Predictive Modeling

BUSINESS REPORT

Contents

List Of Table

List Of Figure

Problem Statement

Context

The EdTech industry has been surging in the past decade immensely, and according to a forecast, the Online Education market would be worth $286.62bn by 2023 with a compound annual growth rate (CAGR) of 10.26% from 2018 to 2023. The modern era of online education has enforced a lot in its growth and expansion beyond any limit. Due to having many dominant features like ease of information sharing, personalized learning experience, transparency of assessment, etc, it is now preferable to traditional education.
In the present scenario due to the Covid-19, the online education sector has witnessed rapid growth and is attracting a lot of new customers. Due to this rapid growth, many new companies have emerged in this industry. With the availability and ease of use of digital marketing resources, companies can reach out to a wider audience with their offerings. The customers who show interest in these offerings are termed as leads. There are various sources of obtaining leads for Edtech companies, like

- The customer interacts with the marketing front on social media or other online platforms.
- The customer browses the website/app and downloads the brochure
- The customer connects through emails for more information.

The company then nurtures these leads and tries to convert them to paid customers. For this, the representative from the organization connects with the lead on call or through email to share further details.

Objective

ExtraaLearn is an initial stage startup that offers programs on cutting-edge technologies to students and professionals to help them upskill/reskill. With a large number of leads being generated on a regular basis, one of the issues faced by ExtraaLearn is to identify which of the leads are more likely to convert so that they can allocate resources accordingly. You, as a data scientist at ExtraaLearn, have been provided the leads data to:

- Analyze and build an ML model to help identify which leads are more likely to convert to paid customers.
- Find the factors driving the lead conversion process.
- Create a profile of the leads which are likely to convert.

Data Description

The data contains the different attributes of leads and their interaction details with ExtraaLearn. The detailed data dictionary is given below.

Data Dictionary

- ID: ID of the lead
- age: Age of the lead
- current_occupation: Current occupation of the lead. Values include 'Professional','Unemployed',and 'Student'
- first_interaction: How did the lead first interacted with ExtraaLearn. Values include 'Website', 'Mobile App'
- profile_completed: What percentage of profile has been filled by the lead on the website/mobile app. Values include Low - (0-50%), Medium - (50-75%), High (75-100%)
- website_visits: How many times has a lead visited the website
- time_spent_on_website: Total time spent on the website in seconds
- page_views_per_visit: Average number of pages on the website viewed during the visits.
- last_activity: Last interaction between the lead and ExtraaLearn.
    - Email Activity: Seeking for details about the program through email, Representative shared information with the lead like brochure of program, etc
    - Phone Activity: Had a Phone Conversation with the representative, Had conversation over SMS with the representative, etc
    - Website Activity: Interacted on live chat with a representative, Updated profile on the website, etc
- print_media_type1: Flag indicating whether the lead had seen the ad of ExtraaLearn in the Newspaper.
- print_media_type2: Flag indicating whether the lead had seen the ad of ExtraaLearn in the Magazine.
- digital_media: Flag indicating whether the lead had seen the ad of ExtraaLearn on the digital platforms.
- educational_channels: Flag indicating whether the lead had heard about ExtraaLearn in the education channels like online forums, discussion threads, educational websites, etc.
- referral: Flag indicating whether the lead had heard about ExtraaLearn through reference.
- status: Flag indicating whether the lead was converted to a paid customer or not.

1- Define the problem and perform Exploratory Data Analysis

>Shape of the dataset:
There are 4612 rows and 15 columns present in the dataset.

>View first & last 5 rows of the Dataset:

	ID	age	current_occupation	first_interaction	profile_completed	website_visits	time_spent_on_website	page_views_per_visit	last_activity	print_media_type1	print_media_type2	digital_media	educational_channels	referral	status
0	EXT001	57	Unemployed	Website	High	7	1639	1.86100	Website Activity	Yes	No	Yes	No	No	1
1	EXT002	56	Professional	Mobile App	Medium	2	83	0.32000	Website Activity	No	No	No	Yes	No	0
2	EXT003	52	Professional	Website	Medium	3	330	0.07400	Website Activity	No	No	Yes	No	No	0
3	EXT004	53	Unemployed	Website	High	4	464	2.05700	Website Activity	No	No	No	No	No	1
4	EXT005	23	Student	Website	High	4	600	16.91400	Email Activity	No	No	No	No	No	0

Table 1- Top 5 rows of the dataset

	ID	age	current_occupation	first_interaction	profile_completed	website_visits	time_spent_on_website	page_views_per_visit	last_activity	print_me
4607	EXT4608	35	Unemployed	Mobile App	Medium	15	360	2.17000	Phone Activity	
4608	EXT4609	55	Professional	Mobile App	Medium	8	2327	5.39300	Email Activity	
4609	EXT4610	58	Professional	Website	High	2	212	2.69200	Email Activity	
4610	EXT4611	57	Professional	Mobile App	Medium	1	154	3.87900	Website Activity	
4611	EXT4612	55	Professional	Website	Medium	4	2290	2.07500	Phone Activity	

