MACHINE LEARNING

GRADED PROJECT

GUIDED

List of Contents

SL. No.	DESCRIPTION	PAGE NO.
1	Problem Statement (Context)	7
2	Data Description	7
3	Data Overview	9
4	Statistical Summary of the data	10
5	Exploratory Data Analysis (EDA) Q&A as per Rubrics	12
6	Univariate Analysis	17
7	Data Preprocessing	62
8	Data Preparation for Modelling	64
9	Model Building	66
10	Naive Bayes Classifier	72
11	KNN Classifier	77
12	Decision Tree Classifier	82
13	Model Performance Improvement	86
14	KNN Classifier Performance with different K values	109
15	Decision Tree Classifier (Pre- running)	116
16	Visualizing The Decision Tree	122
17	Model Performance Comparison and Final Selection	124
18	Actionable Insights and Recommendations	130

List of Tables

SL. NO	DESCRIPTION	PAGE NO.
1	Statistical Summary of Data	9
2	Summary Statistics by Current Occupation	44
3	Train & Test set in the ratio of 75:25	64
4	Percentage of classes in the training set	65
5	Percentage of classes in the test set	65
6	Logistic Regression Results	66
7	Checking Logistic Regression Model Performance on the Training Set	67
8	Checking Naive Bayes Classifier Performance on the Training Set	72
9	Checking Naive Bayes Classifier Performance on the Test Set	74
10	Checking KNN Classifier Performance on the Training Set	77
11	Checking KNN Classifier Performance on the Test Set	80
12	Checking Decision Tree Classifier Performance on Training Set	82
13	Checking Decision Tree Classifier Performance on Test Set	84
14	Variance Inflation Factor (VIF) Table	87
15	Logistic Regression Result	100
16	Checking Tuned Logistic Regression Model Performance on Training Set	104
17	Checking Tuned Logistic Regression Model Performance on Test Set	107
18	KNN Classifier	110

19	Checking Tuned KNN Model Performance on Training Set	111
20	Checking Tuned KNN Model Performance on Test Set	114
21	Checking Tuned Decision Tree Classifier Performance on Training Set	116
22	Checking Tuned Decision Tree Classifier Performance on Test Set	119
23	Performance Comparison of Model on Training Set	124
24	Performance Comparison of Model on Test Set	128

List of Figures

SL. NO	DESCRIPTION	PAGE NO.
1	Plot: Lead Conversion Rate by Current Occupation	13
2	Plot: Lead Conversion Rate by First Interaction Channel	14
3	Plot: Success Rate by First Interaction Mode	16
4	Plot: Lead Conversion Rate by Source Channel	17
5	Bar Plot for success rates for Interaction Channels	19
6	Plot: Observation on Age	21
7	Plot: Observation on Website Visits	23
8	Plot: Observation on Number of Time spent on website	25
9	Plot: Observation on Number of page views per visit	27
10	Plot: Observation on Number of Adults	29
11	Plot: Observation on Profile Completed	31
12	Plot: Observation on Print Media Type 1	33
13	Plot: Observation on Print Media Type 2	34
14	Plot: Observation on Digital Media	36
15	Plot: Observation on Educational Channels	38
16	Plot: Observation on Referrals	39
17	Observation on Status	42
18	Plot of Heatmap	43
19	Stacked Bar plot of Current Occupation vs Status	47
20	Box plot of Age by Current Occupation	49
21	Stacked Bar plot of First interaction by Target	51
22	Box plot of Time spent on website by Target	53
23	Distribution Plot of website visits by Target Status	55
24	Distribution Plot of Page Views per visit by Target Status	57
25	Stacked Bar plot of Profile Completed by Status	59
26	Stacked Bar plot of Last Activity vs Status	60
27	Stacked Bar plot of Digital Media vs Status	61

