Holiday Package

Project – Final Report

Table of Contents

List of Tables

List of Figures

1. INTRODUCTION

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 did not. 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 based on which the company will focus on particular employees to sell their packages

2. DATA DESCRIPTION

Field Name	Description	Data Type
Holiday Package	Opted for Holiday Package yes/no?	Categorical
Salary	Employee salary	Numeric
Age	Age in years	Numeric
Educ	Years of formal education	Numeric
no_of_young_children	The number of young children (younger than 7 years)	Numeric
no_of_older_children	Number of older children (older than 7 years)	Numeric
foreign	Yes/No	Categorical

Table 1: Description of dependent and independent variable

3. EXPLORATORY DATA ANALYSIS

Dataset has 872 rows and 6 features. Foreign and Holiday_package (dependent variable) is objective type and all other are integer.

Sample data

Holliday_ Package	Salary	age	educ	no_young_ children	no_older_ children	Foreign
no	48412	30	8	1	1	No
yes	37207	45	8	0	1	No
no	58022	46	9	0	0	No
no	66503	31	11	2	0	no
no	66734	44	12	0	2	no

Table 2: Sample data

Data Summary

	Holliday_ Package	Salary	Age	educ	no_young_ children	no_older_ children	foreign
count	872	872	872	872	872	872	872
unique	2	-	-	-	-	-	2
top	No	-	-	-	-	-	no
freq	471	-	-	-	-	-	656
mean	-	47729.17	39.96	9.31	0.31	0.98	-
std	-	23418.67	10.55	3.04	0.61	1.09	-
min	-	1322	20	1	0	0	-
0.25 Percentile	-	35324	32	8	0	0	-
0.5 Percentile	-	41903.5	39	9	0	1	-
0.75 Percentile	-	53469.5	48	12	0	2	-
max	-	236961	62	21	3	6	-

Table 3: Data summary

Key Observations:

- Dataset pertains to two type of employees i.e. foreigner and not foreigner, salary of the employees varies from 1322 to 236961, education varies from 1 to 21 years, age is from 20 to 62 years.
- no_young_children and no_older_children are appearing as an integer field. To get an impact of the dependent variable for children in both the categories, these two features would be converted into object type.

3.1 Univariate Analysis

3.1.1 Categorical Variable

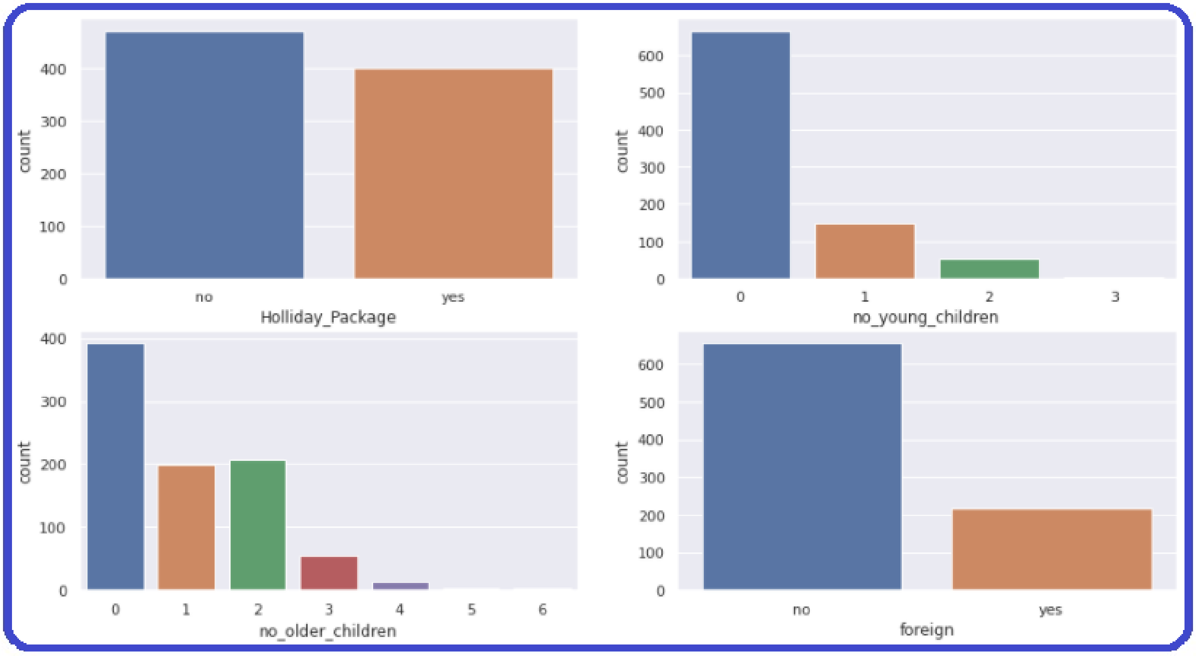


Figure 1: Univariate analysis – Categorical variable

Key Observations:

- 46% of employees(401 out of 872) opted for holiday package (Target variable is balanced)
- Employees with no children are in majority (young children as well as older ones)
- 25 % of employees are foreigner