Table 2- Last 5 rows of the dataset

>Data types of the Dataset:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4612 entries, 0 to 4611
Data columns (total 15 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   ID                     4612 non-null   object 
 1   age                    4612 non-null   int64  
 2   current_occupation     4612 non-null   object 
 3   first_interaction      4612 non-null   object 
 4   profile_completed      4612 non-null   object 
 5   website_visits         4612 non-null   int64  
 6   time_spent_on_website  4612 non-null   int64  
 7   page_views_per_visit   4612 non-null   float64
 8   last_activity          4612 non-null   object 
 9   print_media_type1      4612 non-null   object 
 10  print_media_type2      4612 non-null   object 
 11  digital_media          4612 non-null   object 
 12  educational_channels   4612 non-null   object 
 13  referral               4612 non-null   object 
 14  status                 4612 non-null   int64  
dtypes: float64(1), int64(4), object(10)
memory usage: 540.6+ KB
```

- Age, website_visits, time_spent_on_website, page_views_per_visit, and status are of numeric type while rest of the columns are of object type.
- There is no null value are present in the dataset.

>Checking duplicate Value:
There is no duplicate value present in the dataset.

Exploratory Data Analysis(EDA)-
>Statistical summary of the Dataset:

	count	mean	std	min	25%	50%	75%	max
age	4612.00000	46.20121	13.16145	18.00000	36.00000	51.00000	57.00000	63.00000
website_visits	4612.00000	3.56678	2.82913	0.00000	2.00000	3.00000	5.00000	30.00000
time_spent_on_website	4612.00000	724.01127	743.82868	0.00000	148.75000	376.00000	1336.75000	2537.00000
page_views_per_visit	4612.00000	3.02613	1.96812	0.00000	2.07775	2.79200	3.75625	18.43400
status	4612.00000	0.29857	0.45768	0.00000	0.00000	0.00000	1.00000	1.00000

Tabale 3- Statistical summary of the Dataset

Observations of statistical summary;
- The average age of a lead is 46, with a minimum of 18 and a maximum of 63 years old. At least 75% of the leads are 57 years of age, which means most of the leads are older adults.
- The maximum value for website visits are 30 times. This is a big difference in the 75th percentile of 5 times. This might indicate outliers present. The minimum of 0 is interesting.
- The average time spent on the website is 724 seconds. However, at least 75% are spending almost twice that time with 1336 seconds. Which means the minimum value of 0 might be impacting the mean and we should look at those values more closely.
- Most of the leads visit at least 3 or more pages on the website. Although, there is a big difference from the 75 percentile and the maximum value of 18 pages visited. This might suggest outliers.
- Status is either 1 or 0 depending on whether they became a paid customer or not. The mean of approx. 30 is higher than the 50 percentile which means the data is skewed slightly to the right. The std of about 0.46 means the data is dispersed.

>Checking Unique value of all categorical variables:
- Most of the leads are working professionals.
- Almost an equal percentage of profile completions are categorized as high and medium that is 49.1% and 48.6%, respectively. Only 2.3% of the profile completions are categorized as low.
- Approx 49.4% of the leads had their last activity over email, followed by 26.8% having phone activity. This implies that the majority of the leads prefer to communicate via email.
- We see that each ID has an equal percentage of values. Let's check the number of unique values in the ID column.

>Checking Unique value of ID column:
- Total number of unique value is 4612 so, we simply say that all values of the ID column are unique.
- We can drop this column as it was not adding any value to our analysis.

	age	current_occupation	first_interaction	profile_completed	website_visits	time_spent_on_website	page_views_per_visit	last_activity	print_media_type1	pri
0	57	Unemployed	Website	High	7	1639	1.86100	Website Activity	Yes	
1	56	Professional	Mobile App	Medium	2	83	0.32000	Website Activity	No	
2	52	Professional	Website	Medium	3	330	0.07400	Website Activity	No	
3	53	Unemployed	Website	High	4	464	2.05700	Website Activity	No	
4	23	Student	Website	High	4	600	16.91400	Email Activity	No	

Table 4- Dataset after dropping the ID column

Univariate Analysis-

>Observations on Age:

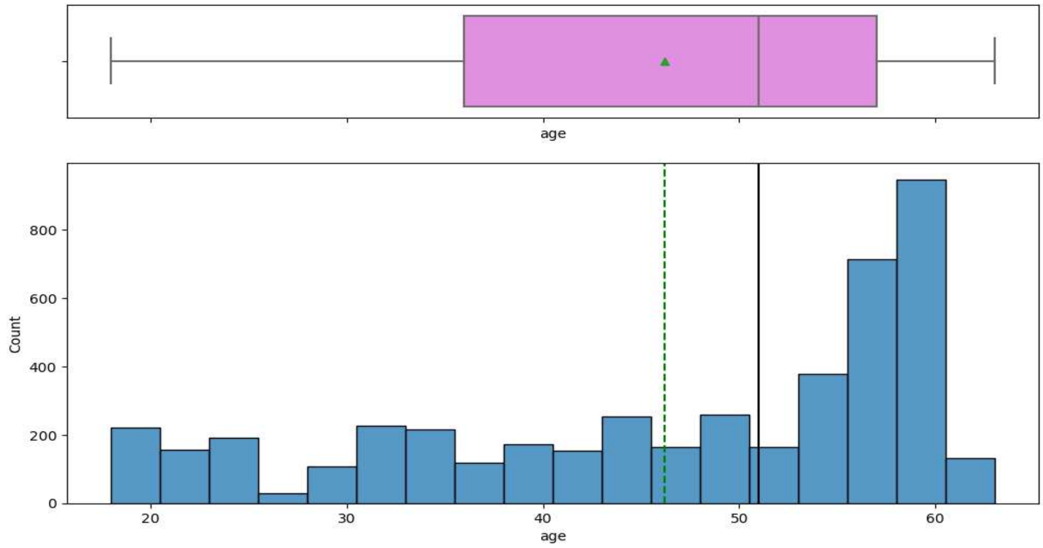

Figure 1

>Observations on Website_visits:

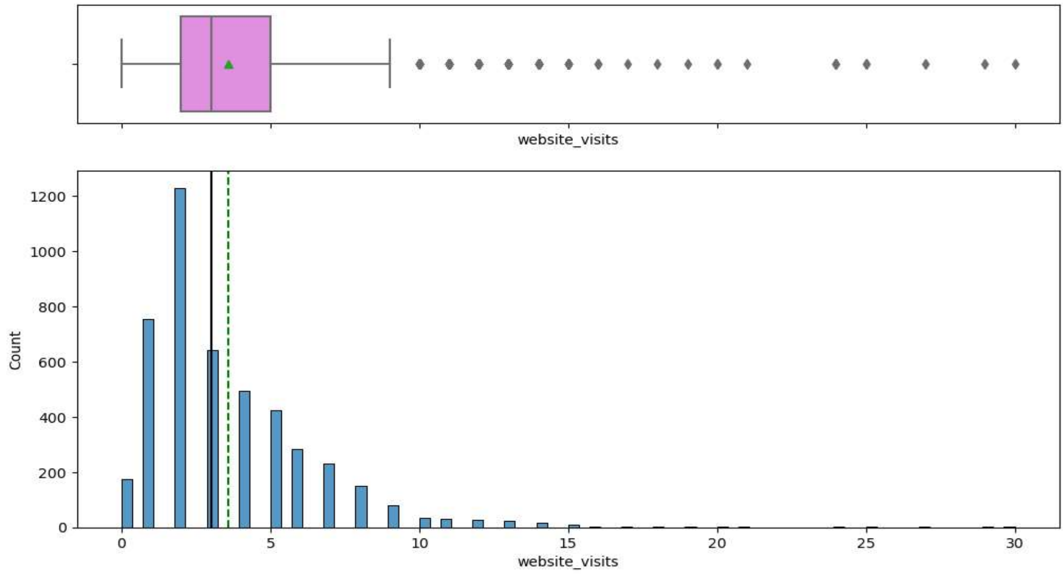

Figure 2

>Observations on number of Time_spent_on_website:

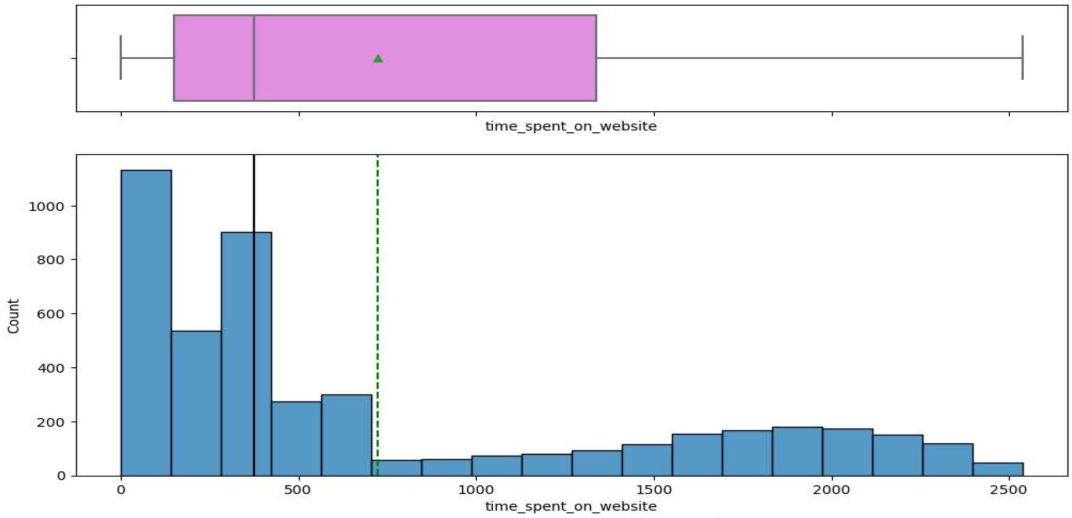

Figure 3

>Observations on number of page_views_per_visit:

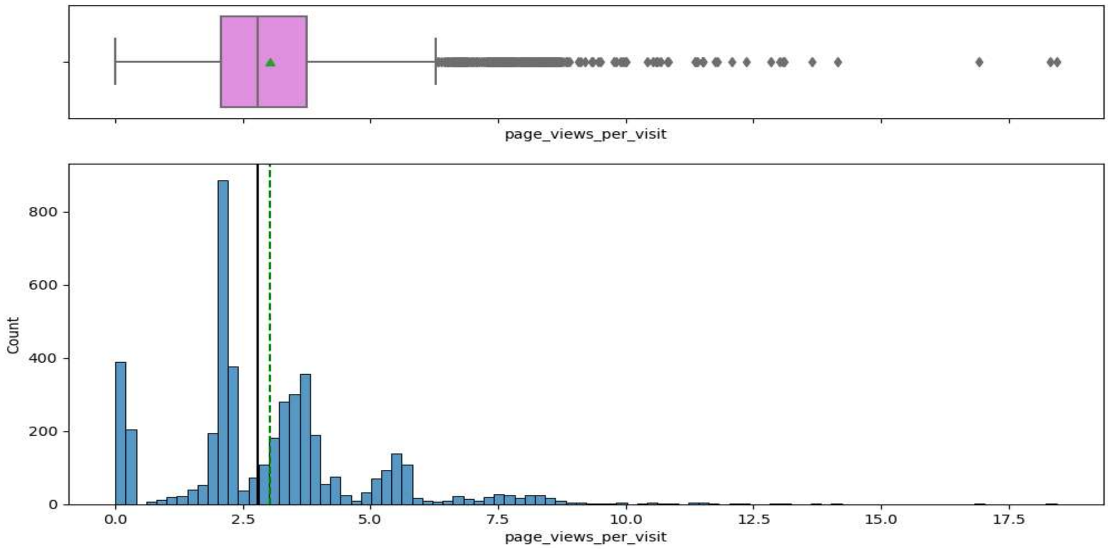

Figure 4

Overall observations from the above plots;

- The distribution of age is left-skewed which shows the majority of leads are 55 - 65 years old.
- Website visits is right-skewed which shows the majority of visits range from 0 to approximately 7 times. There are some outliers. Which means that some leads visited the website from 10 to even 30 times.
- Time spent on the website is right-skewed which means that most of the leads spent less than 700 seconds (~12 min) on the website.
- Page views per visit distribution was approximately normal. Most leads visited 2.5 to 3.5 pages. However, there were many outliers that visited from 7.5 to more than 17.5 pages.

>Observations on the current occupation:

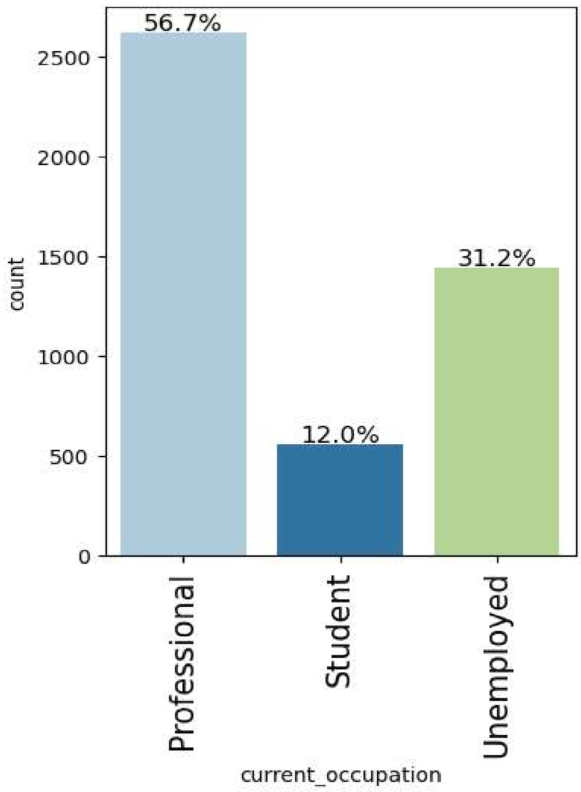

Figure 5

>Observations on number of first_interaction:

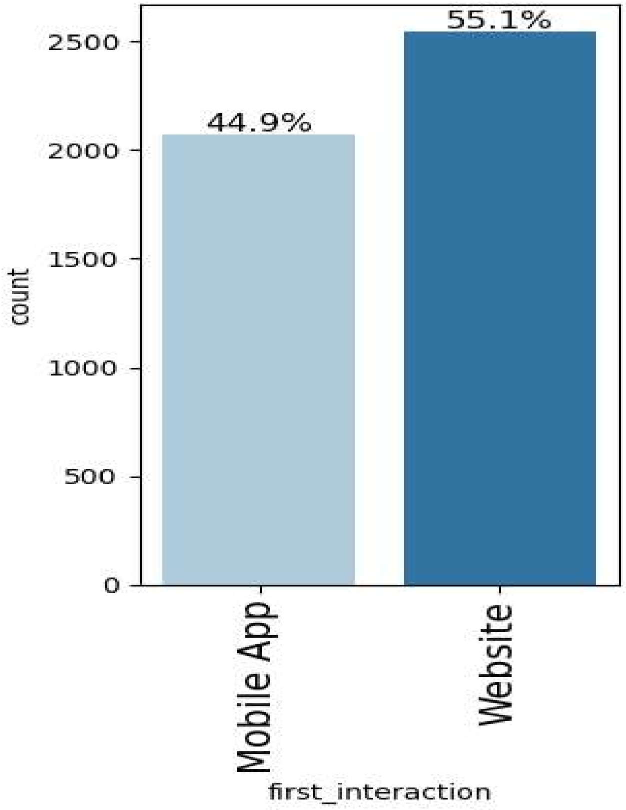

Figure 6

>Observations on profile_completed:

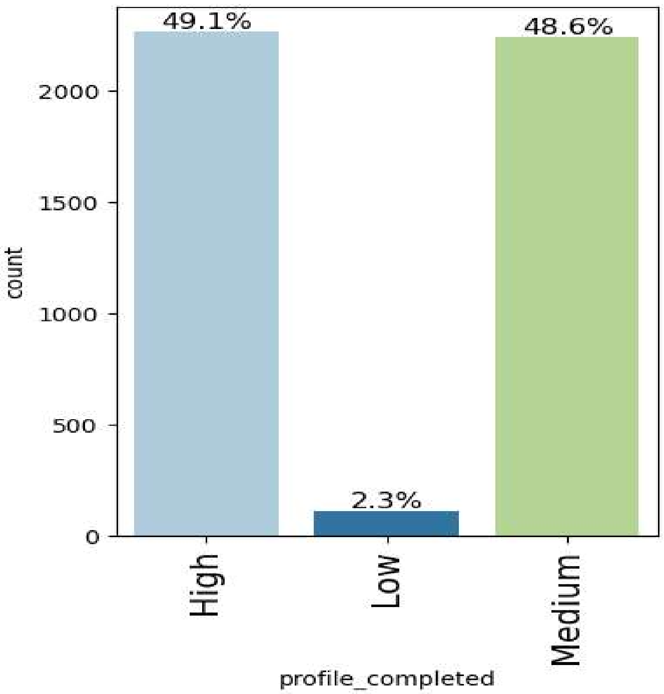

Figure 7

>**Observations on last_activity:**

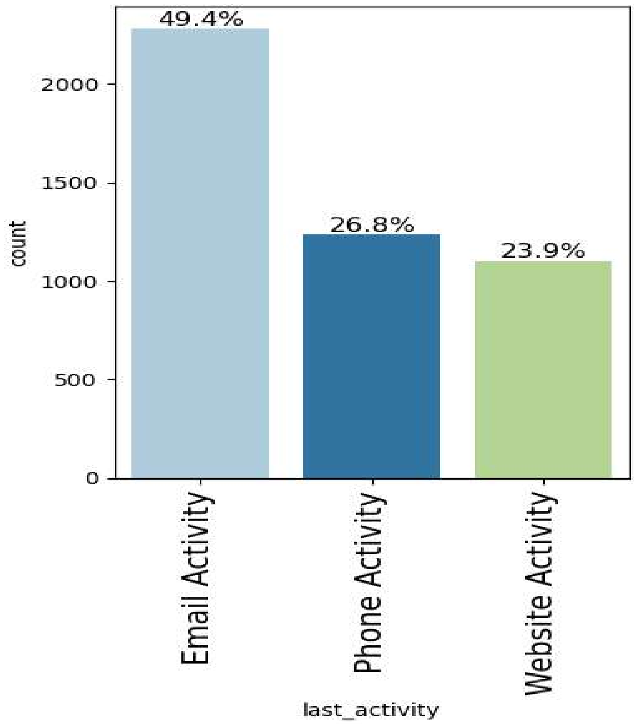

Figure 8

>**Observations on print_media_type1:**

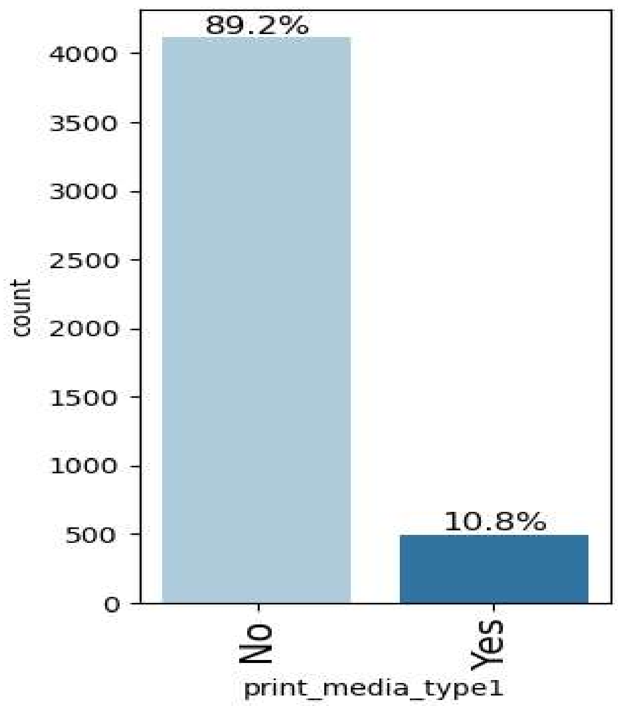

Figure 9

>Observations on print_media_type2:

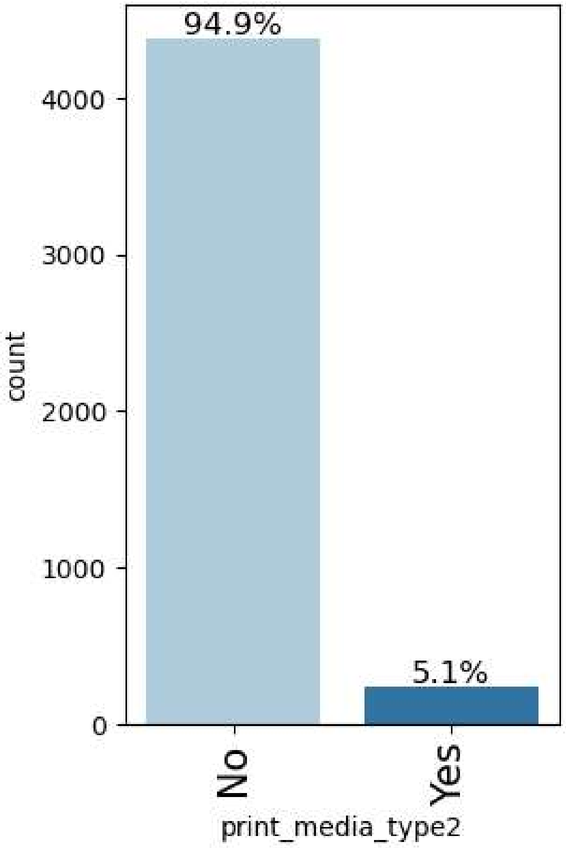

Figure 10

>Observations on digital_media:

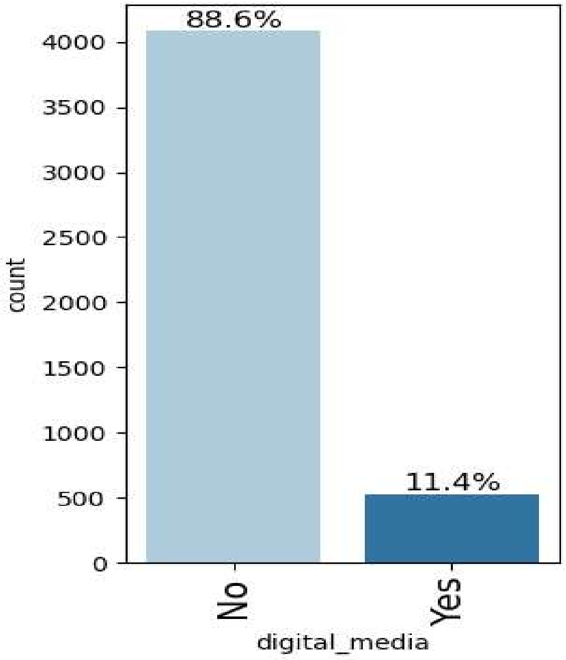

Figure 11

>Observations on educational_channels:

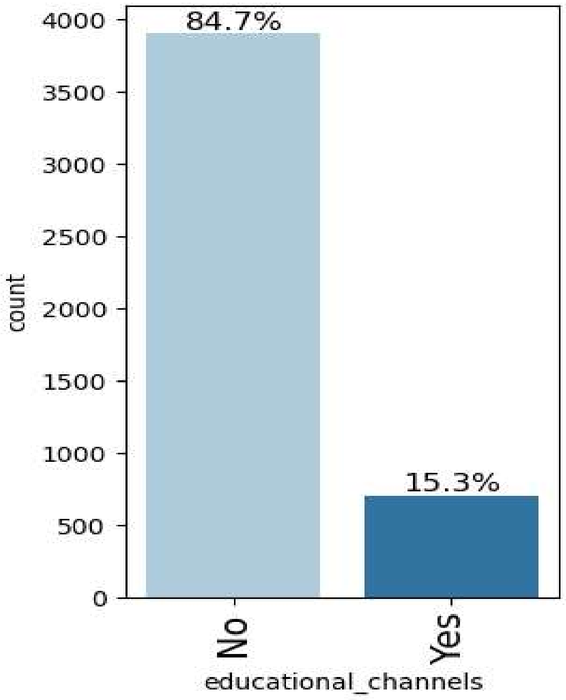

Figure 12

>Observations on referral:

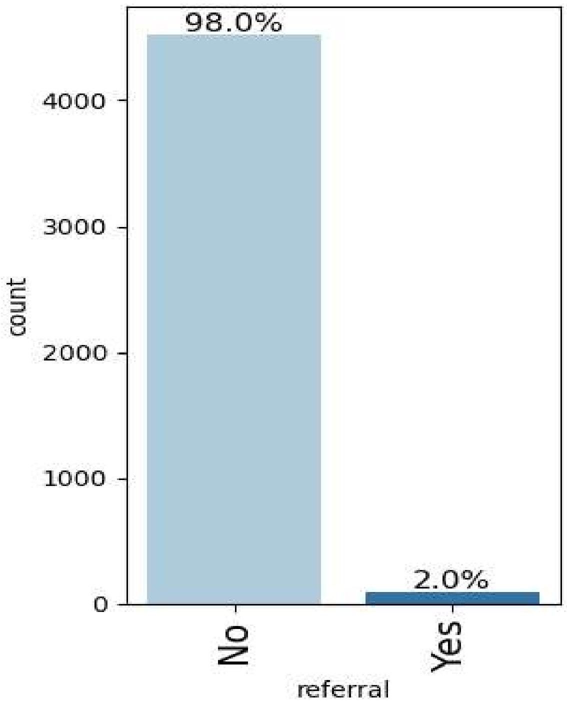

Figure 13

>Observations on status:

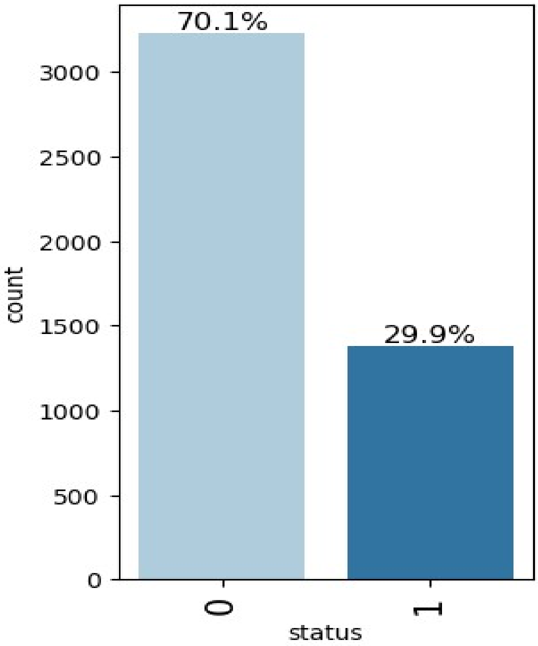

Figure 14

Overall Observations of the above plots;

- The plot shows that working professional leads are more likely to opt for a course offered by the organization and the students are least likely to be converted.
- Majority of the leads around 55.1% who interacted through websites were converted to paid customers, while only a small number around 44.9% of leads, who interacted through mobile app, converted.
- The leads whose profile completion level is high converted more in comparison to other levels of profile completion like 49.1%.
- The low levels of profile completion saw comparatively very less conversions like 2.3%.
- The last activity ended by mostly email about 49.4% as compared to phone and website.
- Print media type1 is more popular- 10.8% or digital media- 11.4% as compared to print media type2 which is relatively less likely 5.1%.
- There are a very few referrals likely 2%.
- Company should try to get more leads through referrals by promoting rewards for existing customer base when they refer someone.
-

Bivariate Analysis-

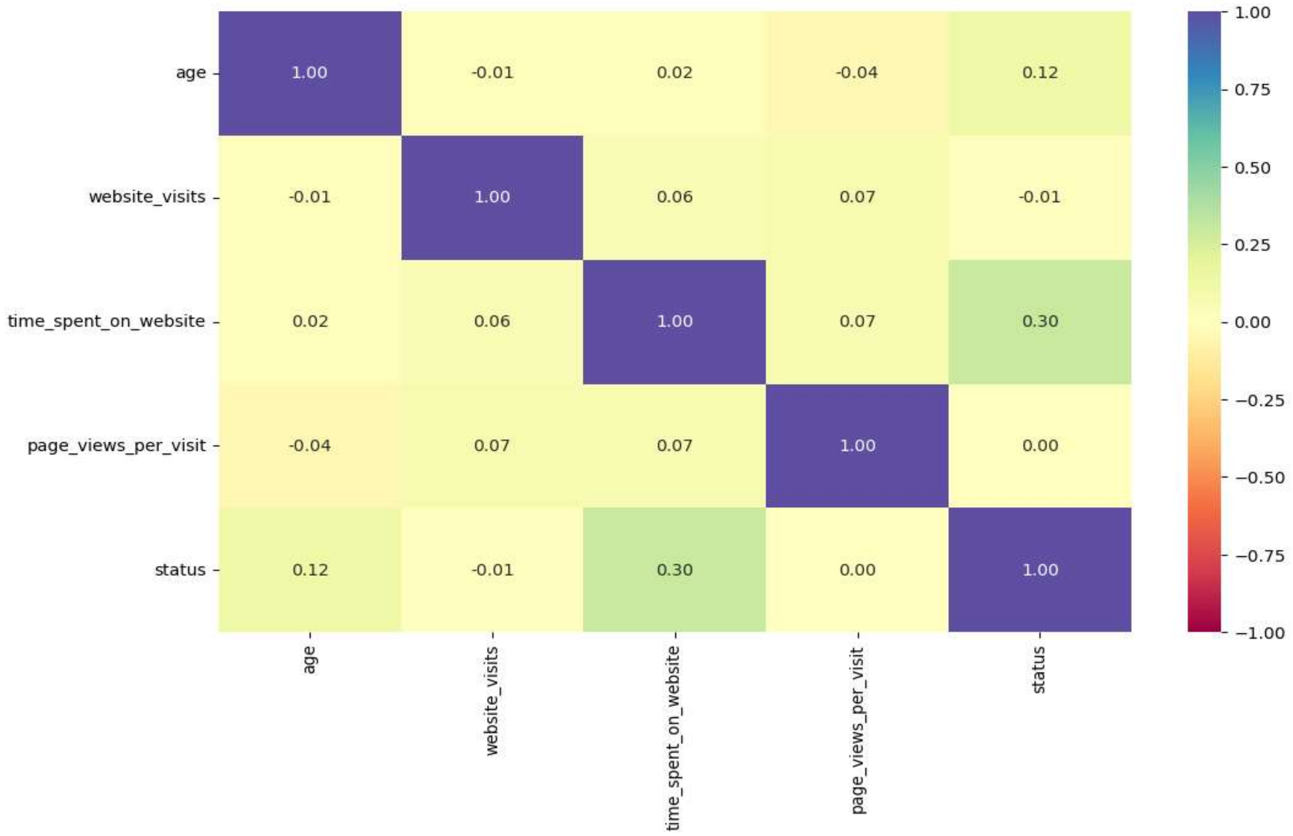

Figure 15

<u>Observation;</u>

- There is a weak positive correlation between time spent on website and status. Which indicates a liklihood that the longer a lead stays on the website the better the chance of converting them to a paid customer.
- There are no other correlations.

Leads will have different expectations from the outcome of the course and the current occupation may play a key role for them to take the program. Let's analyze it

status	0	1	All
current_occupation			
All	3235	1377	4612
Professional	1687	929	2616
Unemployed	1058	383	1441
Student	490	65	555

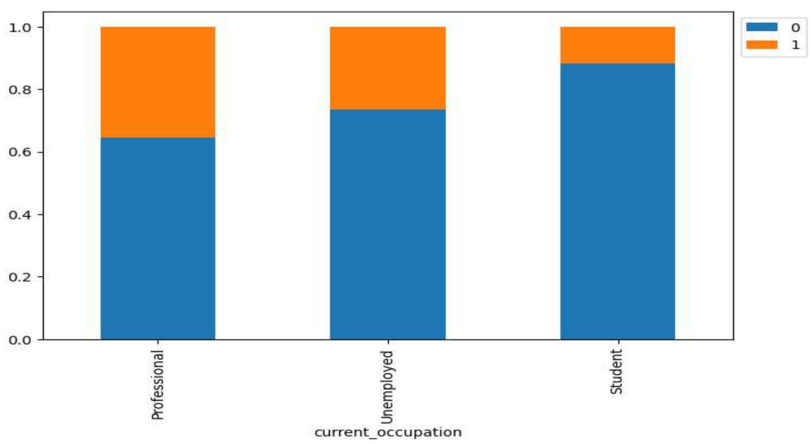

Figure 16

Age can be a good factor to differentiate between such leads

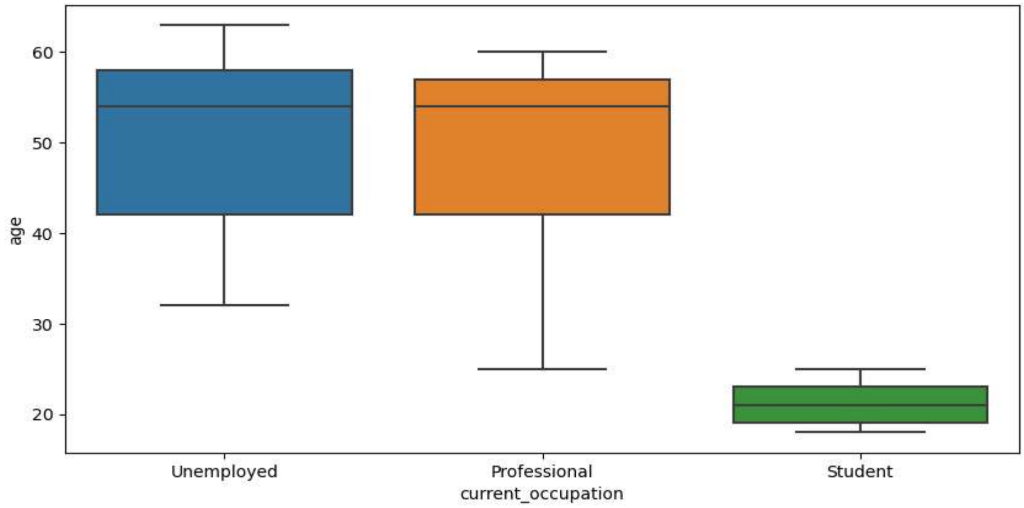

Figure 17

current_occupation	count	mean	std	min	25%	50%	75%	max
Professional	2616.00000	49.34748	9.89074	25.00000	42.00000	54.00000	57.00000	60.00000
Student	555.00000	21.14414	2.00111	18.00000	19.00000	21.00000	23.00000	25.00000
Unemployed	1441.00000	50.14018	9.99950	32.00000	42.00000	54.00000	58.00000	63.00000

Table 5

The company's first interaction with leads should be compelling and persuasive. Let's see if the channels of the first interaction have an impact on the conversion of leads