28	Stacked Bar plot of print media type 1	62
29	Stacked Bar plot of referral vs Status	68
30	Outlier Detection using Box plot	70
31	Plot of Confusion Matrix for Training Set	73
32	Plot of Confusion Matrix on Test Set	76
33	Plot of Confusion Matrix for Training Set	78
34	Plot of Confusion Matrix for Test Set	81
35	Plot of Confusion Matrix for Training Set	83
36	Plot of Confusion Matrix for Test Set	85
37	Plot of Confusion Matrix for Training Set	89
38	Plot of Confusion Matrix for Test Set	90
39	Histogram of Time spent on Website by Student status	91
40	Line plot of Time spent on Website	93
41	Scatter plot of Current Occupation (student vs unemployed)	94
42	Scatter plot of Time spent on website vs Current occupation Student	96
43	Histogram of first_interaction_website	97
44	Histogram of current_occupation_unemployed	98
45	Histogram of current_occupation_student	103
46	Histogram of time_spent_on_website	106
47	Plot of Optimal Threshold using ROC Curve	108
48	Plot of Confusion Matrix on Training Set	113
49	Plot of Confusion Matrix for Test Set	115
50	Plot of Confusion Matrix on Training Set	118
51	Plot of Confusion Matrix for Test Set	120
52	Decision Tree	132

Problem Statement

❑ CONTEXT

ExtraaLearn, an emerging startup in EdTech, focusses on upskilling and reskilling through cutting edge technology programs. With the online education sector rapidly growing, the company faces the challenge of identifying leads most likely to convert into paying customers. To address this, the objective is to leverage data science to:

- Analyze and build a machine learning model to predict the lead conversion
- Identify the key factors influencing the conversion process such as engagement on social media, website/ app interactions and E-mail communications.
- Develop a profile of leads likely to convert, enabling targeted allocation of resources for maximum effectiveness in customer acquisition.

By understanding and utilizing these insights ExtraaLearn aims to optimize its marketing strategies and enhance conversion rates in a competitive online education market.

❑ DATA DESCRIPTION:

The dataset contains various attributes about the leads and their interactions with ExtraaLearn. Here is an overview of the column in the dataset.

- **ID**: ID of the lead
- **age**: Age of the lead
- **current_occupation**: Current occupation of the lead (Professional, Unemployed, Student)
- **first_interaction**: How the lead first interacted with ExtraaLearn (Website, Mobile App)
- **profile_completed**: The percentage of the profile filled by the lead on the website / mobile app (Low: 0-50%, Medium: 50-75%, High: 75-100%)
- **website_visits**: Number of times the lead visited the website
- **time_spent_on_the_website**: Total time spent on the website in seconds
- **page_views_per_visit**: Average number of pages viewed during the visits
- **last_activity**: Last interaction between the lead and ExtraaLearn (Email Activity, Phone Activity, Website Activity)
- **print_media_type1**: Flag indicating whether the lead saw the ad in the newspaper
- **print_media_type2**: Flag indicating whether the lead saw the ad on the digital platform
- **educational_channels**: Flag indicating whether the lead heard about ExtraaLearn through educational channels (online forums, discussion threads, educational websites etc.)
- **referral**: Flag indicating whether the lead heard about ExtraaLearn through reference

- **Status**: Flag indicating whether the lead was converted to a paid customer (1: Yes, 2: No)

❑ DATA OVERVIEW

- The dataset consists of **4612** entries (rows) and **15** columns.
- The dataset is clean with **no missing values** in the columns.
- The data types of columns include **10 object data type, 4 int64, and 1 float data type.**
- There are **no duplicate values** in the dataset.
- There are **4612 unique value** columns in the dataset.

❑ STATISTICAL SUMMARY OF THE DATA

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

TABLE 1

Insights:

- The age distribution of the lead's ranges from 18 to 63 with **median age of 51**. This shows a wide age range with a concentration around the middle-aged group.
- Most of the leads have **2 to 5 visits.** There are few outliers with up to 30 visits. On an average lead visit the website about **3.57 times.**
- Average time spent on the website is **6 minutes** with large standard deviation. The top quartile of users spends more time indicating deeper level of engagement.
- Leads view an average of **3 pages per visit.**
- The overall conversion rate is approximately **29.86%**

- ✓ The median and the 25th percentile both being 0 shows that more than half of the leads do not convert but approx. 30% that does convert.

❑ EXPLORATORY DATA ANALYSIS (EDA)

We will now proceed with the EDA to answer the questions in Rubrics.

Q1: Leads will have different expectations from the outcome of the course and the current occupation may play a key role in getting them to participate in the program. Find out how current occupation affects lead status.