3.1.2 Continuous Variable

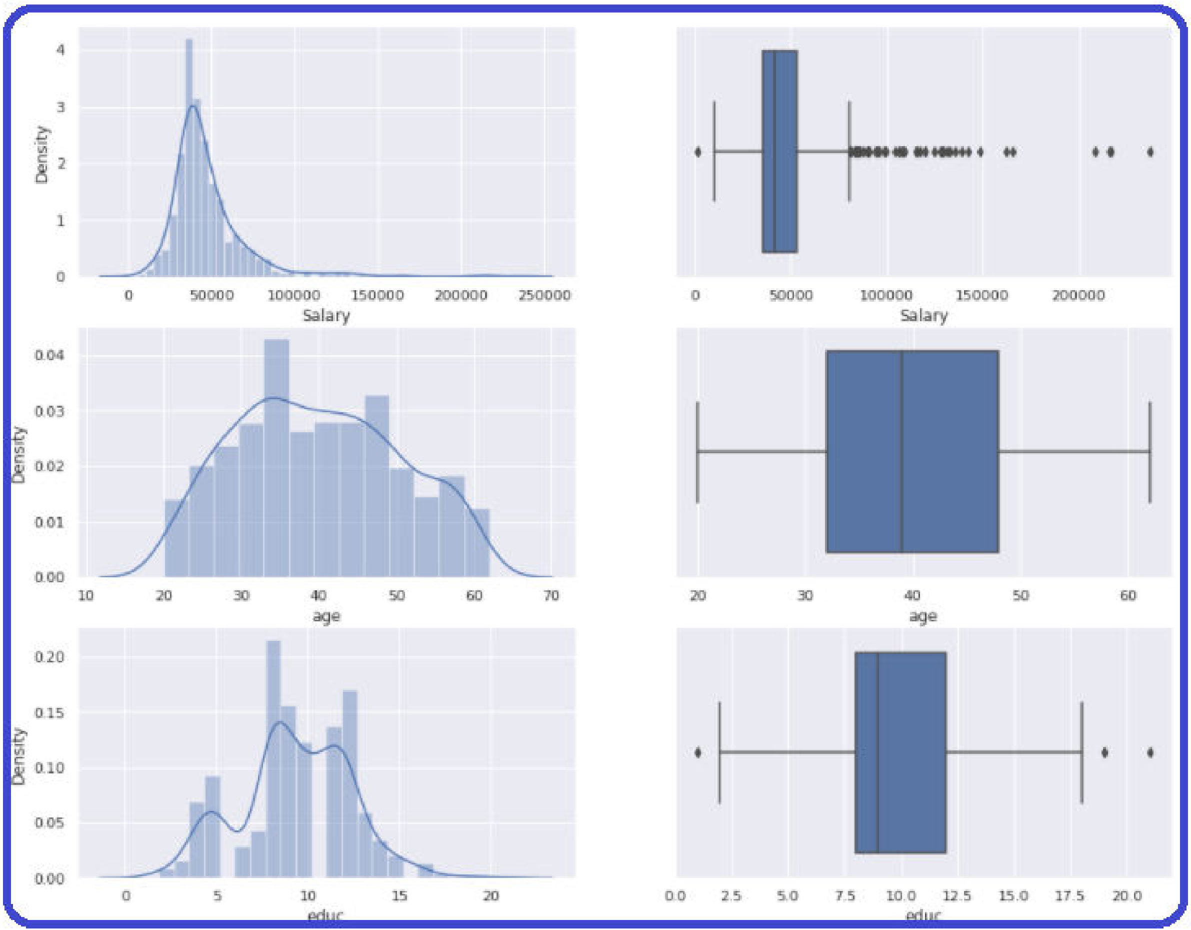


Figure 2: Univariate analysis – Continuous variable

Key Observations:

- Salary is right skewed with high dispersion
- The box plot of salary is also showing some extreme values which is normal with variable like salary
- Age is very close to normal distribution with no extreme data points
- Education is showing multimodal density plot
- Education boxplot shows very few outliers (in acceptable range)

3.2 Bivariate Analysis

3.1.1 Categorical Variable (Independent Variable – Categorical)

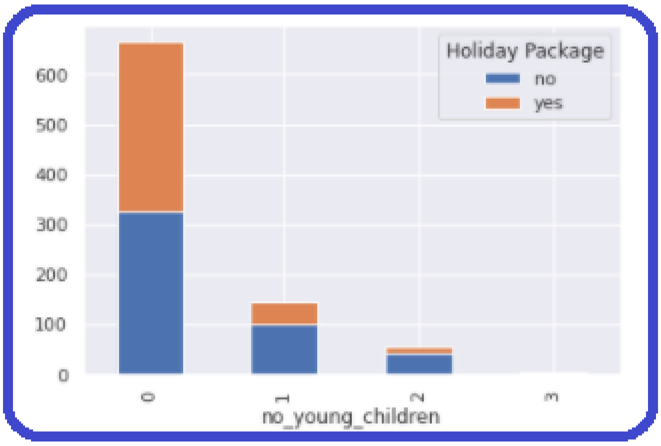


Figure 3: Bivariate analysis (Holiday Package – No. of young children)

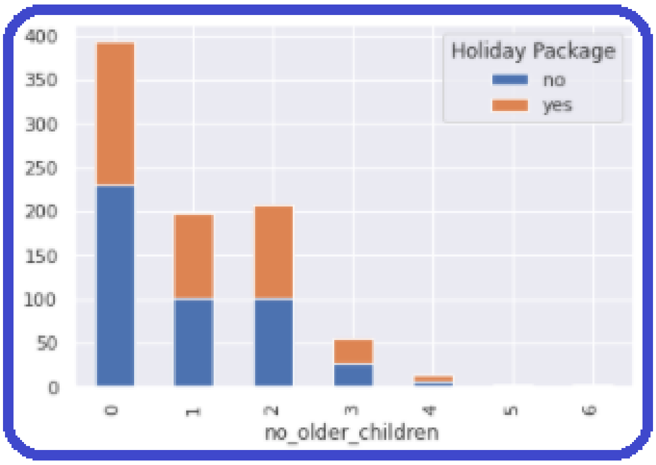


Figure 4: Bivariate analysis (Holiday Package – No. of older children)

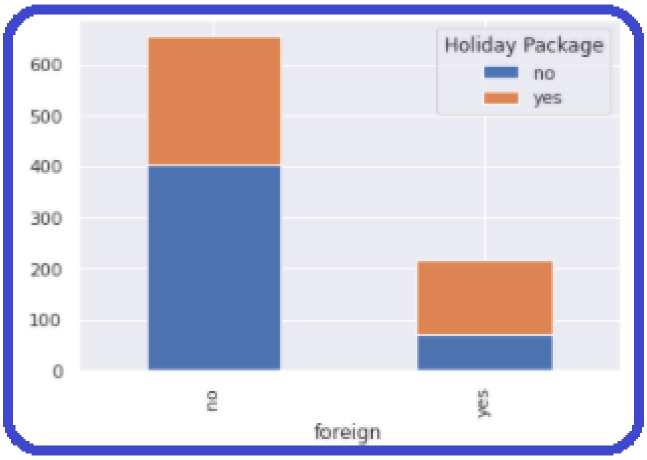


Figure 5: Bivariate analysis (Holiday Package – Foreigner)

Key Observations:

- Employees with one or two young child/children have less probability of choosing holiday package as compared to overall probability. This aspect of the variable could be a differentiator to find out who will opt for holiday package. However, as the majority class are having no children and the ration of people in this category who will opt for package is almost equal to the dataset ration of people who will opt for package, the variable will be a weak differentiator.
- As can be seen from the plot of no_of_old_children – holiday_package, the variable number of old children is not able to differentiate employees who will opt for package and who will not
- Plot foreign – Holiday package shows that if the person is foreigner than higher percentage of people are going for package as compared to average. Which means this variable can be a good differentiator in the model.

3.1.2 Continuous Variable

Salary vs Holiday Package

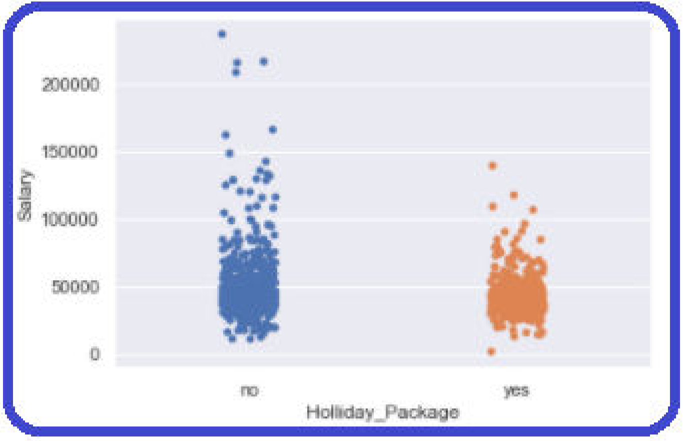


Figure 6: Scatter plot (Holiday package – Salary)