```
status               0     1   All
first_interaction
All               3235  1377  4612
Website           1383  1159  2542
Mobile App        1852   218  2070
```

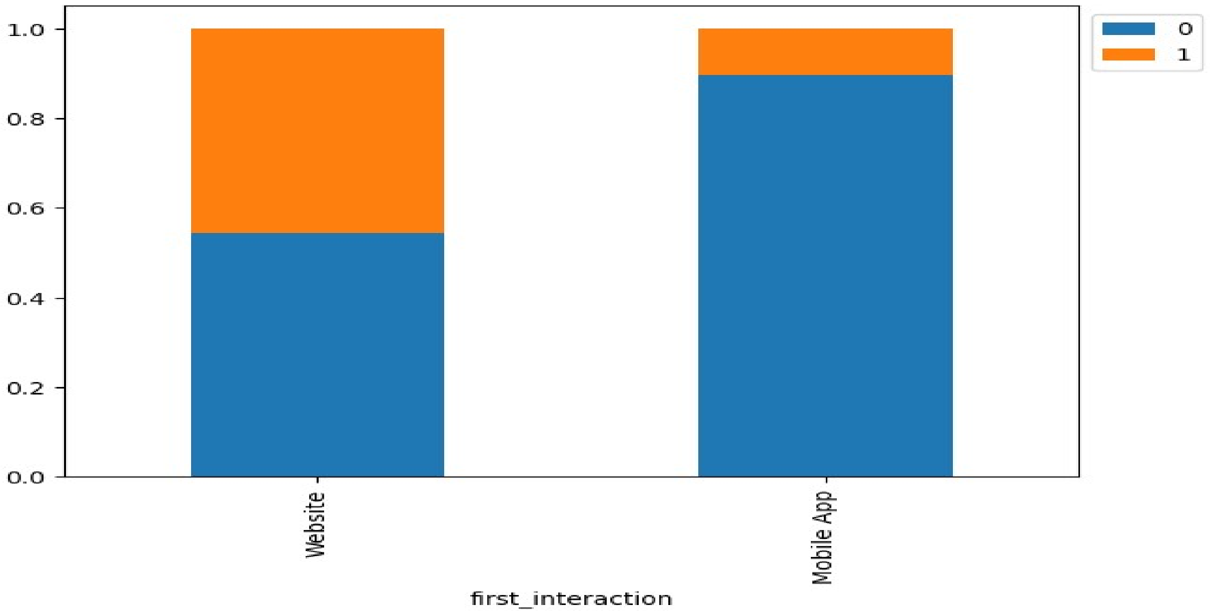

Figure 18

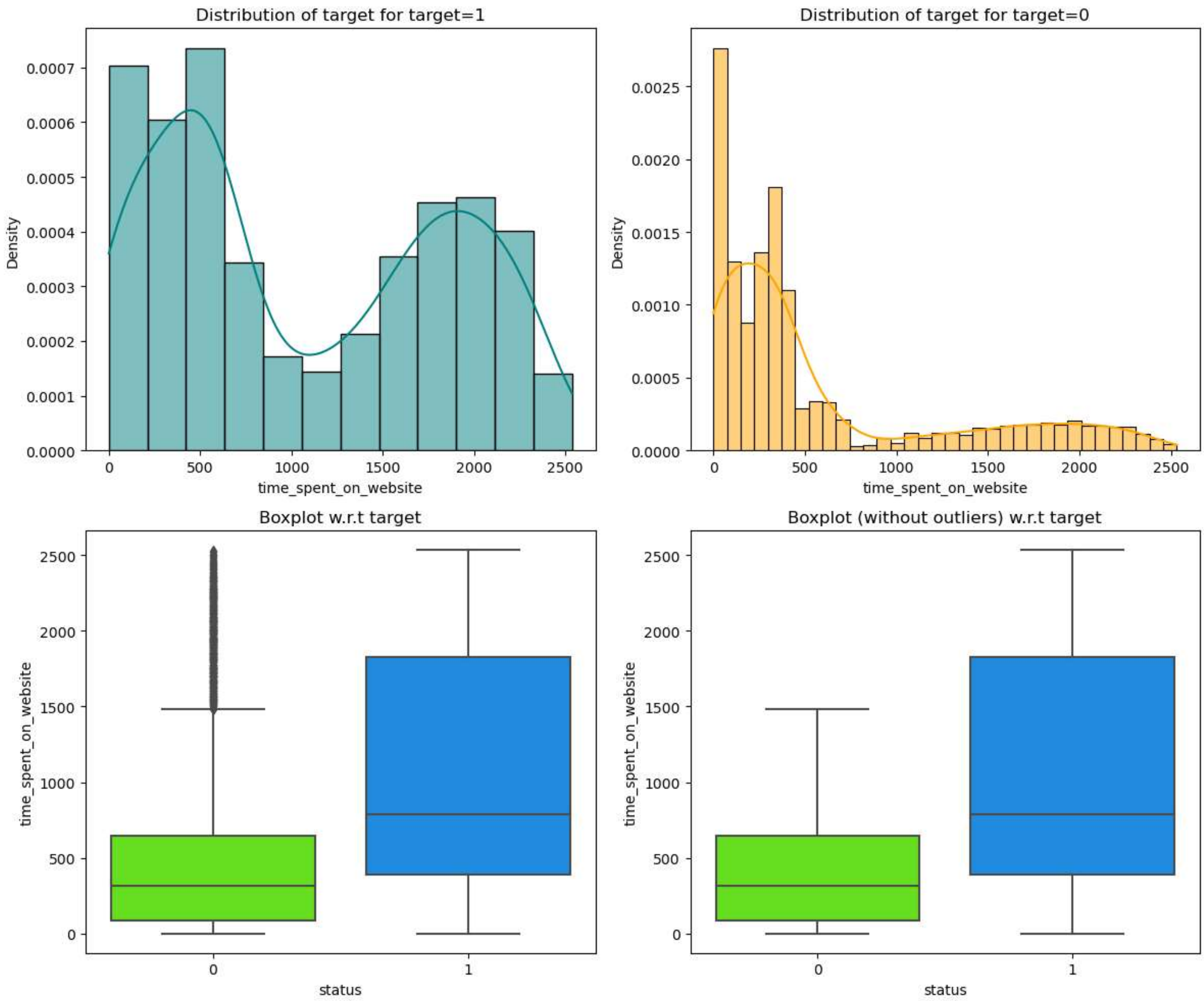

Figure 19

<u>Observations:</u> From the above plots, we can observe that customers who made a purchase tend to spend more time on the website compared to those who didn't. This is so much valuable for understanding customer behavior and potentially optimizing the website to increase conversions.

>Checking the median value;