Answer 1:

- **Lead Conversion Rate by Current Occupation:**

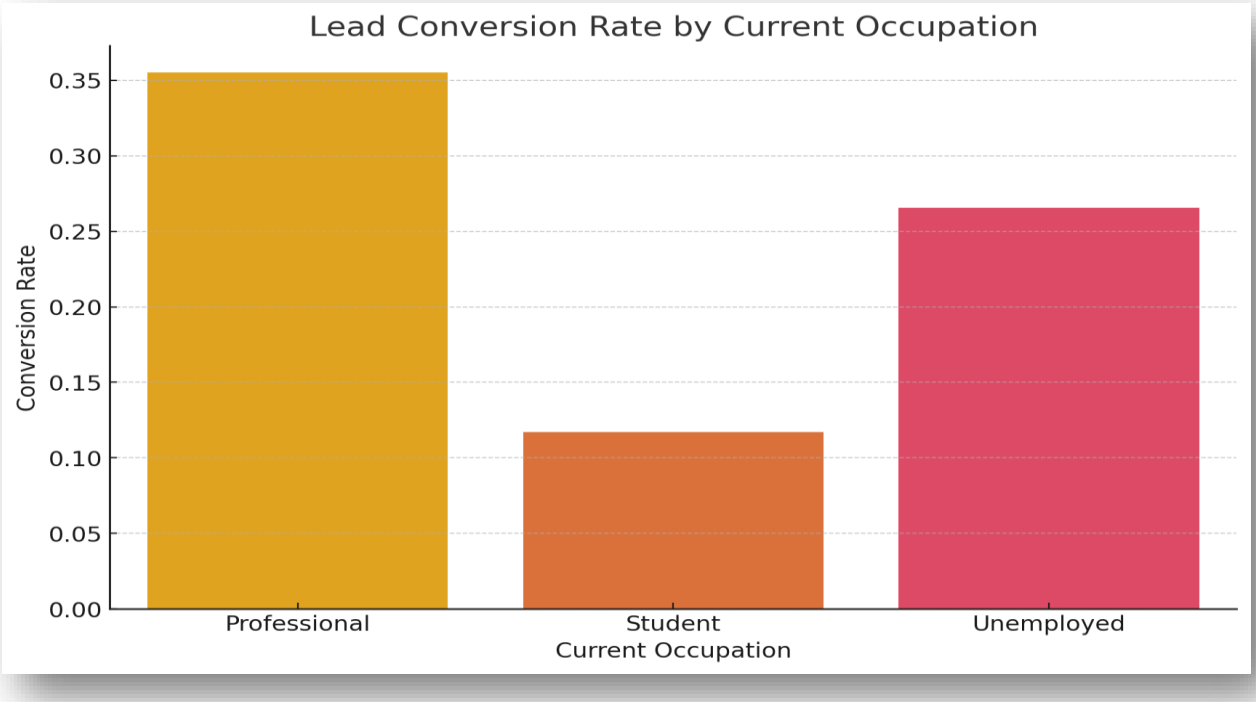


FIGURE1

Insights:

- The bar plot shows the conversion rate by current occupation.
- **'Professional'** leads have the highest conversion rate.
- **'Students'** have a lower conversion rate compared to the professionals.
- **'Unemployed'** leads have the lowest conversion rate.
- This suggests that the **professionals** are more likely to convert into paid customers compared to **students** and **unemployed** individuals.

Q2: The company's first impression on the customer must have an impact. Do the first channels of interaction have an impact on the lead status?

Answer 2: Let's analyze the impact of the first channel of interaction on lead status.