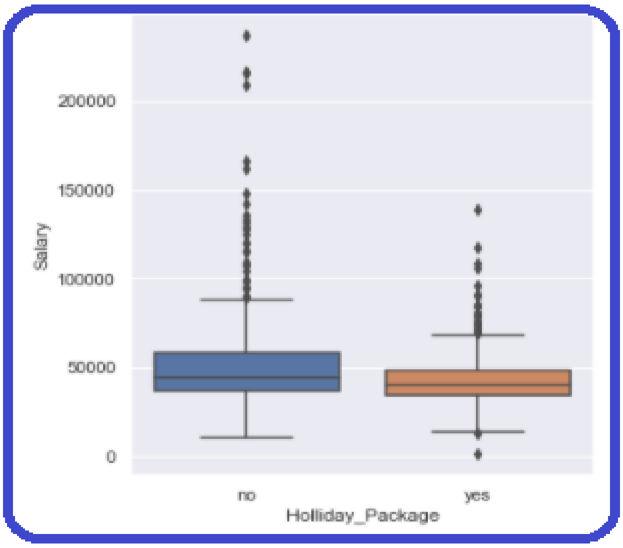


Figure 7: Boxplot (Holiday package – Salary)

Age vs Holiday Package

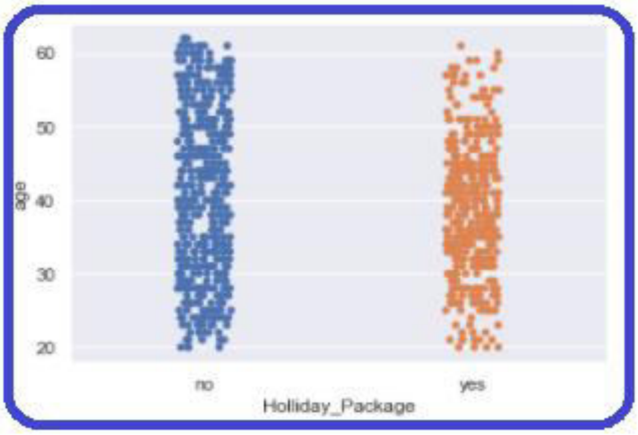


Figure 8: Scatter plot (Holiday Package - Age)

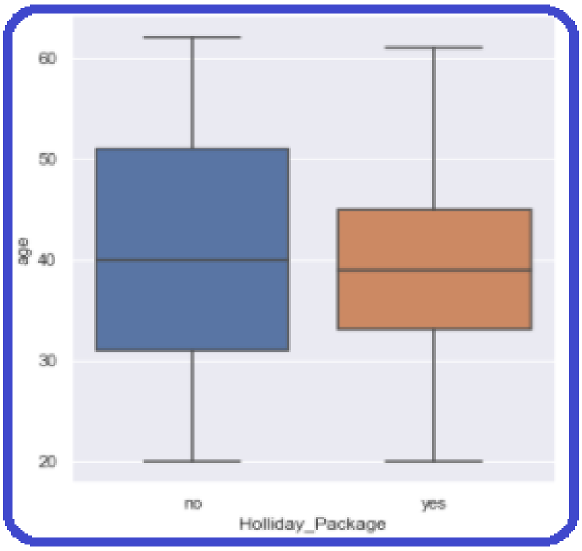


Figure 9: Boxplot (Holiday Package - Age)

Education vs Holiday Package

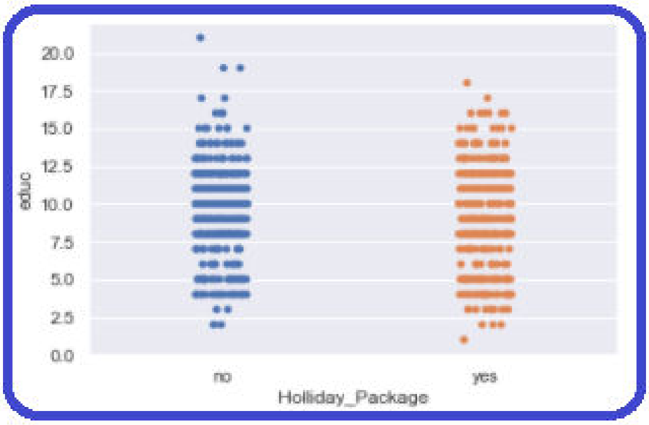


Figure 10: Scatter plot (Holiday package – Education)

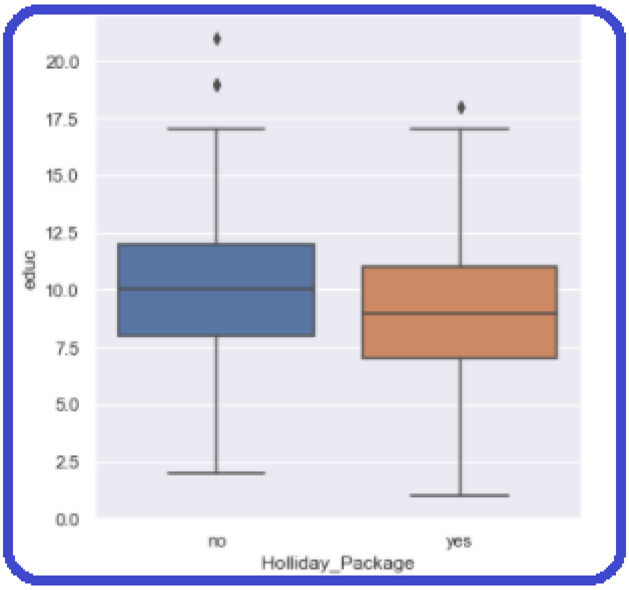


Figure 11: Boxplot (Holiday package – Education)

Key Observations:

- Hardly any big influence of these independent variable can be seen on our target variable