```
status
0    317.00000
1    789.00000
Name: time_spent_on_website, dtype: float64
```

Let's do a similar analysis for time spent on website and page views per visit.

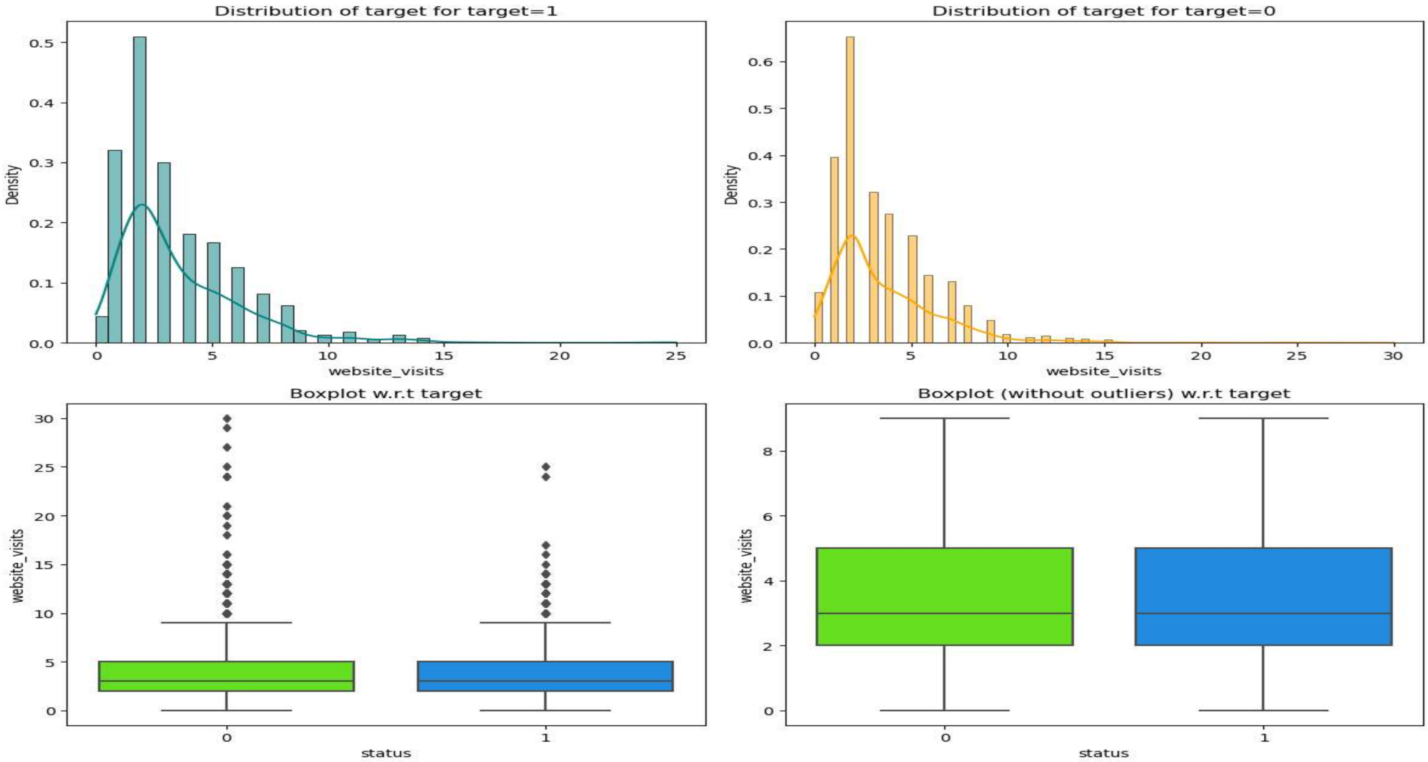

Figure 20

Observation; From analyzing these above plots we can discover that customers who made a purchase tend to have a higher number of website visits compared to those who didn't. This is so much valuable for understanding user behavior and refining marketing strategies to attract and retain potential customers.

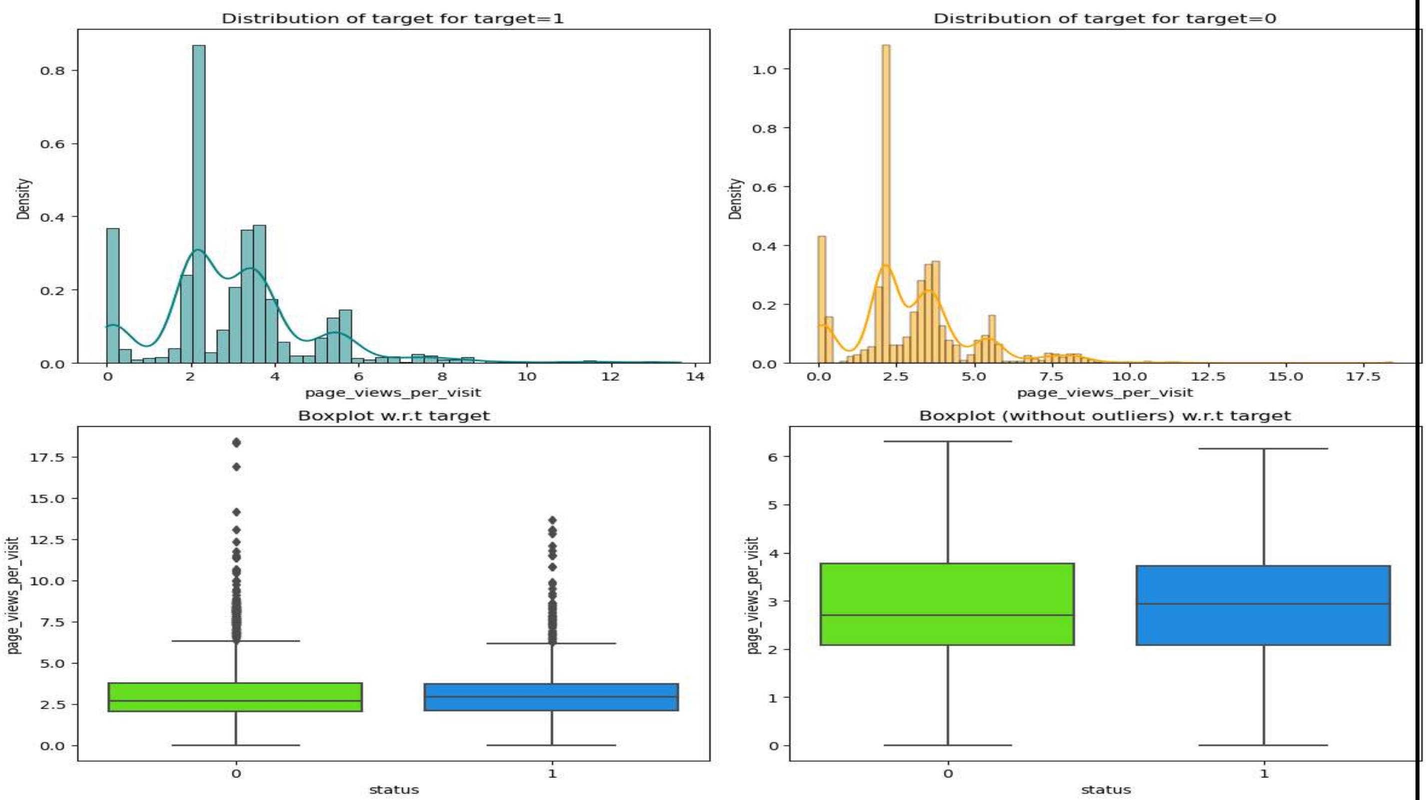

Figure 21

Observation; From the above plots we can observe that customers who made a purchase tend to view more pages per visit compared to those who didn't. This is very useful for understanding user engagement and optimizing the website to encourage more page views, which might lead to increased conversions.

People browsing the website or the mobile app are generally required to create a profile by sharing their personal details before they can access more information. Let's see if the profile completion level has an impact on lead status

```
status                 0     1   All
profile_completed
All                 3235  1377  4612
High                1318   946  2264
Medium              1818   423  2241
Low                   99     8   107
```

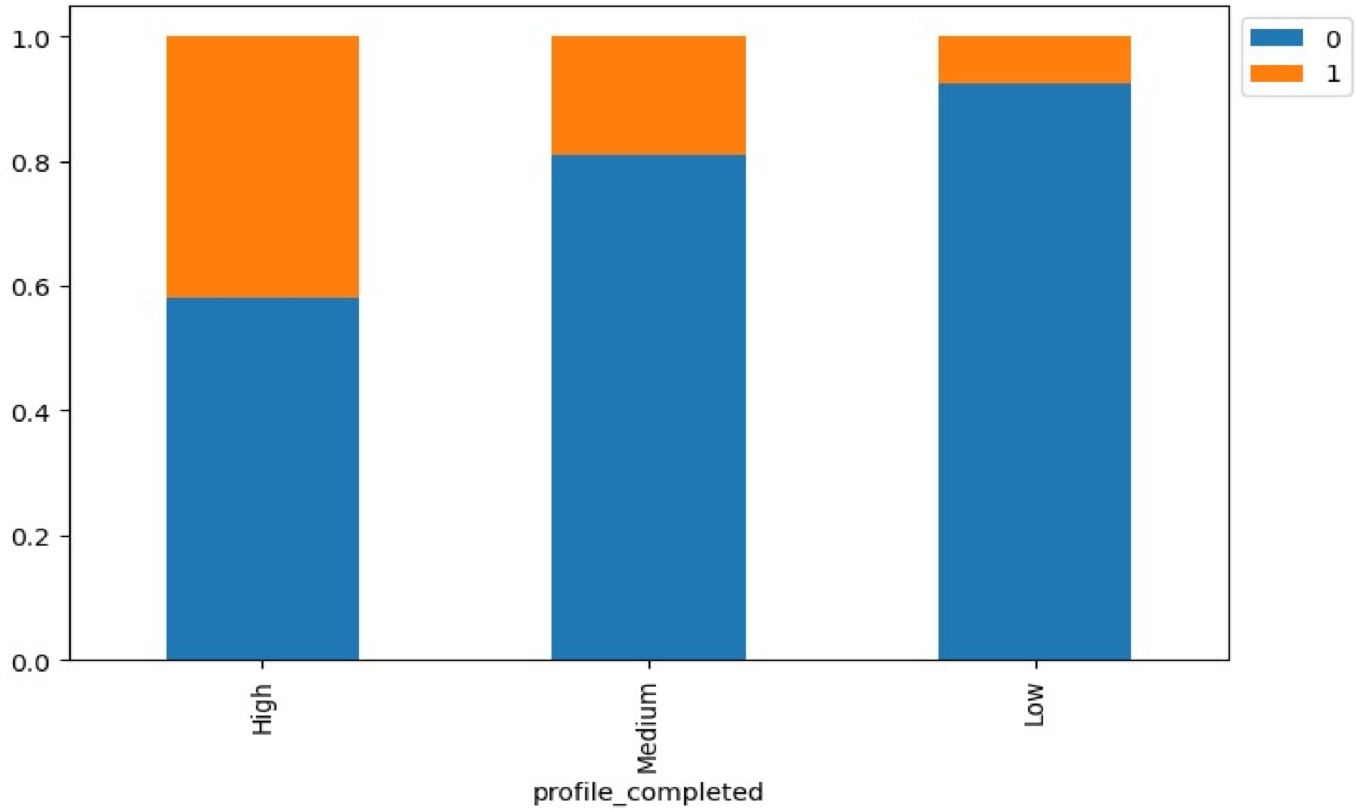

Figure 22

Observations;

- In the "High" profile completion category, there are 1318 observations with status 0 and 946 observations with status 1, totaling 2264 observations.
- In the "Medium" profile completion category, there are 1818 observations with status 0 and 423 observations with status 1, totaling 2241 observations.
- In the "Low" profile completion category, there are 99 observations with status 0 and 8 observations with status 1, totaling 107 observations.
- Overall, across all profile completion levels, there are 3235 observations with status 0 and 1377 observations with status 1, totaling 4612 observations.
- From the above stacked bar plot we see that a higher proportion of customers who completed their profiles made a purchase compared to those who didn't complete their profiles. This is so much valuable for understanding the impact of profile completion on conversion rates and informing strategies to encourage more users to complete their profiles.

After a lead shares their information by creating a profile, there may be interactions between the lead and the company to proceed with the process of enrollment. Let's see how the last activity impacts lead conversion status