- **Lead Conversion Rate by First Interaction Channel:**

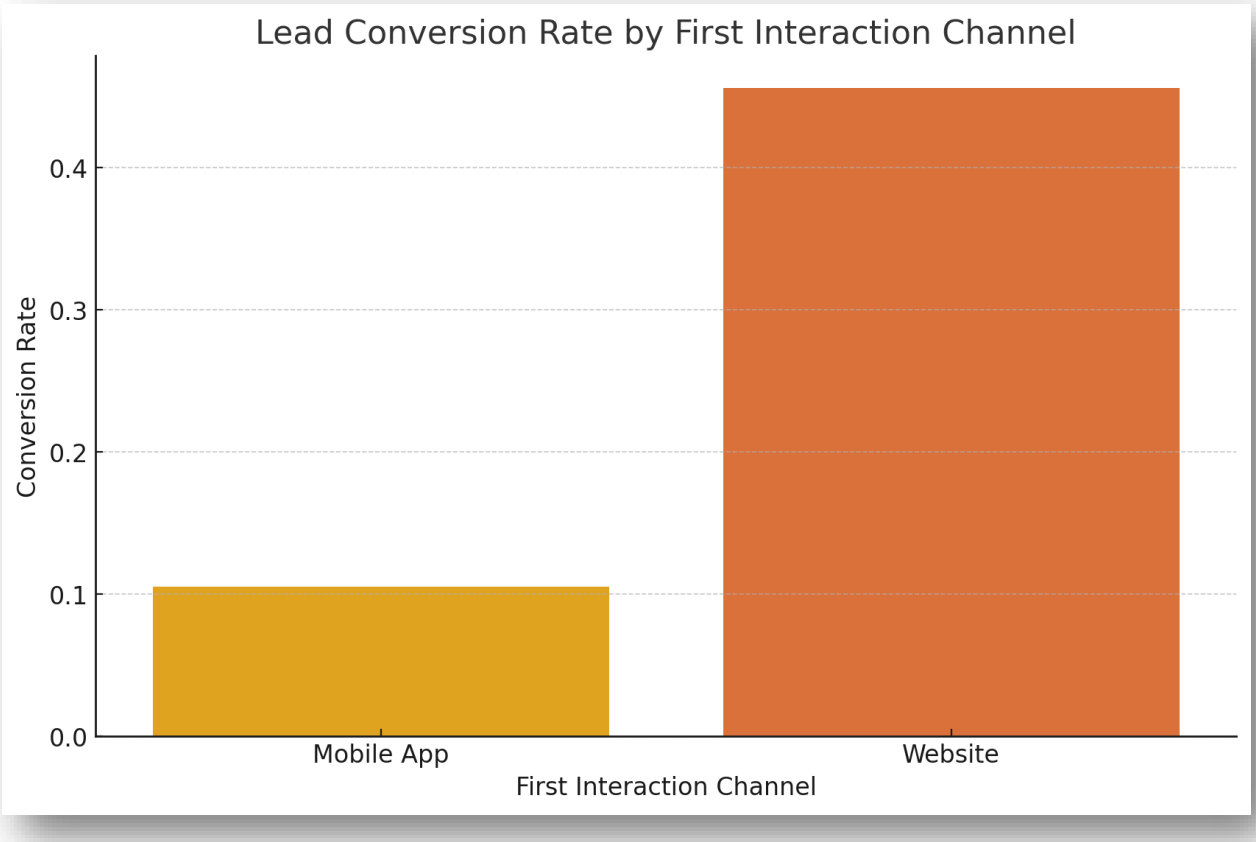


FIGURE 2

Insights:

- The bar plot indicates the conversion rate by First Interaction Channel.
- The conversion rate for leads first interacting through the website is significantly higher than those interacting through the mobile app.
- The website conversion rate is above 0.4 while the mobile app is just above 0.1
- This indicates that the website is more effective channel for converting leads compared to mobile app.

- Given the lower conversion rate for the mobile app there might be potential areas for improvement. This could include user experience enhancement, better call-to-action placement, or more personalized interaction within the app.

Q3: The company uses multiple modes to interact with prospects. Which way of interaction works best?

Answer 3: The dataset includes various columns related to interaction with the prospects such as ‘first_interaction’, ‘website_visits’, ‘time_spent_on_website’, ‘page_views_per_visit’ and ‘last_activity’.

To determine which interaction medium works best we analyzed ‘first_interaction’ column in relation with ‘status’ column which indicates whether the prospect became customer (1) or not (0)

Let's create a plot to visualize the effectiveness of different interaction modes.

