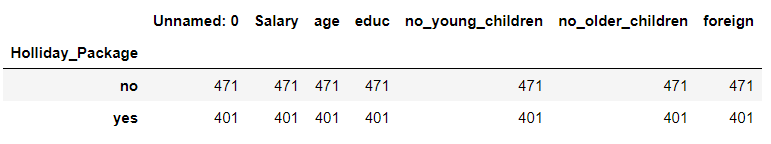
 



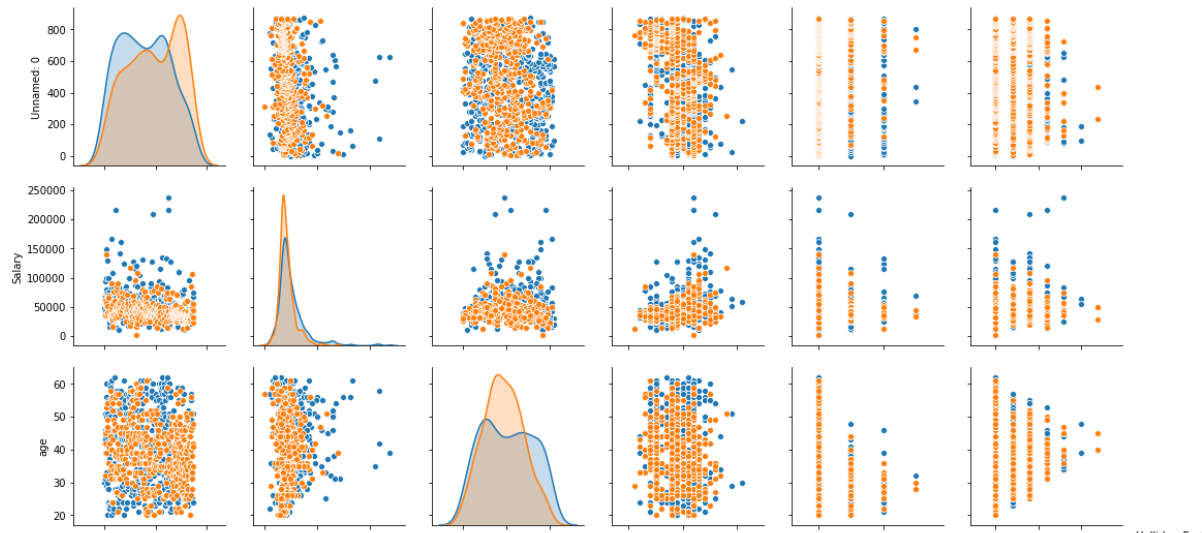
 

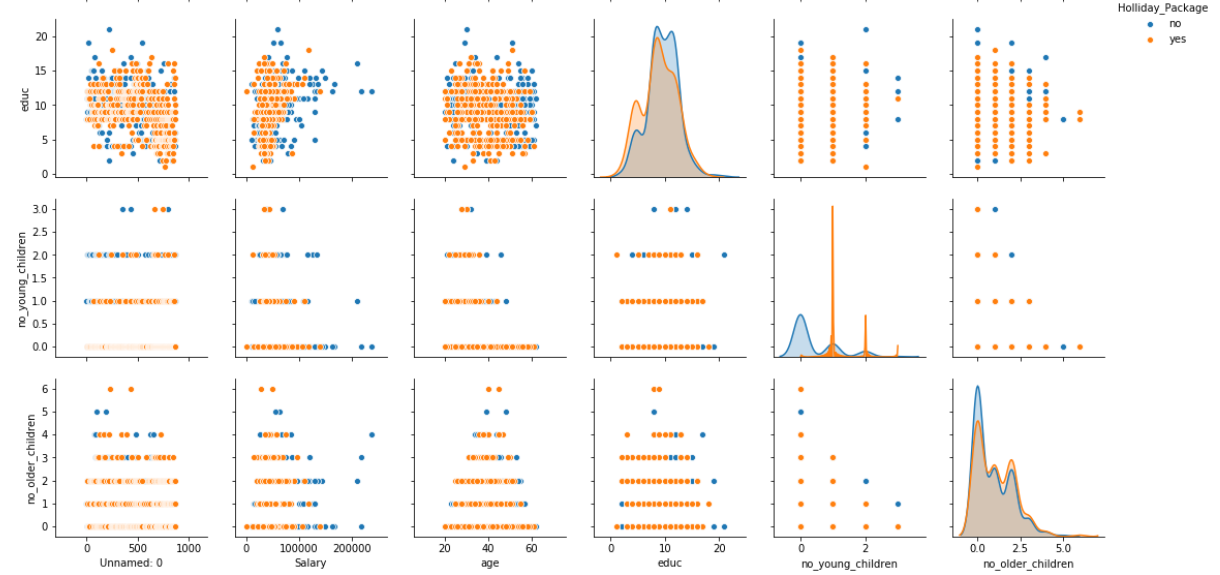
 



Pair Plot

Pair plot visualizes given data to find the relationship between them where the variables can be continuous or categorical.

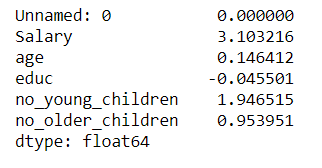




If we look at the distribution of record for holliday package opted and not opted , the opted part being interest to us is under representation of 401 as compared to not opted holliday package of 471. This kind of skew in the number of record we have in each class makes our overall accuracy score less reliable.

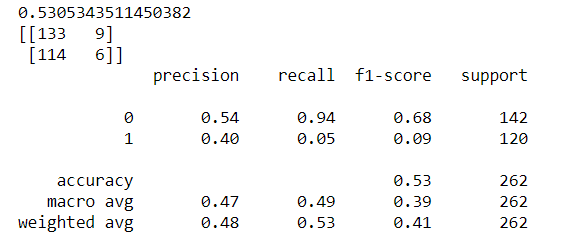
I we now observe our pair plot, What we see in the diagonal we see the estimates of density for each attributes. If we see the