```
status                0     1   All
last_activity
All                3235  1377  4612
Email Activity     1587   691  2278
Website Activity    677   423  1100
Phone Activity      971   263  1234
```

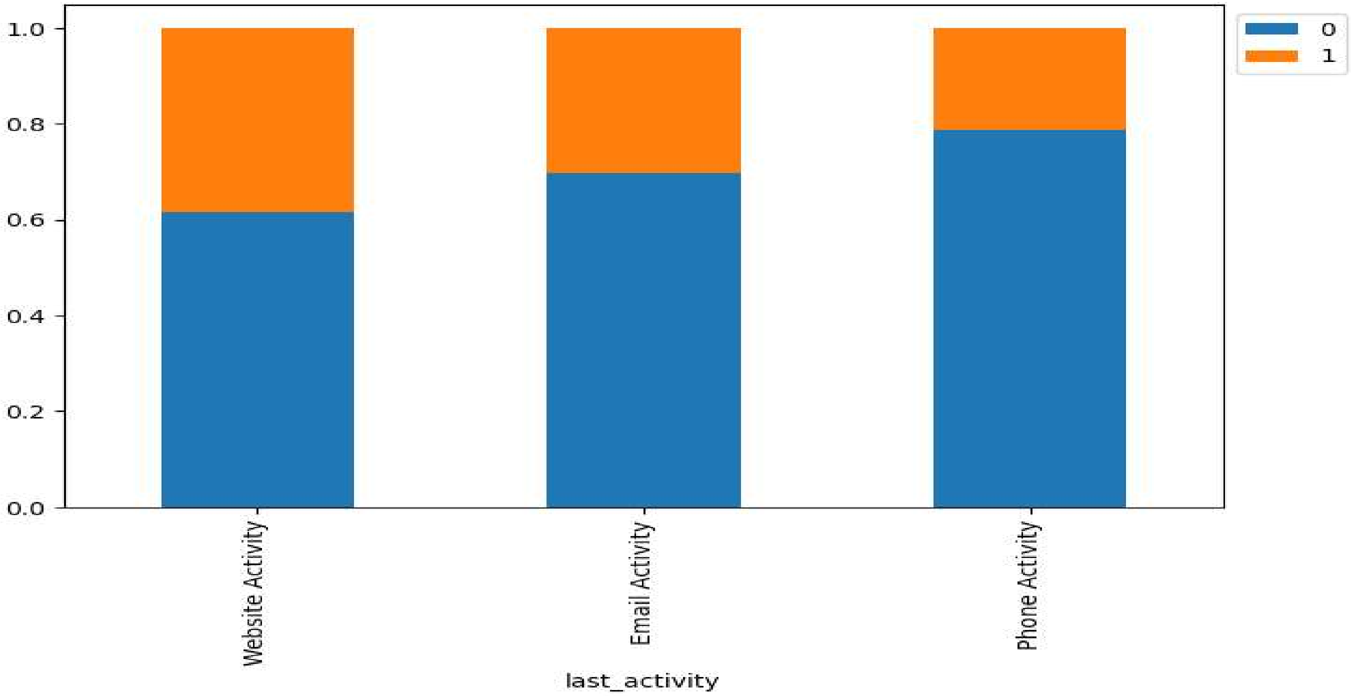

Figure 23

Observation;

- For "Email Activity" as the last activity, there are 1587 observations with status 0 and 691 observations with status 1, totaling 2278 observations.
- For "Website Activity" as the last activity, there are 677 observations with status 0 and 423 observations with status 1, totaling 1100 observations.
- For "Phone Activity" as the last activity, there are 971 observations with status 0 and 263 observations with status 1, totaling 1234 observations.
- Overall, across all types of last activity, there are 3235 observations with status 0 and 1377 observations with status 1, totaling 4612 observations.
- From the stacked bar plot we observe that certain types of last activity have a higher proportion of status 1 compared to others. This is inform us the strategies to prioritize or optimize certain types of activities based on their effectiveness in driving desired outcomes.

Let's see how advertisement and referrals impact the lead status
>Print_media_type1 vs Status:

```
status                0     1   All
print_media_type1
All                3235  1377  4612
No                 2897  1218  4115
Yes                 338   159   497
```

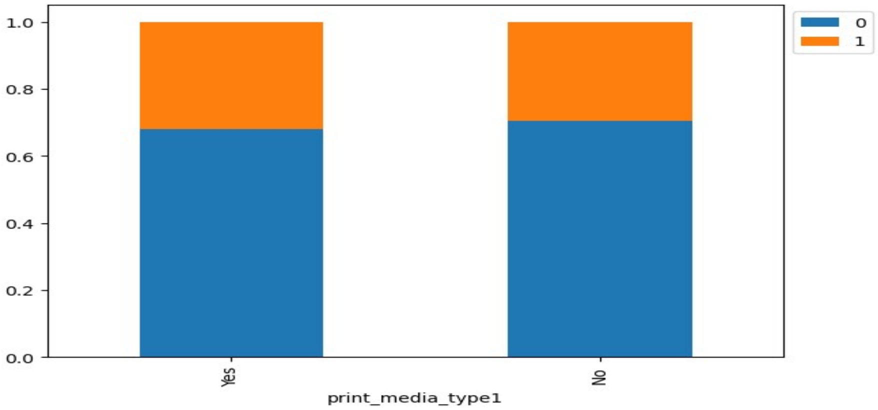

Figure 24

>Print_media_type2 vs Status:

status	0	1	All
print_media_type2			
All	3235	1377	4612
No	3077	1302	4379
Yes	158	75	233

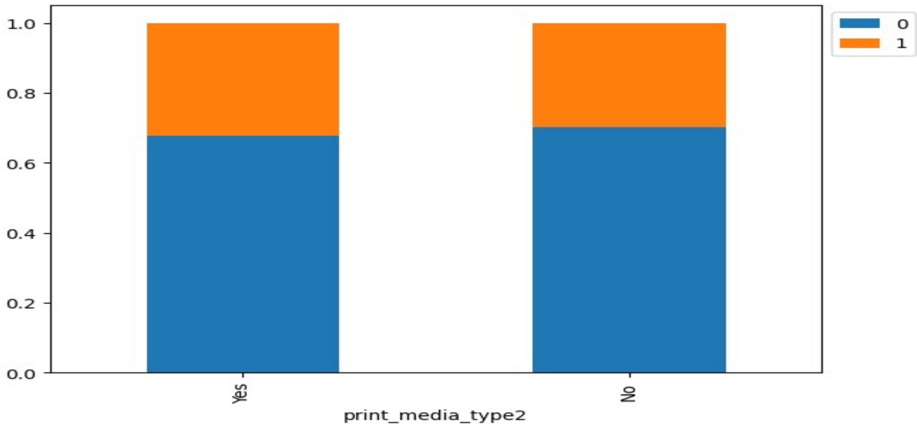

Figure 25

>Digital_media vs Status:

status	0	1	All
digital_media			
All	3235	1377	4612
No	2876	1209	4085
Yes	359	168	527

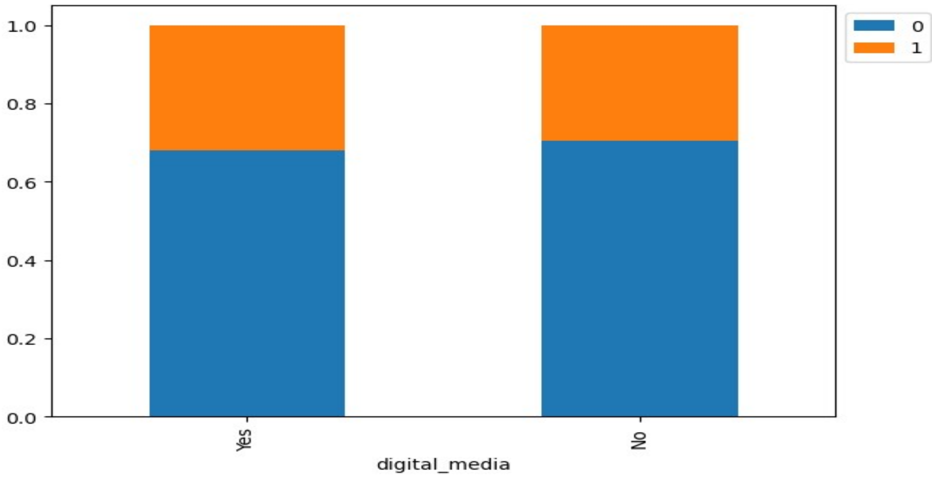

Figure 26

>Educational_channels vs Status:

status	0	1	All
educational_channels			
All	3235	1377	4612
No	2727	1180	3907
Yes	508	197	705

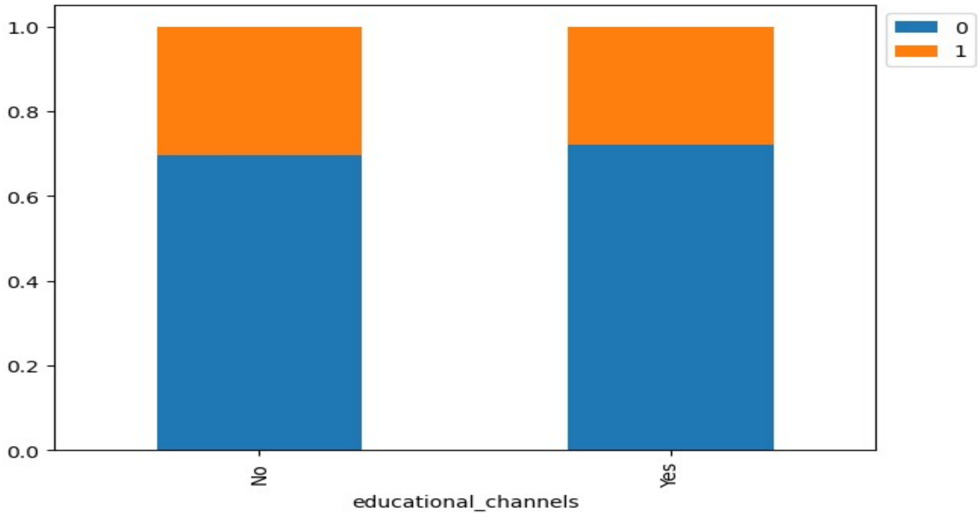

Figure 27

>Referral vs Status:

status	0	1	All
referral			
All	3235	1377	4612
No	3205	1314	4519
Yes	30	63	93

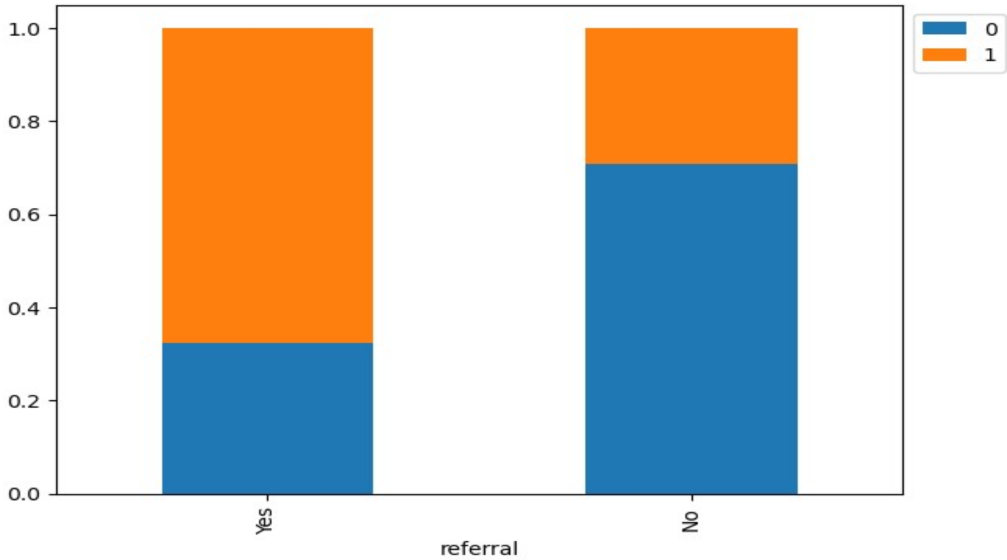

Figure 28

2- Data Pre-processing

Outlier Check:

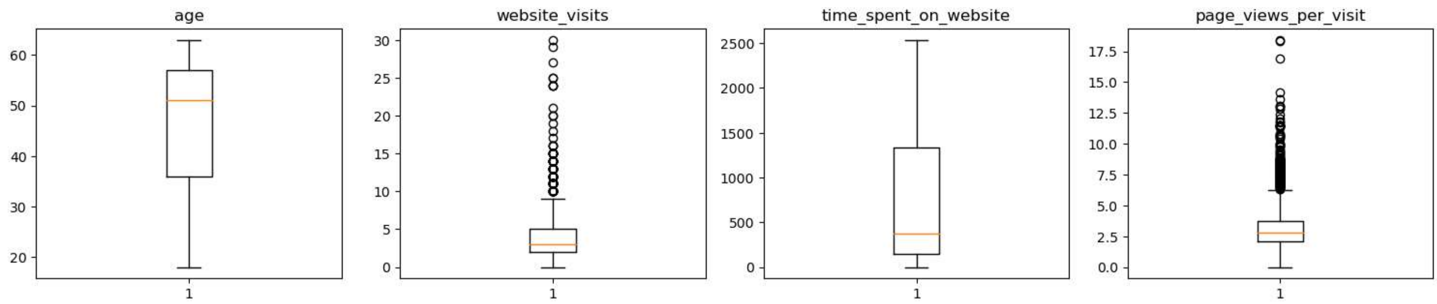

Figure 29

Observation; We see that website_visits and page_views_per_visit shows some outliers but age and time_spent_on_website have no outlier.

Data Preparation for Modeling:

- We want to predict which lead is more likely to be converted.
- Before we proceed to build a model, we'll have to encode categorical features.
- We'll split the data into train and test to be able to evaluate the model that we build on the train data.