- **Bar plot shows the success rates for each first interaction mode:**

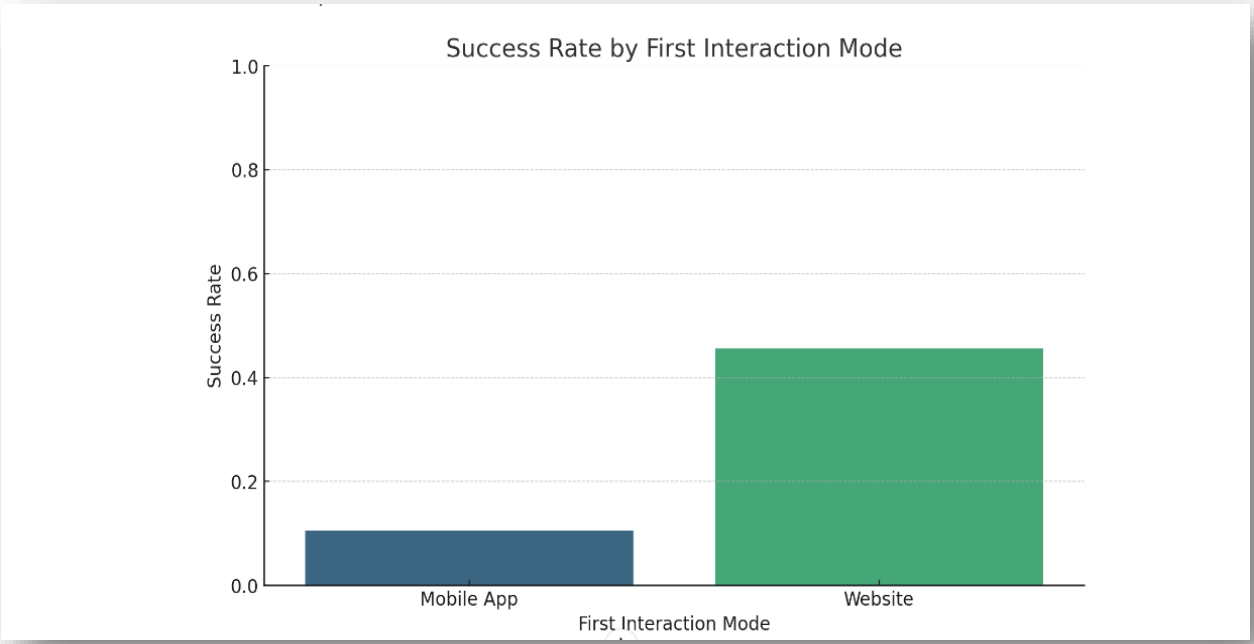


FIGURE 3

Insights:

- Website: Appears to have a higher success rate compared to other modes of interaction.
- Mobile App: It has a moderate success rate.
- Other modes: Any other modes not specified here have a lower or less consistent success rates.
- Interacting with prospects via the website tends to lead to a higher conversion rate, making it the most effective initial interaction mode.

Q4: The company gets leads from various channels such as print media, digital media, referral etc. Which of these channels has the highest lead conversion rate?

Answer 4:

- **Plot of Lead Conversion Rate by Source Channel:**

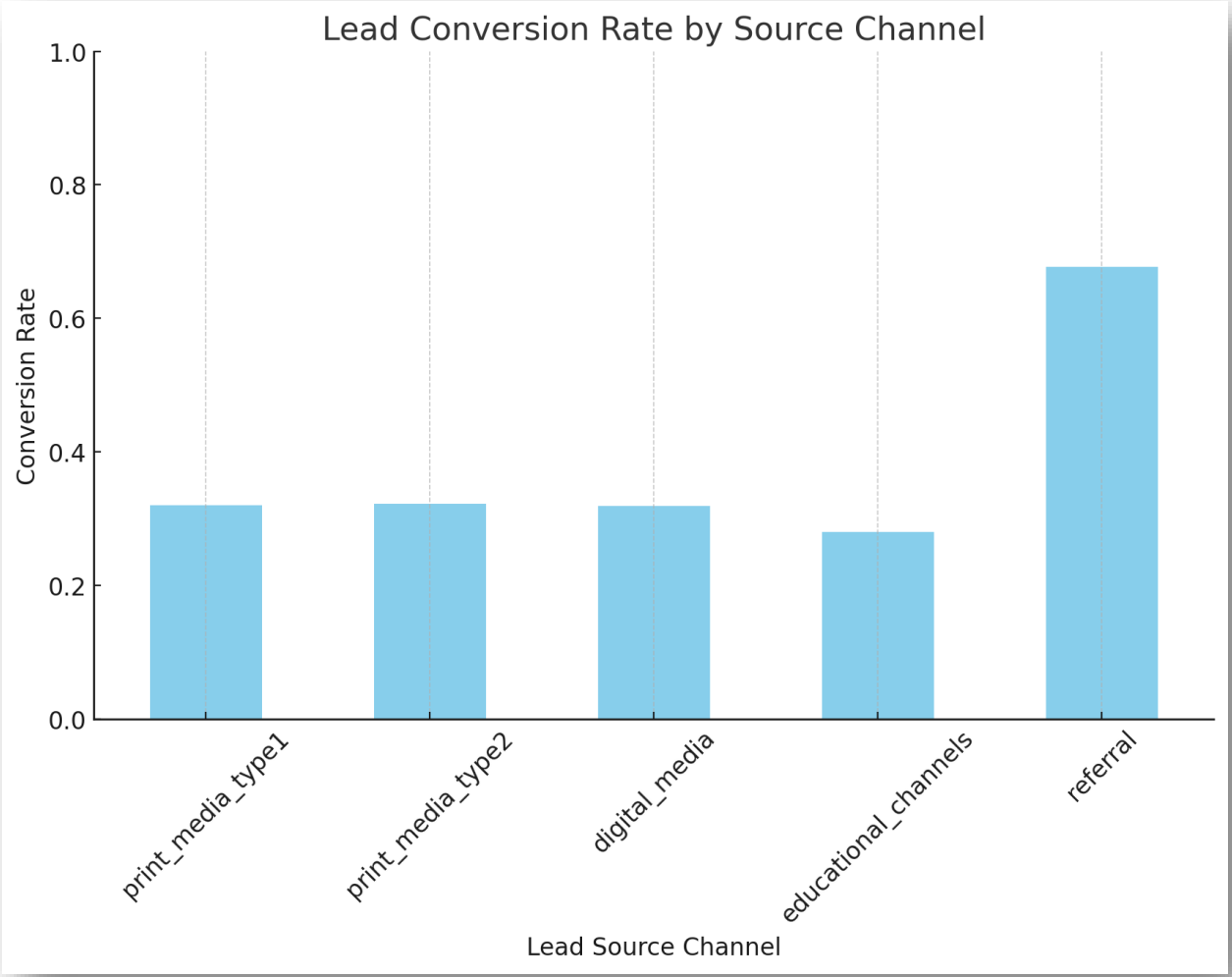


FIGURE 4

Insights:

1. Highest Conversion Rate:

- Referral: Leads coming through the referrals have the highest conversion rate at approx. 67.7% This indicates

that the referral leads are highly valuable and have a higher likelihood of converting compared to other channels.

2. **Moderate Conversion Rate:**

- Print media type 1: The conversion rate is around 32%
- Print media type 2: The conversion rate is approx. 32.2%
- Digital Media: The conversion rate is around 31.9%

3. **Lowest Conversion Rate:**

- Educational channel: Leads from the educational channel have the lowest conversion rate at approx. 27.9%
- The bar plot clearly shows that referral is the most effective channel for converting leads.
- Print media and digital media have similar conversion rates, indicating that these channels are equally active.
- Educational channels have a relatively lower conversion rate indicating potential areas for improvement.

Q5: People browsing the website on mobile application are generally required to create a profile by sharing their data before they can access additional information. Does having more details about a prospect increases the chances of conversion?

Answer 5: To analyze whether having more details about a prospect increases the chances of conversion, we will focus on following steps:

1. Compare conversion rates based on level of profile completion.
2. Generate visualization to show trends and patterns.
3. Insights based on the analysis.