Pair plot

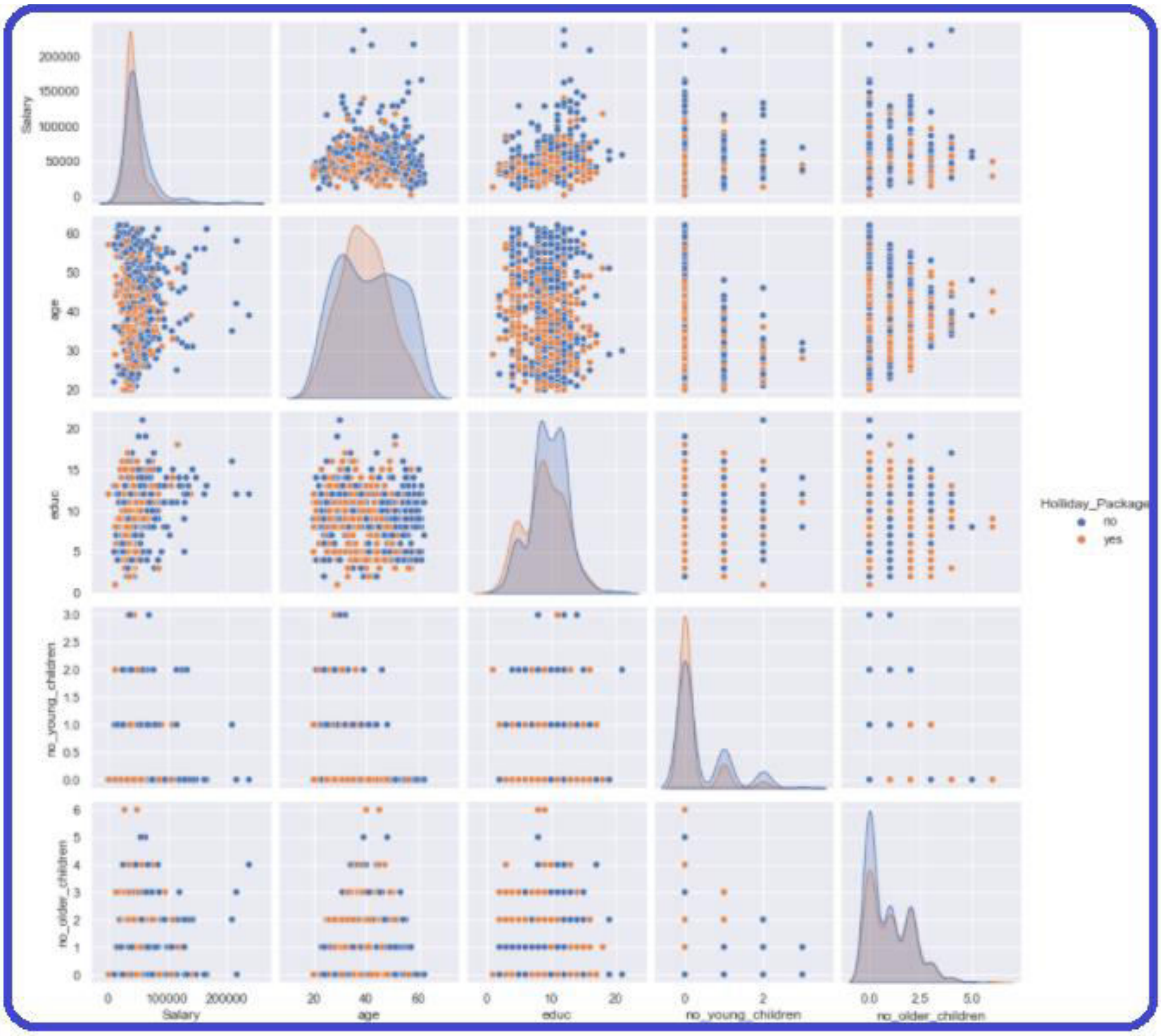


Figure 12: Pair plot

Key Observations:

- As we can see in diagonal of matrix, the distribution of employees opting for holiday package is overlapping with employees who are not, making it evident that none of the variable are strong predictor of target variable (holiday_package) individually.

4 MODEL DEVELOPMENT

4.1 Logistic Regression Model with Performance Metrics

Initial logistic model built without any specific parameter setting

Model

Coefficient of the model

```
The intercept for our model is 0.0005015138821658308
The coefficient for Salary is -9.783257449496757e-06
The coefficient for age is 0.005665659173597112
The coefficient for educ is 0.0030172843796183475
The coefficient for no_young_children is -0.0011201922417316747
The coefficient for no_older_children is 0.0016322405212726336
The coefficient for foreign is 0.0013590723444338504
```

Important features (None of the features has strong influence on target variable alone, as observed in EDA)

Important features identified are as follows

- Age
- Education
- No_older_children
- Foreign
- No_young_children (As no of young children increases, probability of holiday package decrease)
- Salary (Does not influence, same can be seen in bivariate analysis)

Performance metrics

ROC Curve – Train data

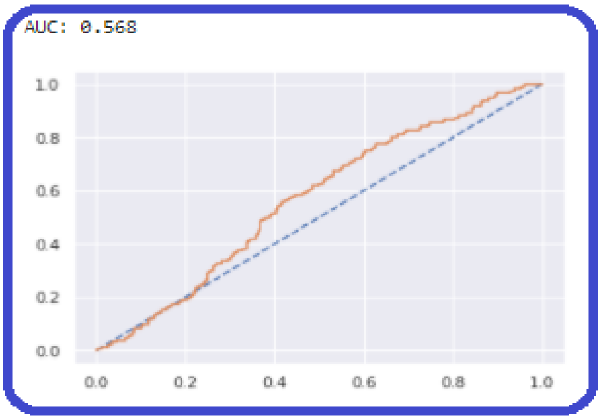


Figure 13: ROC Curve train data– Logistic Regression

ROC Curve – Test data

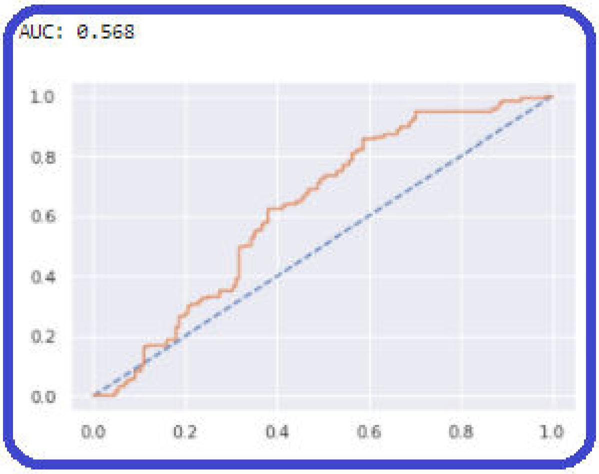
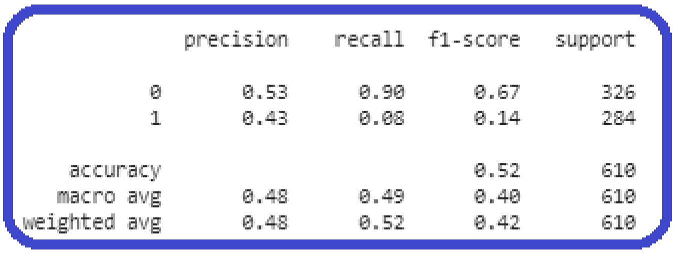


Figure 14: ROC Curve test data– Logistic Regression

Comment – The ROC curve with the AUC value of both train and test shows that the model performance will be not up to mark. The same can be seen in classification report.

<u>Classification report on Train data</u>



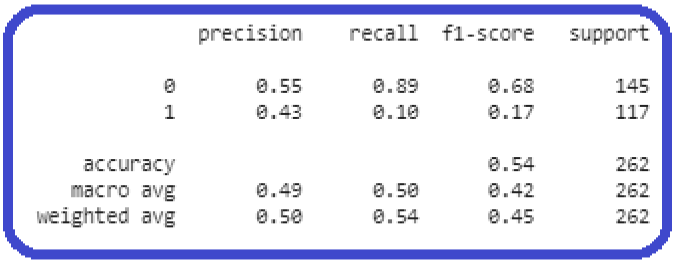
```
              precision    recall  f1-score   support

           0       0.53      0.90      0.67       326
           1       0.43      0.08      0.14       284

    accuracy                           0.52       610
   macro avg       0.48      0.49      0.40       610
weighted avg       0.48      0.52      0.42       610
```

Table 4: Classification report – Logistic Regression Train data

Inference:

- For predicting employees who will not opt for package
  - Precision (53%) – 53% of prediction of employees who will not opt for package are correct
  - Recall (90%) – 90% of of employees who will not opt for package are correctly predicted
- For predicting employees who opt for package
  - Precision (43%) – 43% of prediction of employees who will opt for package are correct
  - Recall (8%) – 8% of of employees who will opt for package are correctly predicted
- Overall accuracy of the model – 52% of total prediction are correct

Classification report on Test data



```
              precision    recall  f1-score   support

           0       0.55      0.89      0.68       145
           1       0.43      0.10      0.17       117

    accuracy                           0.54       262
   macro avg       0.49      0.50      0.42       262
weighted avg       0.50      0.54      0.45       262
```

Table 5: Classification report – Logistic Regression Test data

Inference:

- For predicting employees who will not opt for package
  - Precision (55%) – 55% of prediction of employees who will not opt for package are correct
  - Recall (89%) – 89% of of employees who will not opt for package are correctly predicted
- For predicting employees who opt for package
  - Precision (89%) – 89% of prediction of employees who will opt for package are correct
  - Recall (10%) – 10% of of employees who will opt for package are correctly predicted
- Overall accuracy of the model – 54% of total prediction are correct

4.2 LDA Model with Performance Metrics

Performance metrics

ROC Curve – Train data

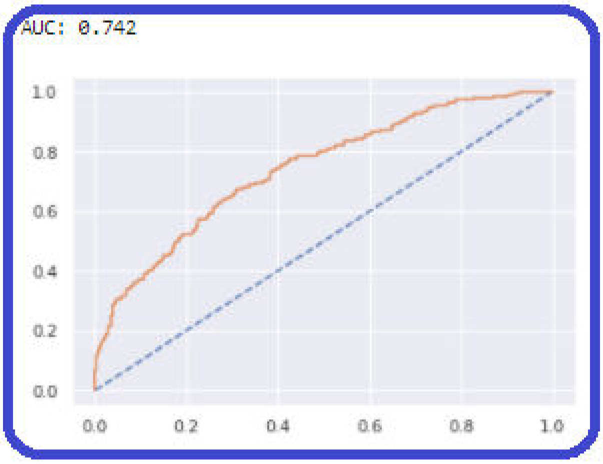


Figure 15: ROC Curve train data– LDA

ROC Curve – Test data

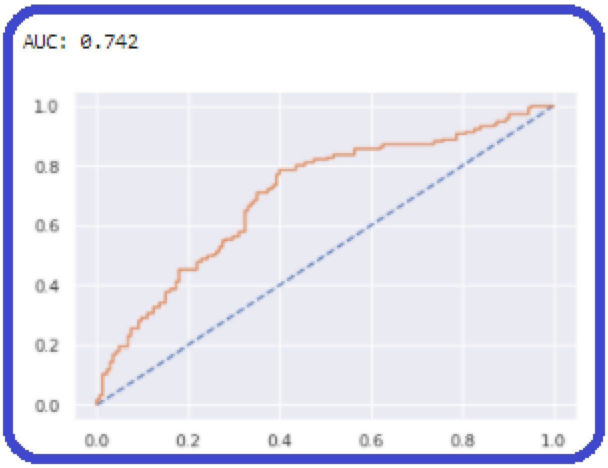
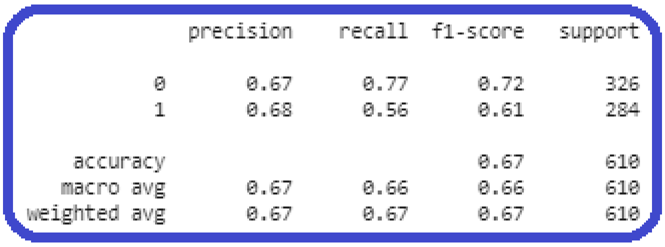


Figure 16: ROC Curve test data– LDA

<u>Classification report on Train data</u>



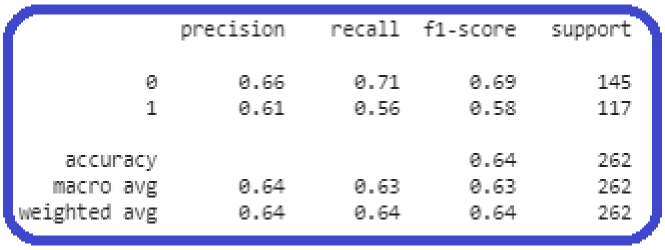
```
              precision    recall  f1-score   support

           0       0.67      0.77      0.72       326
           1       0.68      0.56      0.61       284

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

Table 6: Classification report – LDA Train data

Inference:

- For predicting employees who will not opt for package
  - Precision (67%) – 67% of prediction of employees who will not opt for package are correct
  - Recall (77%) – 77% of of employees who will not opt for package are correctly predicted
- For predicting employees who opt for package
  - Precision (68%) – 68% of prediction of employees who will opt for package are correct
  - Recall (56%) – 56% of of employees who will opt for package are correctly predicted
- Overall accuracy of the model – 67% of total prediction are correct

Classification report on Test data



```
              precision    recall  f1-score   support

           0       0.66      0.71      0.69       145
           1       0.61      0.56      0.58       117

    accuracy                           0.64       262
   macro avg       0.64      0.63      0.63       262
weighted avg       0.64      0.64      0.64       262
```

Table 7: Classification report – LDA Test data

Inference:

- For predicting employees who will not opt for package
  - Precision (66%) – 66% of prediction of employees who will not opt for package are correct
  - Recall (71%) – 71% of of employees who will not opt for package are correctly predicted
- For predicting employees who opt for package
  - Precision (61%) – 61% of prediction of employees who will opt for package are correct
  - Recall (56%) – 56% of of employees who will opt for package are correctly predicted
- Overall accuracy of the model – 64% of total prediction are correct

4.3 Logistic Regression Model using grid search solver with Performance Metrics

Hyperparametrs selected from grid search (code in appendix)

Best Penalty: l1
Best C: 10000

Performance metrics

ROC Curve – Train data

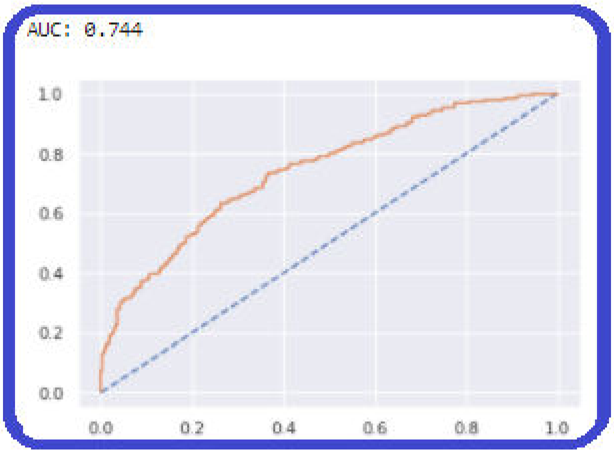


Figure 17: ROC Curve train data – Logistic Regression using grid search

ROC Curve – Test data

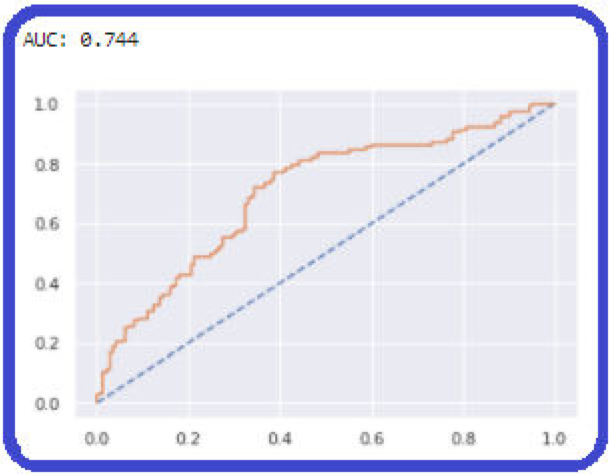
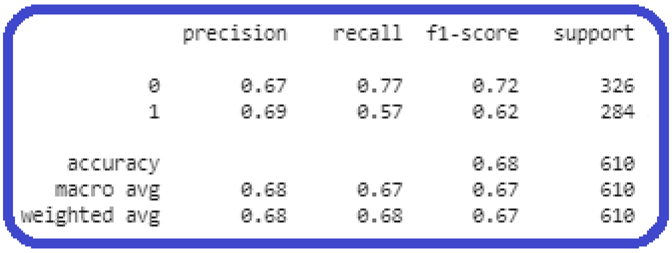


Figure 18: ROC Curve test data – Logistic Regression using grid search

Comment: We can see a vast difference in our earlier logistic regression mode and logistic model after applying grid search. With the new hyperparameters, we have improve AUC from 0.56 to 0.744.

<u>Classification report on Train data</u>



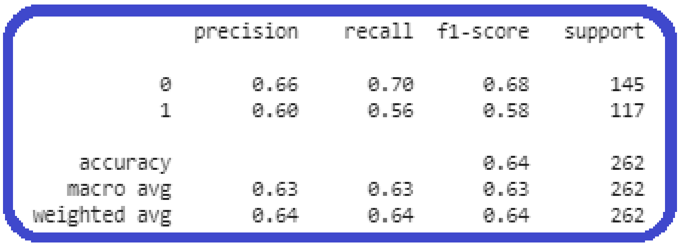
```
              precision    recall  f1-score   support