```
Shape of Training set :  (3228, 16)
Shape of test set :  (1384, 16)
Percentage of classes in training set:
status
0   0.70415
1   0.29585
Name: proportion, dtype: float64
Percentage of classes in test set:
status
0   0.69509
1   0.30491
Name: proportion, dtype: float64
```

Model Evaluation Criterion:

>Model can make wrong predictions as:

1. Predicting a lead will not be converted to a paid customer in reality, the lead would have converted to a paid customer.
2. Predicting a lead will be converted to a paid customer in reality, the lead would not have converted to a paid customer.

Which case is more important?

- If we predict that a lead will not get converted and the lead would have converted then the company will lose a potential customer.
- If we predict that a lead will get converted and the lead doesn't get converted the company might lose resources by nurturing false-positive cases.

Losing a potential customer is a greater loss.

How to reduce the losses?

- Company would want Recall to be maximized, greater the Recall score higher are the chances of minimizing False Negatives.

First, let's create functions to calculate different metrics and confusion matrix so that we don't have to use the same code repeatedly for each model.

- The model_performance_classification_statsmodels function will be used to check the model performance of models.
- The confusion_matrix_statsmodels function will be used to plot the confusion matrix.

defining a function to compute different metrics to check performance of a classification model built using statsmodels

- **Thresholding prediction:** The predictions are compared against a threshold to classify them into binary classes. The threshold determines whether a predicted probability corresponds to class 1 or class 0.
- **Accuracy:** The proportion of correctly classified instances.
- **Recall (Sensitivity):** The proportion of actual positive cases that were correctly identified.
- **Precision:** The proportion of predicted positive cases that were correctly identified.
- **F1 Score**: The harmonic mean of precision and recall. It's a balance between precision and recall.

Splitting the Data

Dropping the column "status" from the DataFrame X and assign it to the variable Y. Then, it will create dummy variables for categorical features in X. Finally, it will split the data into training and testing sets with a 70:30 ratio using train_test_split, where X_train and y_train will contain 70% of the data for training, and X_test and y_test will contain 30% of the data for testing. The random_state parameter ensures reproducibility by fixing the random seed to 1.

3- Model Building - Logistic Regression

Building Model

First, let's create functions to calculate different metrics and confusion matrix so that we don't have to use the same code repeatedly for each model.

- The model_performance_classification_sklearn function will be used to check the model performance of models.
- The confusion_matrix_sklearn function will be used to plot the confusion matrix.

Building Logistic Regression Model

```
LogisticRegression
LogisticRegression(random_state=1)
```

>Checking model performance on training set:

	Accuracy	Recall	Precision	F1
0	0.82125	0.63770	0.72500	0.67855

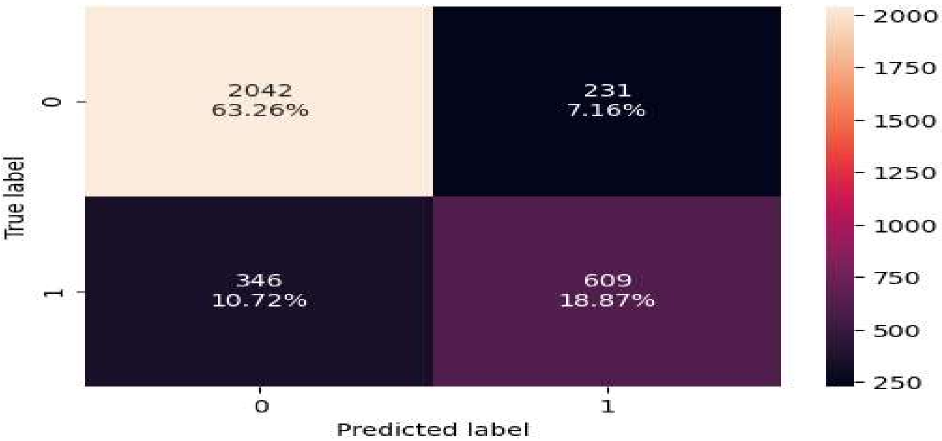

Figure 30

Observations;

- The model seems to be performing quite well overall, as evidenced by the relatively high number of true positives (bottom right quadrant) and true negatives (top left quadrant).
- The number of false positives (top right quadrant) appears to be relatively low compared to true positives, indicating that the model is not overly aggressive in predicting positive cases.
- Similarly, the number of false negatives (bottom left quadrant) also seems to be relatively low, suggesting that the model does not miss many positive cases.
- Overall, the distribution in the confusion matrix indicates a balanced performance of the model across both classes.

>Checking model performance on test set:

	Accuracy	Recall	Precision	F1
0	0.80925	0.61848	0.71703	0.66412

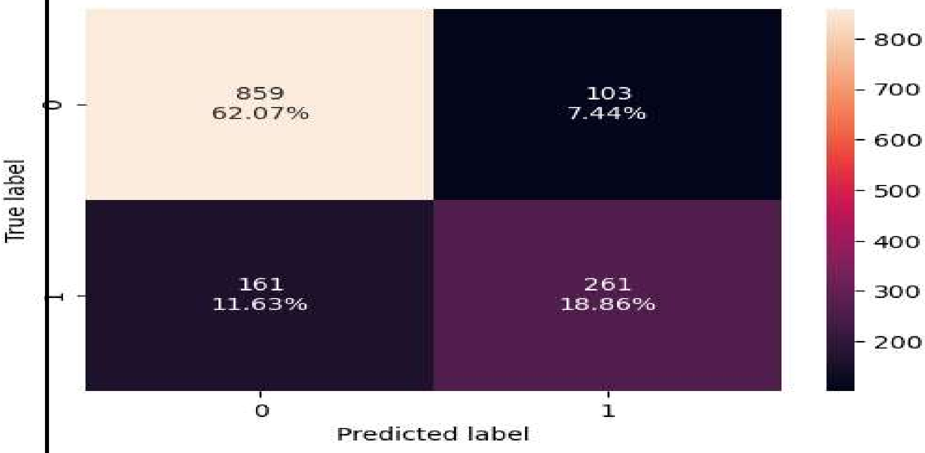

Figure 31

ROC-AUC

Train ROC-AUC score is: 0.8762545981393228

Test ROC-AUC score is: 0.860983978874974

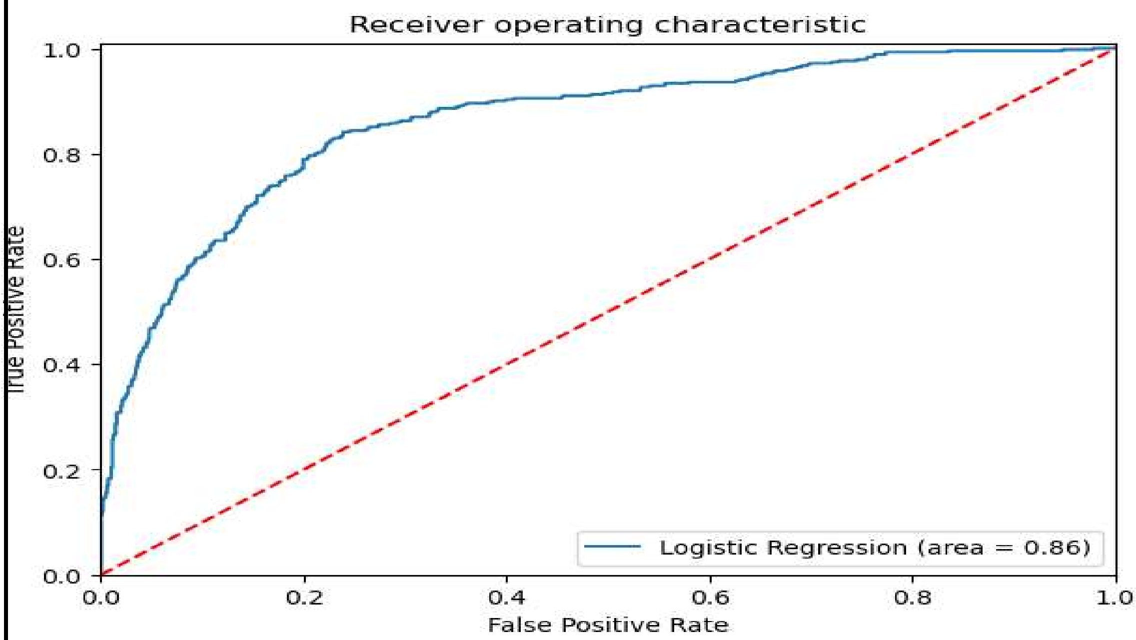

Figure 32

Observations; The ROC curve is well designed so the model is good.

4- Model Performance evaluation and improvement

Using GridSearch for Hyperparameter tuning of our logistic regression model

- Let's see if we can improve our model performance even more.

```
LogisticRegression
LogisticRegression(random_state=1, solver='liblinear', tol=0.0003)
```

>Checking performance on training set:

	Accuracy	Recall	Precision	F1
0	0.82125	0.63770	0.72500	0.67855

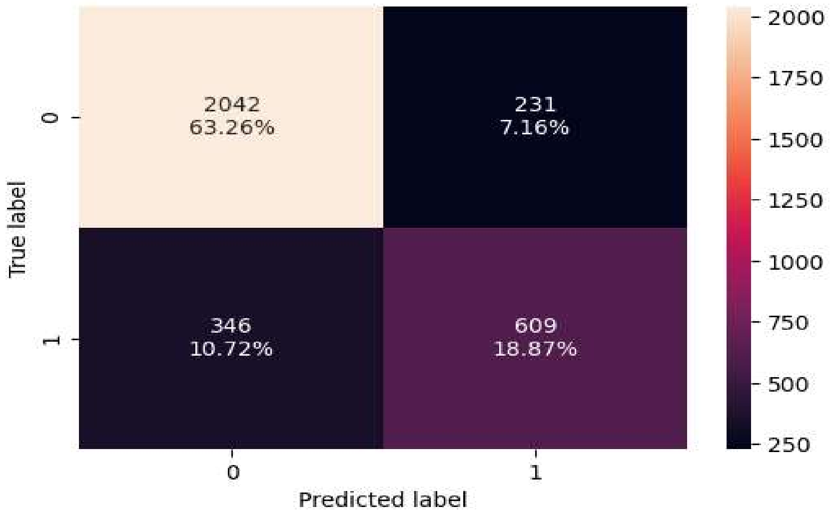

Figure 33

>Checking model performance on test set:

	Accuracy	Recall	Precision	F1
0	0.80925	0.61848	0.71703	0.66412

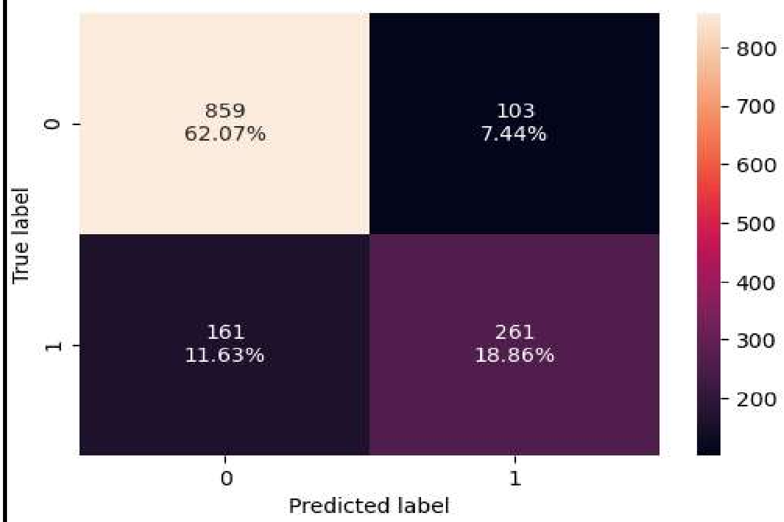

Figure 34

>Feature Importance:

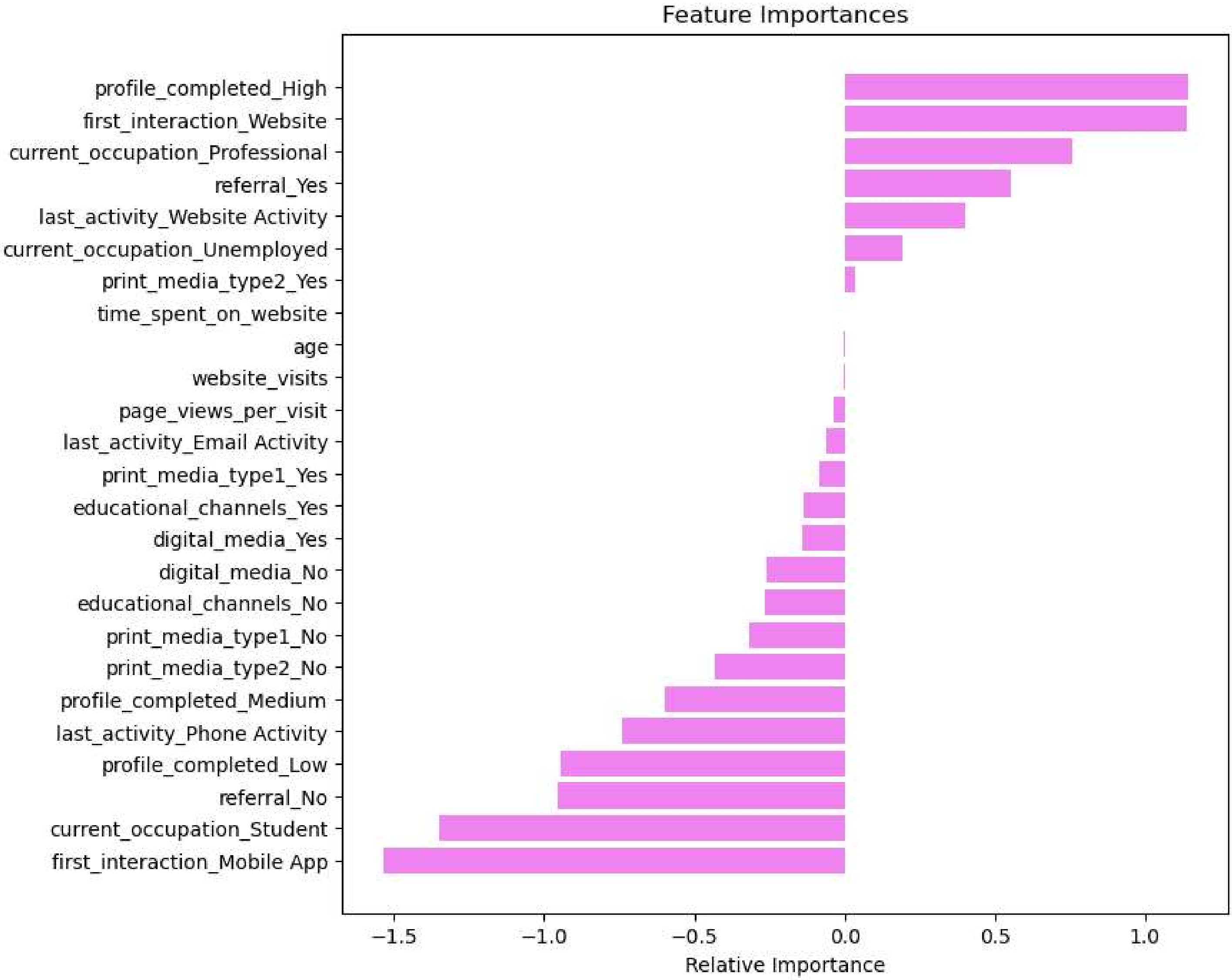

Figure 35

<u>Observations</u>; Some of relative importance are 0 to positive 1 but all the others features are 0 to negative 1.5 relative importance.

5- Model Building - Linear Discriminant Analysis

Building Linear Discriminant Analysis Model;