♦ **Plot of Lead Conversion Rate by Profile Completion:**

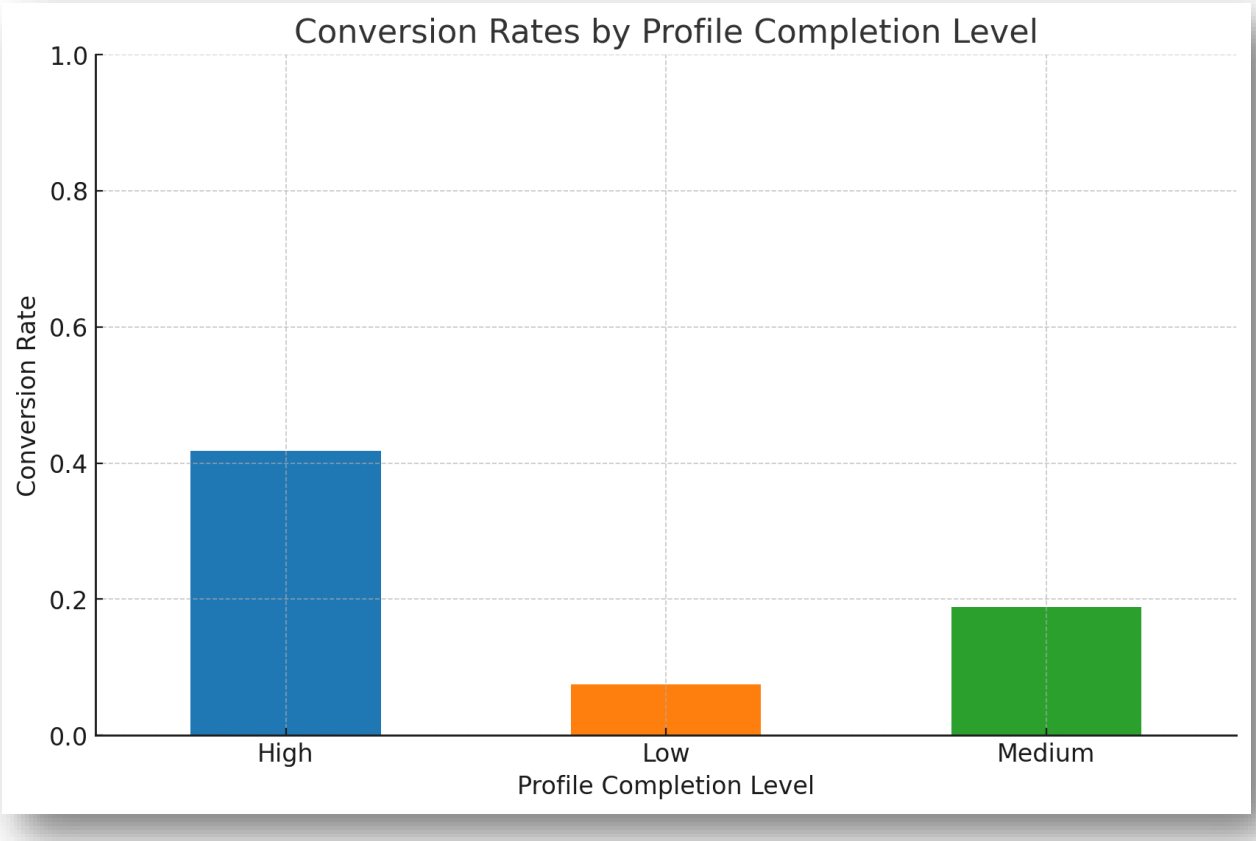


FIGURE 5

Insights:

1. Profile Completion and Conversion Rates:

- Users with a 'high' level of profile completion have the highest conversion rate. This suggests that more

detailed profiles are associated with a a greater likelihood of conversion.

- Users with a ‘medium’ level of profile completion have a lower conversion rate as compared to those with a ‘high’ level.
- Users with a ‘low’ level of profile completion have the lowest conversion rate.
- Encouraging users to complete their profiles can positively impact conversion rates.
- While profile completion is an important factor, other variables such as age, current occupation, and first interaction type (website or mobile app) might also influence the conversion rates.

❑ UNIVARIATE ANALYSIS

Now we will do further analysis.