           0       0.67      0.77      0.72       326
           1       0.69      0.57      0.62       284

    accuracy                           0.68       610
   macro avg       0.68      0.67      0.67       610
weighted avg       0.68      0.68      0.67       610
```

Table 8: Classification report – Logistic Regression using grid search Train data

Inference:

- For predicting employees who will not opt for package
  - Precision (67%) – 67% of prediction of employees who will not opt for package are correct
  - Recall (77%) – 77% of of employees who will not opt for package are correctly predicted
- For predicting employees who opt for package
  - Precision (69%) – 69% of prediction of employees who will opt for package are correct
  - Recall (57%) – 57% of of employees who will opt for package are correctly predicted
- Overall accuracy of the model – 68% of total prediction are correct

Classification report on Test data



```
              precision    recall  f1-score   support

           0       0.66      0.70      0.68       145
           1       0.60      0.56      0.58       117

    accuracy                           0.64       262
   macro avg       0.63      0.63      0.63       262
weighted avg       0.64      0.64      0.64       262
```

Table 9: Classification report – Logistic Regression using grid search Test data

Comment: A huge improvement in recall after applying grid search.

Inference:

- For predicting employees who will not opt for package
  - Precision (66%) – 66% of prediction of employees who will not opt for package are correct
  - Recall (70%) – 70% of of employees who will not opt for package are correctly predicted
- For predicting employees who opt for package
  - Precision (60%) – 60% of prediction of employees who will opt for package are correct
  - Recall (56%) – 56% of of employees who will opt for package are correctly predicted
- Overall accuracy of the model – 64% of total prediction are correct

4.4 LDA Model using grid search solver with Performance Metrics

Performance metrics

ROC Curve – Train data

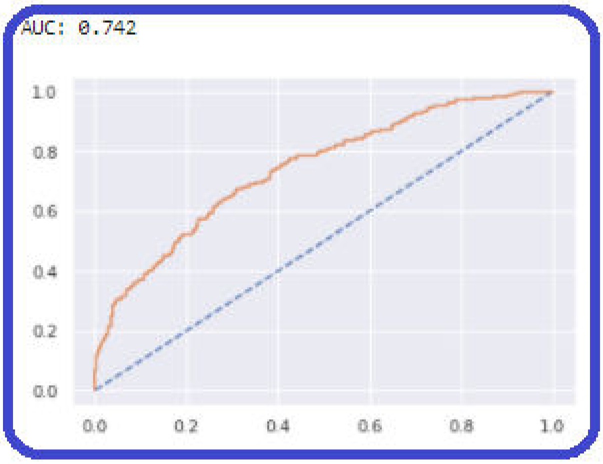


Fig 19: ROC Curve train data – LDA using grid search

ROC Curve – Train data

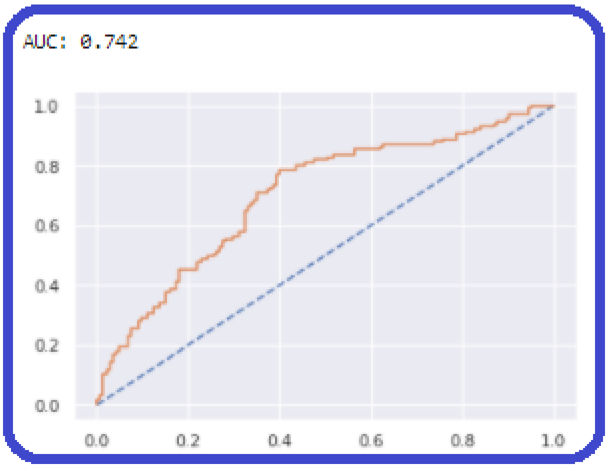
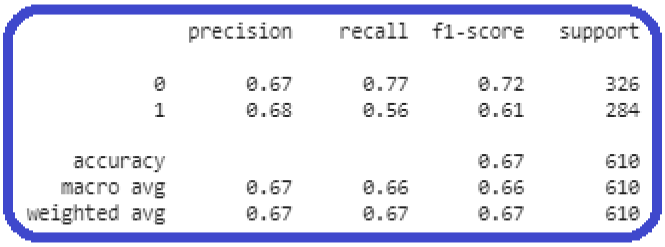


Fig 20: ROC Curve test data – LDA using grid search

Classification report on Train data



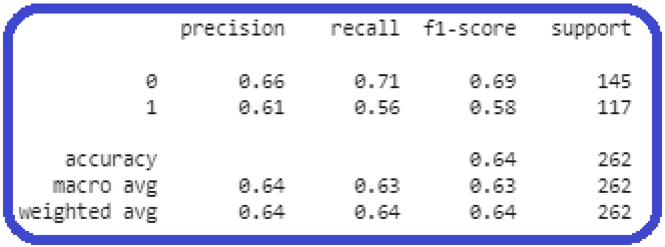
```
              precision    recall  f1-score   support

           0       0.67      0.77      0.72       326
           1       0.68      0.56      0.61       284

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

Table 10: Classification report – LDA using grid search Train data

Inference:

- For predicting employees who will not opt for package
  - Precision (67%) – 67% of prediction of employees who will not opt for package are correct
  - Recall (77%) – 77% of of employees who will not opt for package are correctly predicted
- For predicting employees who opt for package
  - Precision (68%) – 68% of prediction of employees who will opt for package are correct
  - Recall (56%) – 56% of of employees who will opt for package are correctly predicted
- Overall accuracy of the model – 67% of total prediction are correct

Classification report on Test data



```
              precision    recall  f1-score   support

           0       0.66      0.71      0.69       145
           1       0.61      0.56      0.58       117

    accuracy                           0.64       262
   macro avg       0.64      0.63      0.63       262
weighted avg       0.64      0.64      0.64       262
```

Table 11: Classification report – LDA using grid search Test data

Inference:

- For predicting employees who will not opt for package
  - Precision (66%) – 66% of prediction of employees who will not opt for package are correct
  - Recall (71%) – 71% of of employees who will not opt for package are correctly predicted
- For predicting employees who opt for package
  - Precision (61%) – 61% of prediction of employees who will opt for package are correct
  - Recall (56%) – 56% of of employees who will opt for package are correctly predicted
- Overall accuracy of the model – 64% of total prediction are correct

5 MODEL COMPARISON

Comparing Confusion Matrix of all model (train data)

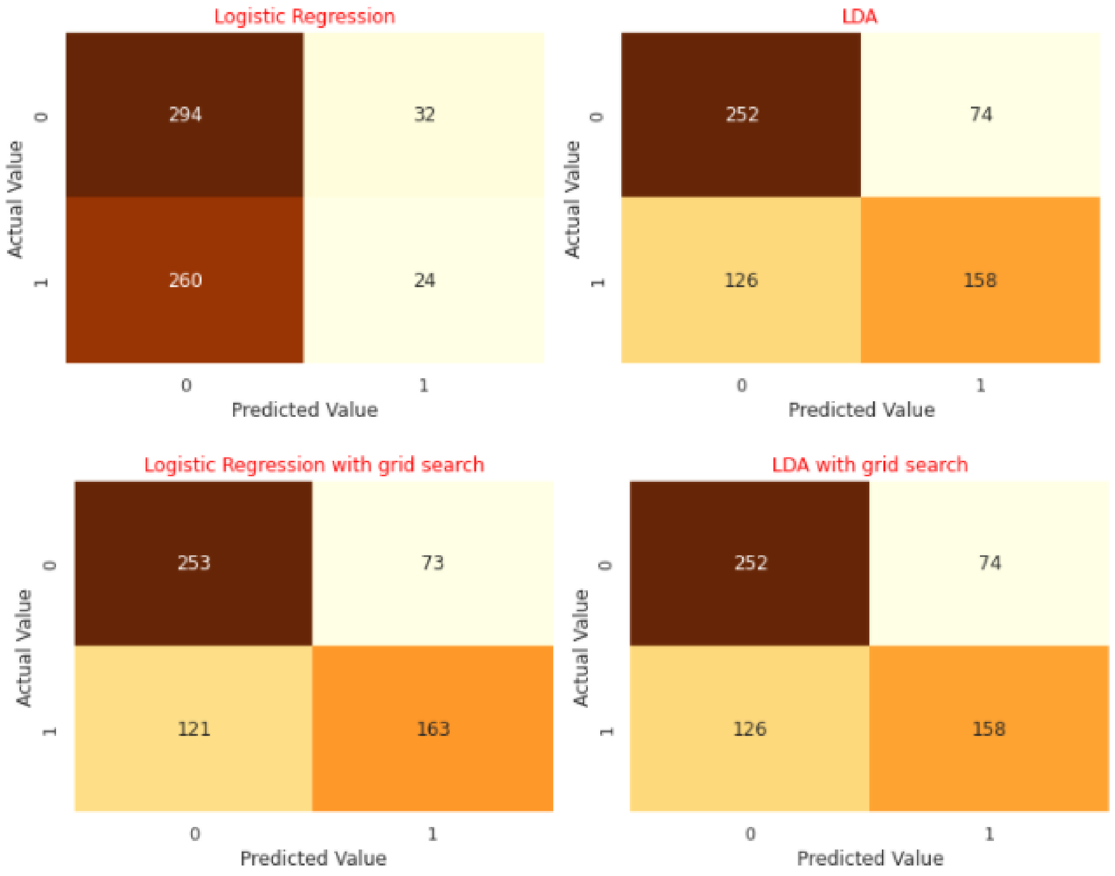
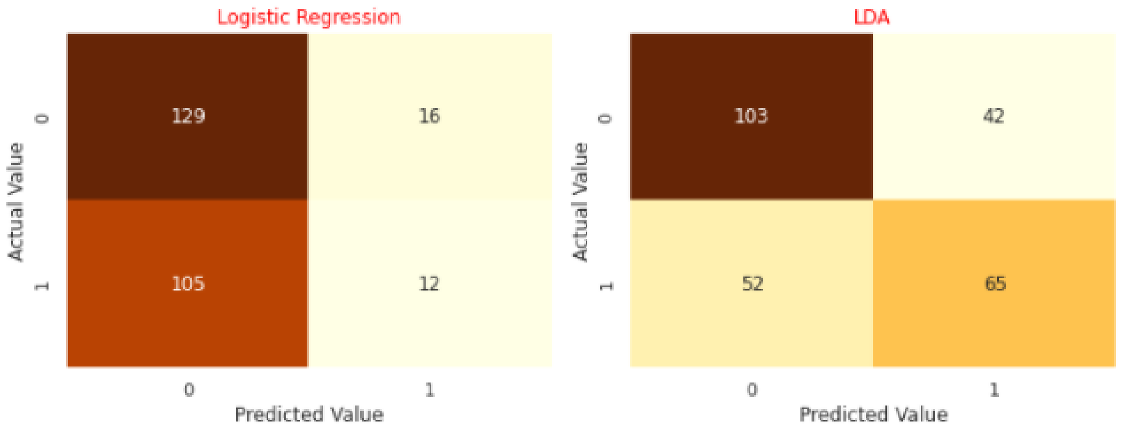


Figure 21: Confusion matrix train data

Comparing Confusion Matrix of all model (test data)