```
▾ LinearDiscriminantAnalysis
LinearDiscriminantAnalysis()
```

>Checking model performance on training set:

	Accuracy	Recall	Precision	F1
0	0.82404	0.65969	0.72165	0.68928

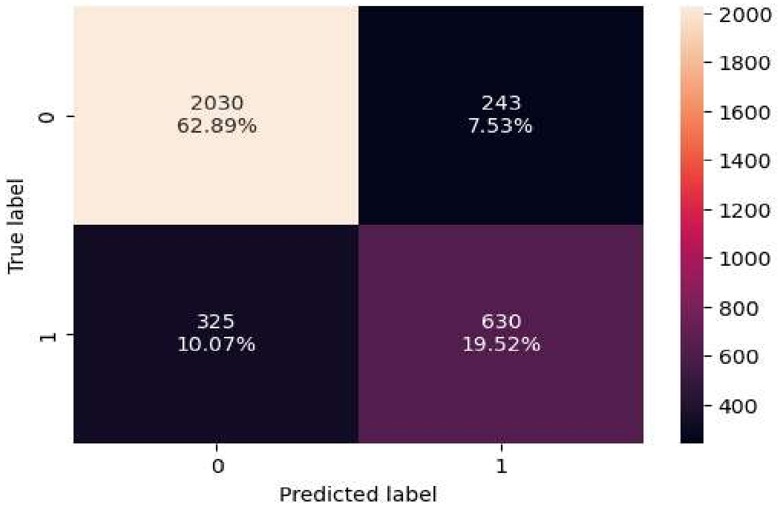

Figure 36

>Checking model performance on test set:

	Accuracy	Recall	Precision	F1
0	0.81069	0.62322	0.71858	0.66751

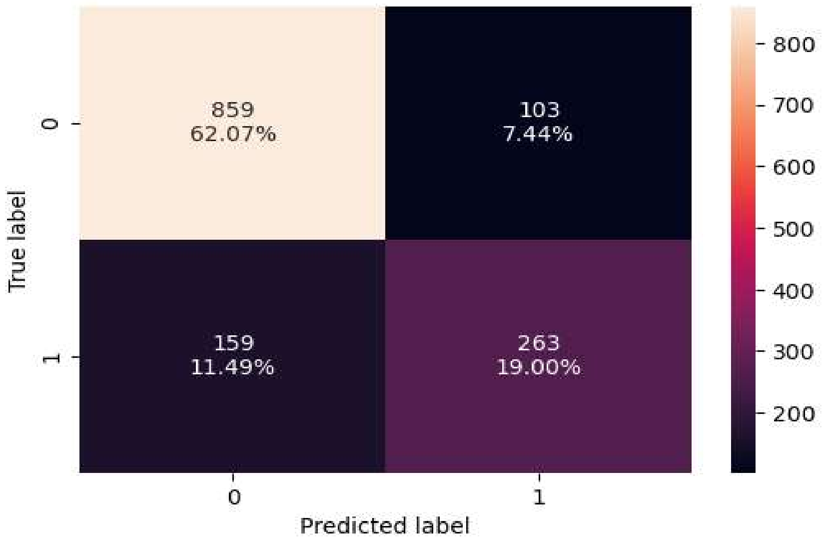

Figure 37

ROC-AUC;

Train ROC-AUC score is : 0.877286516193973

Test ROC-AUC score is : 0.8609544196037087

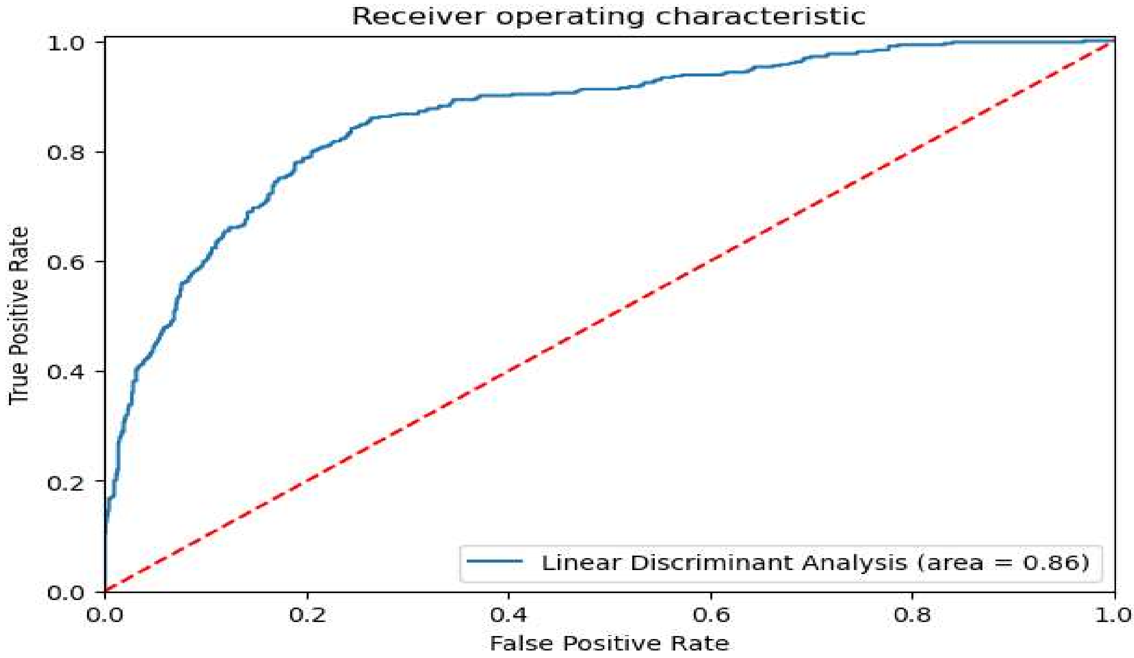

Figure 38

6- Model Performance evaluation and improvement

Using GridSearch for Hyperparameter tuning of our LDA model;

- Let's see if we can improve our model performance even more.

```
LinearDiscriminantAnalysis
LinearDiscriminantAnalysis(shrinkage=0.0, solver='lsqr')
```

>Checking model performance on training set:

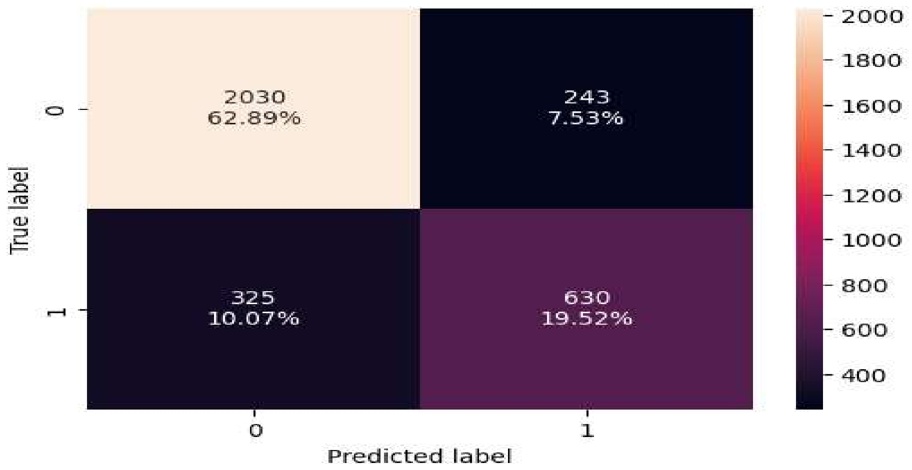

Figure 39

>Checking model performance on test set:

	Accuracy	Recall	Precision	F1
0	0.81069	0.62322	0.71858	0.66751

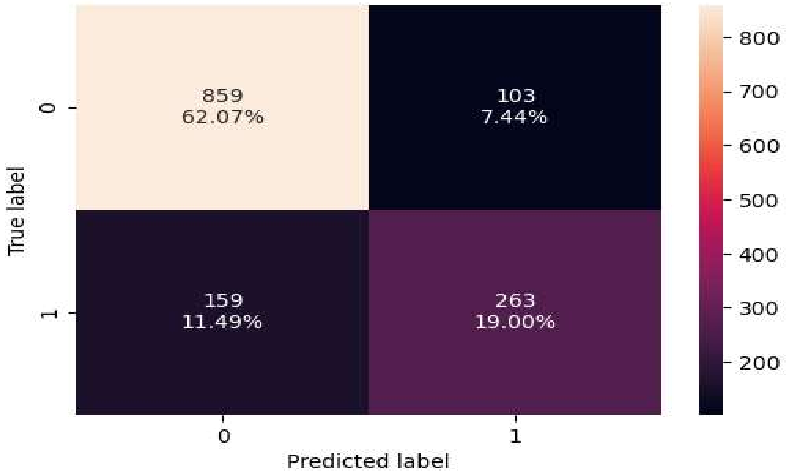

Figure 40

>Feature Importance:

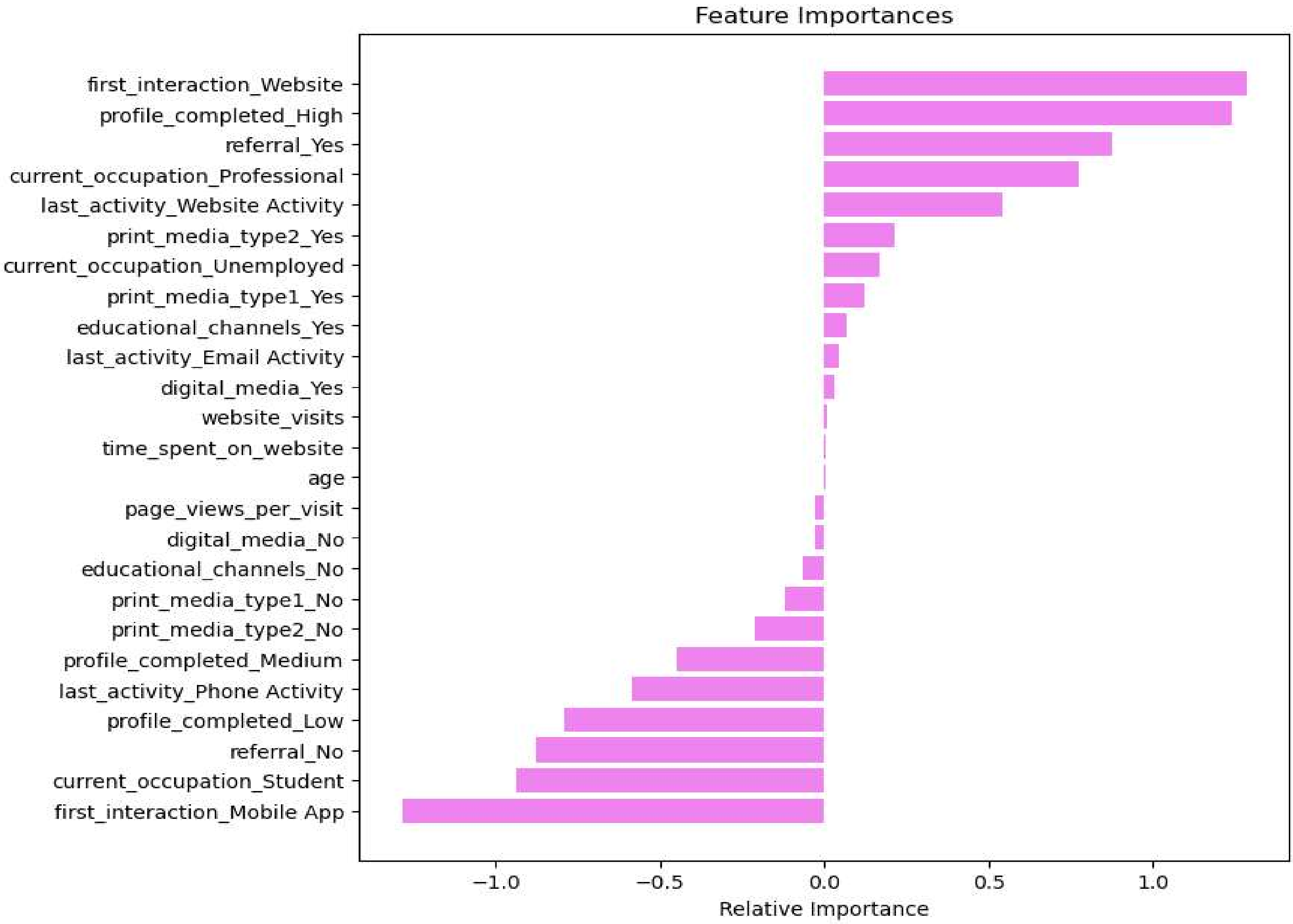

Figure 41

Observations;

- The plot visualizes the relative importance of each feature in the Logistic Regression model.
- Features with larger coefficients have a more significant impact on the model's predictions.
- Positive coefficients indicate features that positively contribute to the target variable, while negative coefficients indicate features that negatively contribute.
- By observing the plot, one can identify which features are the most influential in predicting the target variable based on their coefficients.

7- Model Building - CART

Building Decision Tree Model;

```
DecisionTreeClassifier
DecisionTreeClassifier(random_state=1)
```

>Checking model performance on training set:

	Accuracy	Recall	Precision	F1
0	1.00000	1.00000	1.00000	1.00000

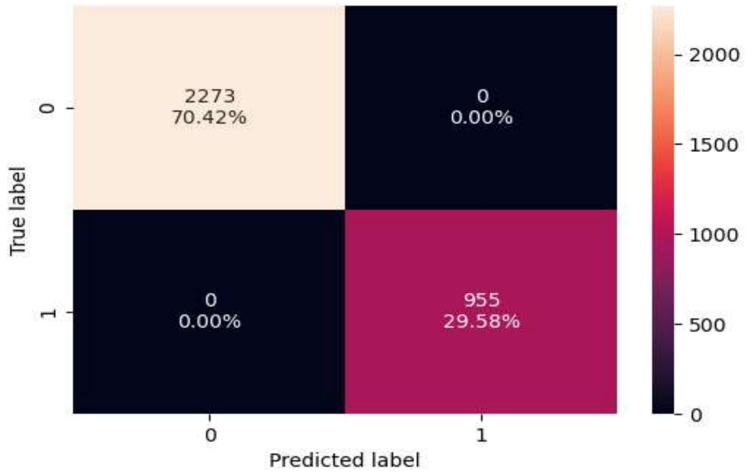

Figure 42

>Checking model performance on test set:

	Accuracy	Recall	Precision	F1
0	0.79769	0.67062	0.66745	0.66903

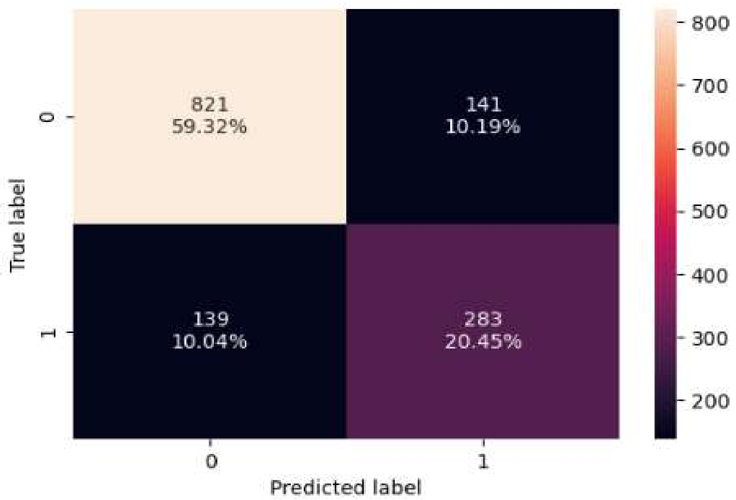

Figure 43

ROC-AUC;
Train ROC-AUC score is : 1.0

Test ROC-AUC score is : 0.7620232335872146

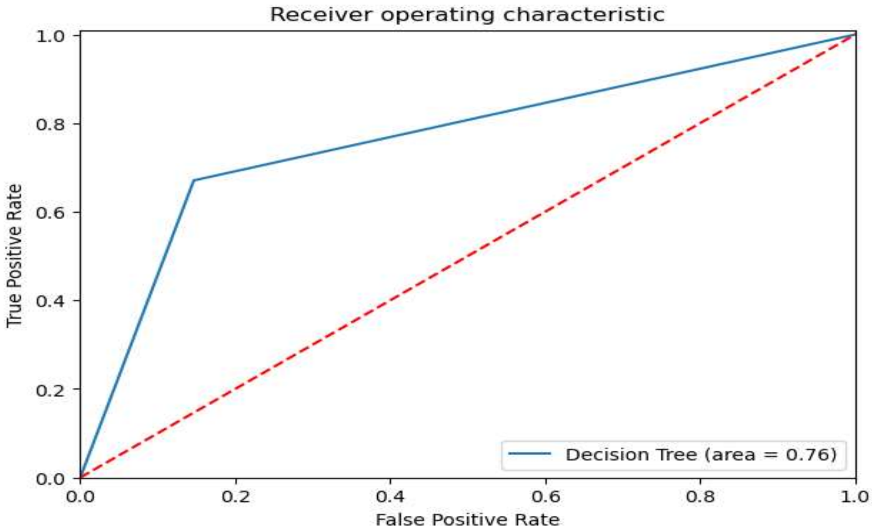

Figure 44

Observations;

- The training ROC-AUC score of 1.0 indicates that the model achieved perfect discrimination on the training set. It is suggest that the model has memorized the training data, which could potentially indicate overfitting.
- The test ROC-AUC score of 0.762 suggests that the model's performance on unseen data is decent but not perfect. A score of 0.762 indicates that the model is able to discriminate between the positive and negative instances in the test set better than a random classifier (which would have an AUC of 0.5), but it's not perfect like the training set performance. It is suggest that the model is generalizing reasonably well to unseen data, but there might still be room for improvement.

8- Model Performance evaluation and improvement

Using GridSearch for Hyperparameter tuning of our Decision Tree model;

- Let's see if we can improve our model performance even more.