salary attribute both our classes are overlapping each other. Such attribute are unable to distinguish between the classes. Hence, they are not good attributes from classification point of view. Same is the case for age, no\_older childrens and educ .

Skewness of every attribute



The data is not properly skewed.

Logistic Regression



0.530 - is overall accuracy . This is not reliable metrics as the classes are skewed.

1st row is for Not opted Holliday package class

2nd row is opted Holliday package class.

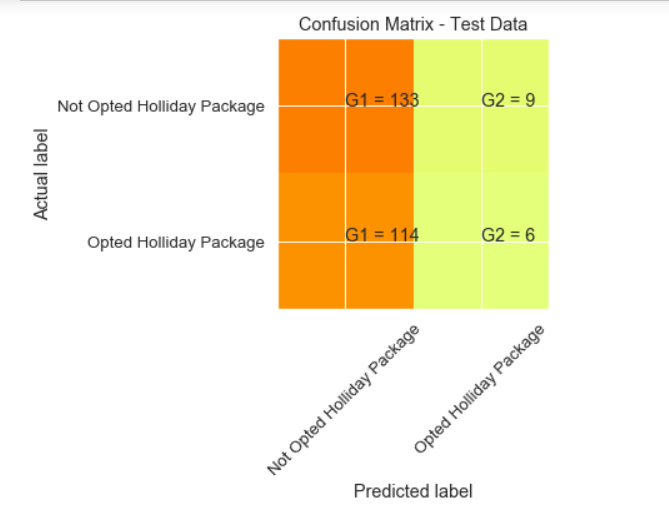
Similarly,

1st Column is for Not opted Holliday package class

2nd Column is for opted Holliday package class.

The score F1 is directly related to precision recall. ( 0- Not opted Holliday package class , 1- opted Holliday package class.)