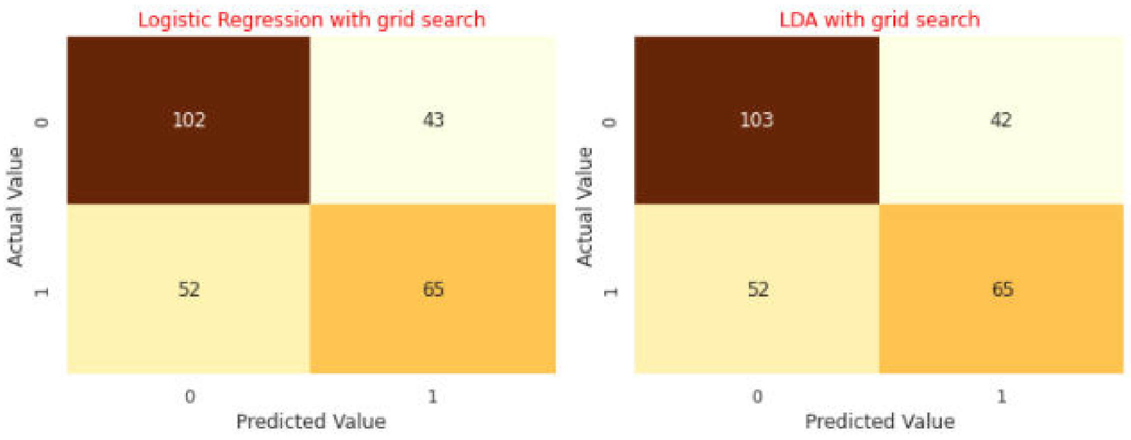


Figure 22: Confusion matrix test data

Model	AUC	Recall		Precision		Accuracy	
		Train	Test	Train	Test	Train	Test
Logistic	0.568	0.08	0.1	0.43	0.43	0.52	0.54
LDA	0.742	0.56	0.56	0.68	0.61	0.67	0.64
Logistic with Grid search	0.744	0.57	0.56	0.69	0.6	0.68	0.64
LDA with Grid search	0.742	0.56	0.56	0.68	0.61	0.67	0.64

Table 12: Model Comparison

If the company is looking for a model, which can help them to target potential customer, and thus helping them to reduce client acquisition cost than model with higher precision makes sense. As in this correctness of predicting employee, going opting for package is more important. Lower the percentage means that the company lost money on the respective lead.

If the company is looking to acquire higher number of customer irrespective of cost of acquisition than recall makes more sense. Here they want to maximise prediction of employee opting for package.

After discussion with the management it was found that company wants to focus on increasing it customer base (expansion mode) and thus the company would be interested in model which will more accurately predicted client who will opt for holiday package. Therefore, as compare to other performance measure recall is most critical.

Taking recall as the criteria, both model, logistics with grid search and LDA are almost giving similar result.

6. CONCLUSION

From Logistic Regression Model (Feature Analysis)

- Age, Education and Number of older children are the continuous variable which have a positive influence of person opting for holiday package
- Employee of higher age have shown more interest in taking the package.
- In the same way, as no of education years increases, probability of an employee going for holiday package increases.
- As number of older children increase the probability of person taking package increase
- If the person is foreigner, the chance of person taking package is higher. In EDA after doing bivariate analysis, we inferred the same.
- As no of young children increases, probability of holiday package decrease.
- Surprisingly Salary are not the criterion for individual to opt for holiday package.

Prediction purpose

As discussed in model comparison section, both model logistic regression with grid search and LDA are doing equally good. However, as logistic regression does not assume any specific shapes of densities in the space of predictor variables and LDA does, we will prefer logistic regression to LDA.

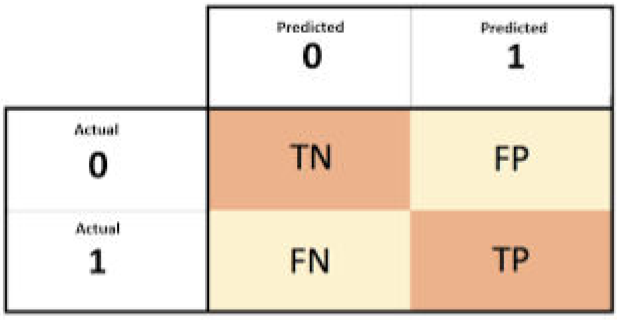
Recommendation

- As the foreigners have shown higher interest towards the package, the company should come up with new packages as per their requirement.
- The company can first target the following two profiles to have a higher conversion rate and thus manage their marketing and operational cost
  - Foreign Employee
  - Employees with high education do not have young children and are in a phase where they do not have to take care of the children (grown children).
- As all the value of all performance parameters is not high, organisation should look at considering some more features to improve the model performance.

Appendix

Confusion Matrix

The confusion matrix is the table that will help us to describe the performance of the classification model on a set of test data and train data for which true values are known. The confusion matrix summarizes the classification of the four groups. The confusion matrix is summarized as follow:



	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

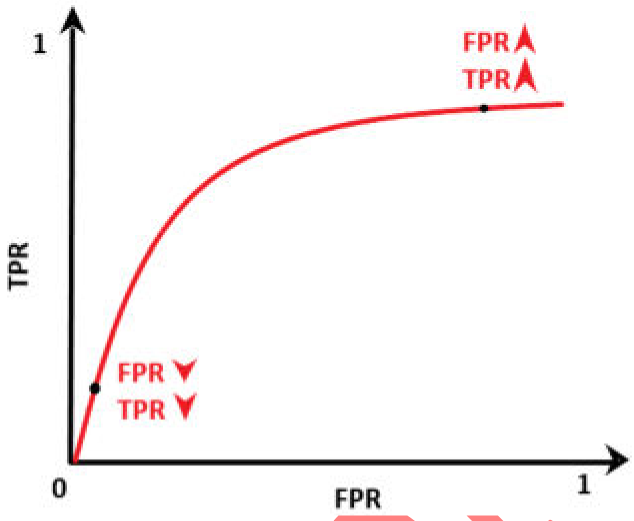
- Using the confusion matrix, we can calculate different metrics.
- Accuracy: this metric gives the fraction of predictions our model got right. Formally, accuracy has the following definition: Accuracy = Number of correct predictions / Total number of predictions.
- Accuracy= (TP+TN)/)/(TP+FP+FN+TN)
- Sensitivity or recall: Known as Fraction of positives that were correctly identified, Recall is the ability of a classifier to find all positive instances. For each class it is defined as the ratio of true positives to the sum of true positives and false negatives. Recall = TP/(TP+FN)
- Precision: it corresponds to the accuracy of positive predictions. It is the ability of a classifier not to label an instance positive that is actually negative. For each class, it is defined as the ratio of true positives to the sum of a true positive and false positive. Precision = TP/(TP + FP)
- F1 Score : The F1 score is defined as the harmonic mean of the model's precision and recall. The use of « Harmonic mean » is because it is not sensitive to extremely large values, unlike simple averages. The formula is as follow :

$$\text{F1 Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

The best score is 1.0 and the worst is 0.0. F1 scores are lower than accuracy measures as they embed precision and recall into their computation. As a rule of thumb, the weighted average of F1 should be used to compare classifier models, not global accuracy.

ROC Curve

The ROC curve is a commonly used graph that summarizes the performance of a classifier over all possible thresholds. It is generated by plotting the True Positive Rate (y-axis) against the False Positive Rate (x-axis) as you vary the threshold for assigning observations to a given class.



The ROC curve shows the trade-off between sensitivity (or TPR) and specificity (1 –FPR). Classifiers that give curves closer to the top-left corner indicate a better performance. As a baseline, a random classifier is expected to give points lying along the diagonal (FPR = TPR). The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

Area Under the Curve (AUC)

To compare different classifiers, it can be useful to summarize the performance of each classifier into a single measure. One common approach is to calculate the area under the ROC curve, which is abbreviated to AUC. It is equivalent to the probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative instance, i.e. it is equivalent to the two-sample Wilcoxon rank-sum statistic.

A classifier with high AUC can occasionally score worse in a specific region than another classifier with lower AUC. However, in practice, the AUC performs well as a general measure of predictive accuracy.