```
DecisionTreeClassifier
DecisionTreeClassifier(class_weight={0: 0.3, 1: 0.7}, max_depth=5
                       max_leaf_nodes=20, random_state=1)
```

>Checking performance on training set:

	Accuracy	Recall	Precision	F1
0	1.00000	1.00000	1.00000	1.00000

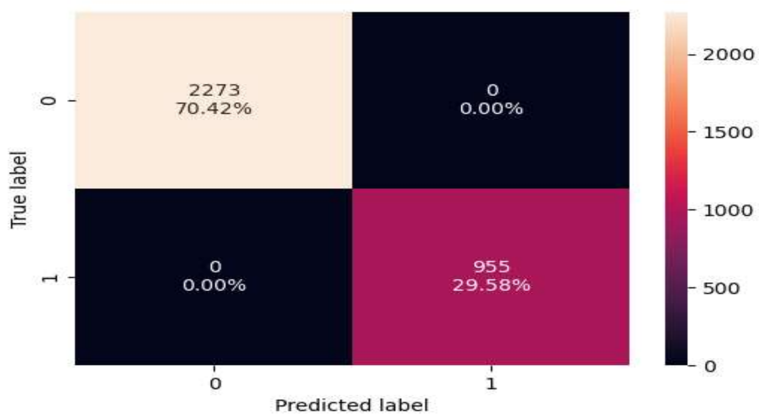

Figure 45

Observations;

- Achieving perfect performance on all metrics, especially on the training set, could be a sign of overfitting. Overfitting occurs when the model learns the training data too well, including its noise and outliers, to the extent that it doesn't generalize well to unseen data.
- While perfect performance on the training set is desirable, it's essential to evaluate the model's performance on an independent test set to ensure it can generalize well to unseen data. A perfect score on the training set doesn't guarantee similar performance on new data.

>Checking performance on test set:

	Accuracy	Recall	Precision	F1
0	0.79769	0.67062	0.66745	0.66903

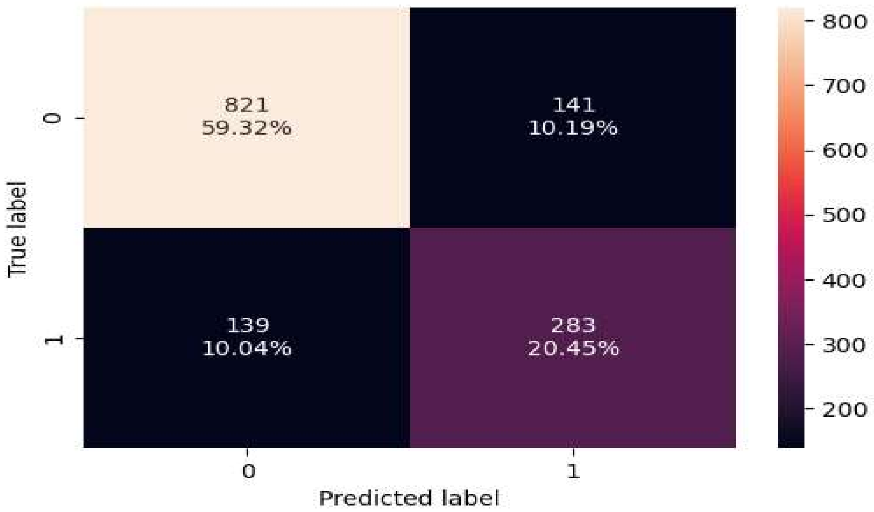

Figure 46

Observations;

- The accuracy is relatively high, indicating that the model performs well in terms of overall correct classifications.
- Recall and precision are similar, which means that the model is making a balanced number of true positive predictions and minimizing false negatives and false positives.
- The F1 score is close to the precision and recall values, indicating a good balance between precision and recall.

Visualizing the Decision Tree;

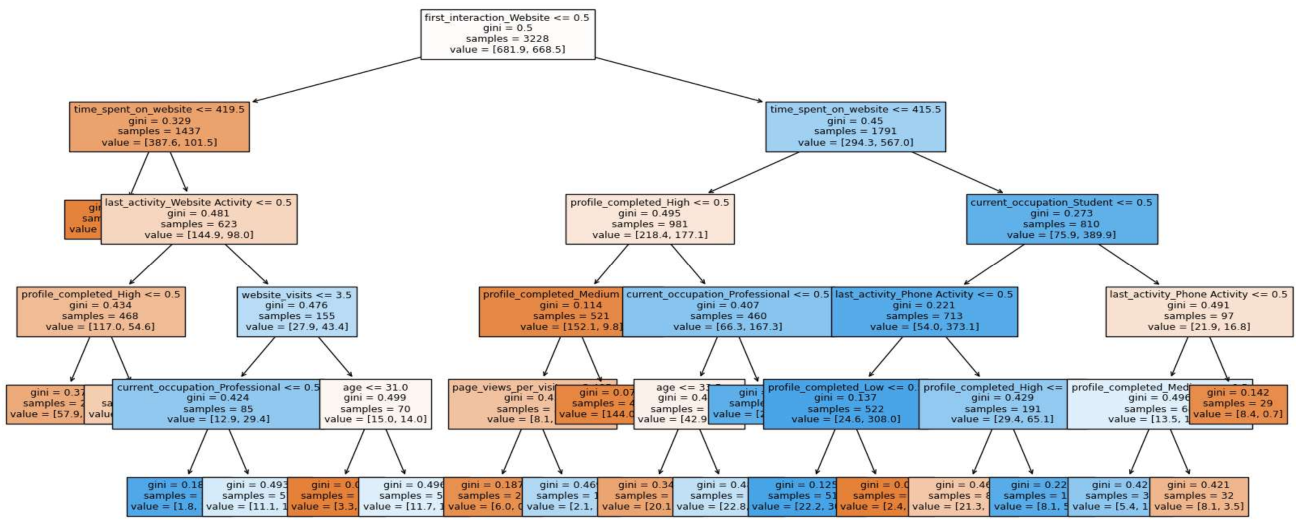

Figure 47

Observation;

- The root node of the decision tree is first_interaction_website and gini value is 0.5, samples 3228, the values are 681.9 and 668.5.
- It seems to be split based on various features like "first_interaction_Website," "time_spent_on_website," "last_activity_Website Activity," "profile_completed_High," etc. and each of the split leads to further subdivisions until reaching a decision.
- From the decision tree model some leads seem to have more significant impacts on the final classification which are "time_spent_on_website," "profile_completed_High".
- The class distribution varies across different paths of the tree, indicating different levels of confidence in predictions.

>Feature Importance:

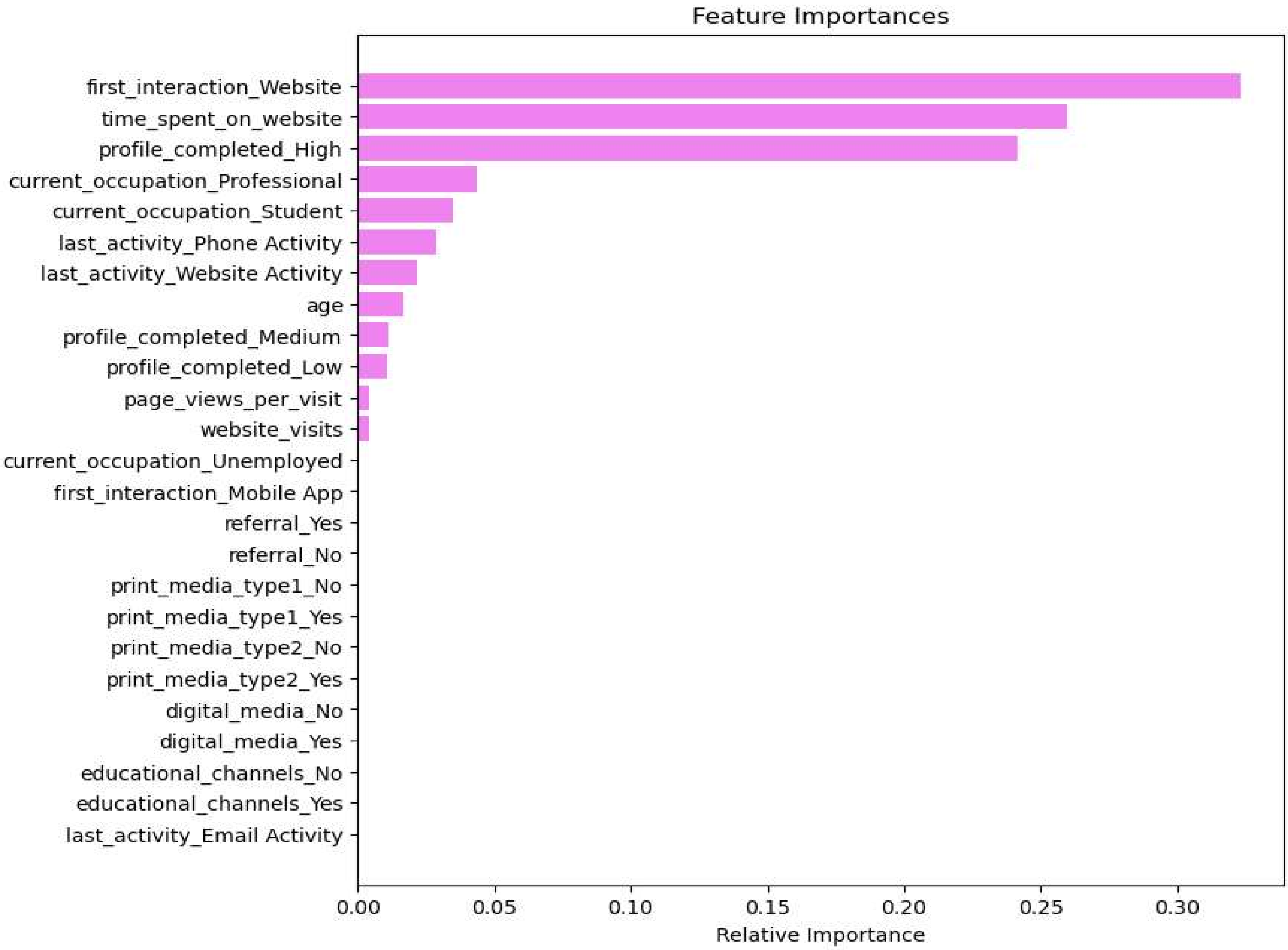

Figure 48

Observations;

- From the above plot visually represents the relative importance of each feature in the decision tree model.
- Features with taller bars indicate higher importance in predicting the target variable.
- This visualization helps in identifying which features are most influential in the model's decision-making process.
- The plot can assist in feature selection or understanding for which features are driving the model's predictions the most, which could be valuable for feature engineering or model interpretation.

7- Actionable Insights & Recommendations

Comparing all the models;

>Training performance comparison:

Training performance comparison:

	Logistic Regression	Logistic Regression After Tuning	LDA
Accuracy	0.82125	0.82125	0.82404
Recall	0.63770	0.63770	0.65969
Precision	0.72500	0.72500	0.72165
F1	0.67855	0.67855	0.68928

	LDA After Tuning	Decision Tree	Decision Tree After Tuning
Accuracy	0.82404	1.00000	1.00000
Recall	0.65969	1.00000	1.00000
Precision	0.72165	1.00000	1.00000
F1	0.68928	1.00000	1.00000

<u>Observations</u>;

- Logistic Regression and LDA show consistent performance before and after tuning, suggesting that tuning didn't significantly impact their performance.
- The decision tree model performs suspiciously well, achieving perfect scores both before and after tuning. This could indicate overfitting, especially if these scores don't generalize well to unseen data.

>Testing performance comparison:

Testing performance comparison:

	Logistic Regression	Logistic Regression After Tuning	LDA
Accuracy	0.80925	0.80925	0.81069
Recall	0.61848	0.61848	0.62322
Precision	0.71703	0.71703	0.71858
F1	0.66412	0.66412	0.66751

	LDA After Tuning	Decision Tree	Decision Tree After Tuning
Accuracy	0.81069	0.79769	0.79769
Recall	0.62322	0.67062	0.67062
Precision	0.71858	0.66745	0.66745
F1	0.66751	0.66903	0.66903

<u>Observations</u>;

- Logistic Regression and LDA demonstrate consistent performance on the test set, similar to their performance on the training set.
- The decision tree's performance is also consistent between the training and test sets, suggesting that it's not overfitting and generalizes reasonably well to unseen data.

Business Recommendations:

- Users higher they spent time on the website which are likely to convert as Leads. Extraalearn Website needs to be more attractive to the users to keep them engaged in the Website about the content and demonstration of thr course.
- Website has a better first time reach with customers compared to the MobileApp. More details in the data are required in order to provide more analysis.
- Profile_completion also takes further significance whereas High & Medium are two important factors contributes to the Lead conversion status. The organization needs to re-evaluate the profile section, if the infomration is totally relevant & signifiant to their business or not and it wasalso helpful to trim down the details which are required at the Profile page and helps the user to complete their profile 100%.
- Most of the leads who converted were belongs to Professional category , this could be due to the course fee as working professionals can afford the learning content more than Unemployed or Student.
- Referral seems do not have much influence on the conversion rate this could be due to either the Institution does not have any communication with Alumni group post the completion of the course or The training courses are not helping students on finding any job opportunities.
- Advertisement Magazine promotions has to be improved more as it's the least when compared to other platfroms such as Digital media & Newspaper advertisements.