Confusion matrix



G2- is true positive which shows 6 peoples were predicted correctly opted for Holliday package

G1- True negative 133

G2 - False positive (9)

G1 – False negative (114)

Performance Matrix

### Training Data and Test Data Confusion Matrix Comparison

Confusion Matrix is a square matrix, which in the ideal case, its main diagonal must be valued and other sides must be none.

| TP | FP |

--------------

| FN | TN |

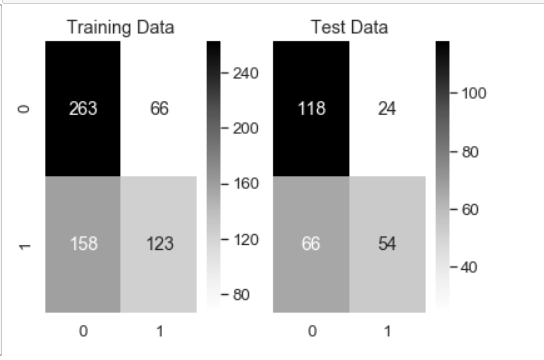
------+-------

TP – True positive (correct data)

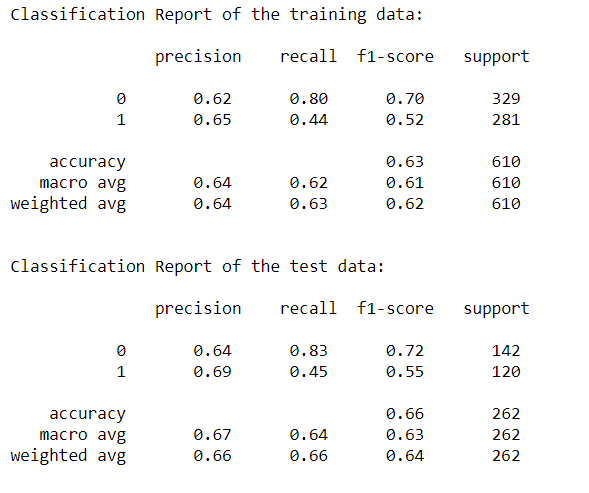
FP-False positive (data which is incorrect however still predicted as correct)

FN – False negative (Data which is correct however still predicted incorrect)

TN – True negative ( Incorrect data)



### Training Data and Test Data Classification Report Comparison



From the above we can conclude that performance metrix for test and train data is is almost same .

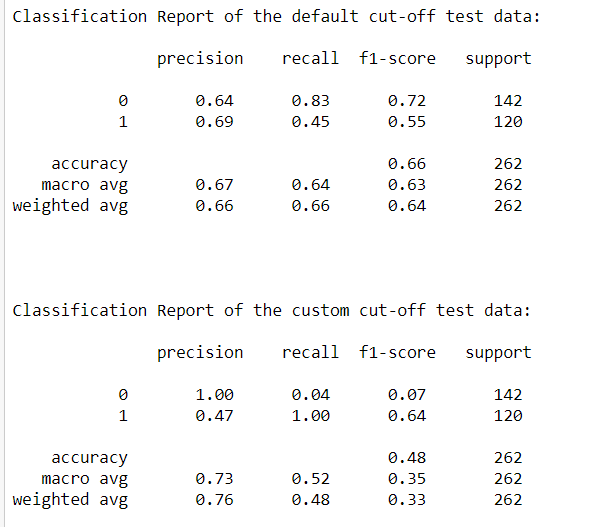
There is only little difference for f1-score , precision , recall for test and train data.

**AUC for Training data and Test data**



The model accuracy on the training as well as the test set is about 70%, which is roughly the same proportion as the class 0 observations in the dataset. This model is affected by a class imbalance problem. Since we only have 872 observations, if re-build the same LDA model with more number of data points, an even better model could be built.

Classification report of the default and custom cut-off test data



Inference :

For both the models i.e. Logistic regression and linear discriminant analysis AUC, recall , precision , accuracy are almost same , hence both the models are best.