♦ Plot: Observation on age:

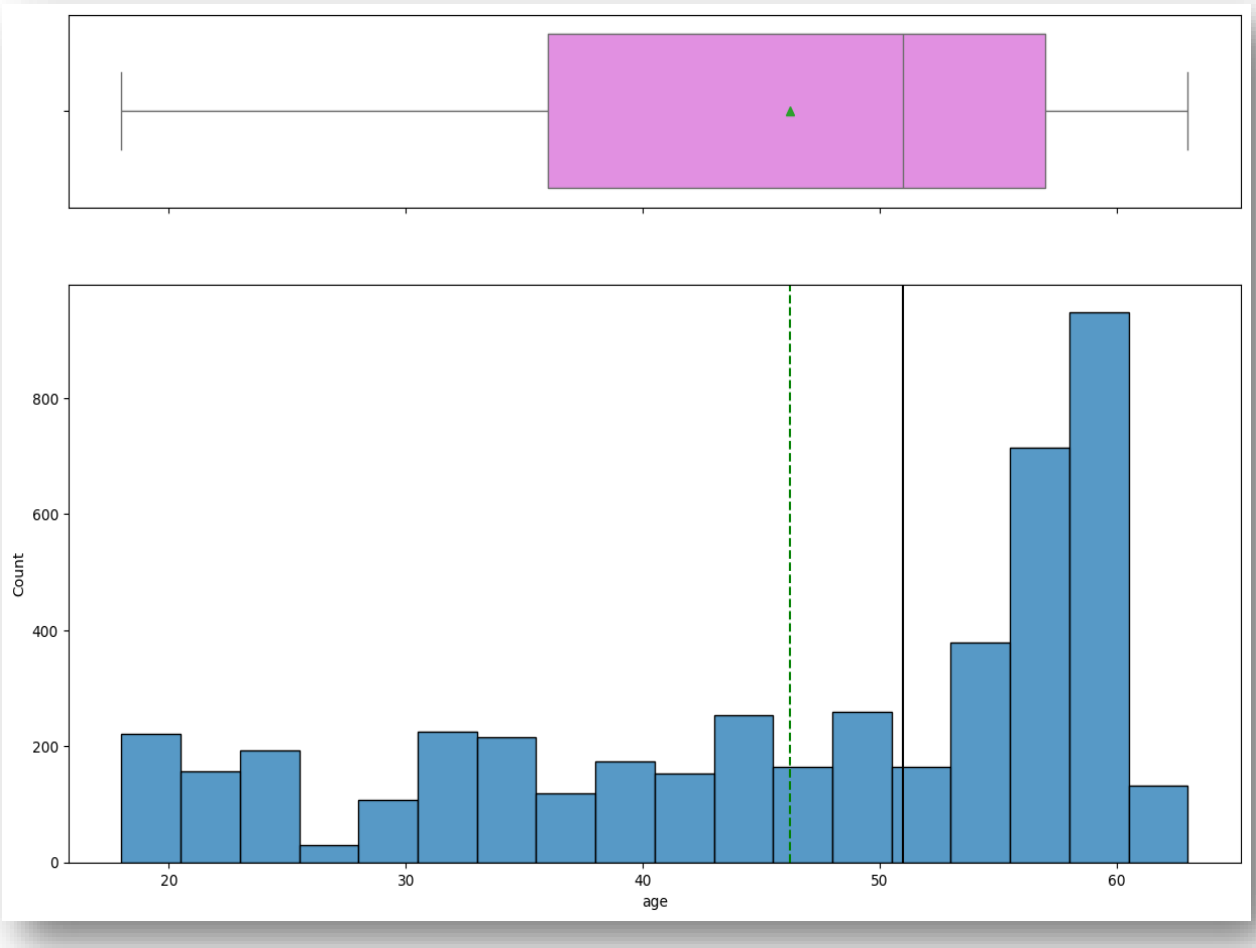


FIGURE 6

Insights on the Box Plot:

- The IQR extends from about 28 to 58 years indicating that the middle 50% of the ages fall within this range.
- The median age is around 53 years.
- The whiskers extend to approx. 10 to 62 years suggesting the range of data without outliers.
- There are no significant outliers present beyond the whiskers.

Insights on the Histogram:

- The age distribution is bimodal, with 2 peaks. One peak is around early 20's whereas second larger one occurs in late 50's to 60's.
- The drop in the number of leads can be seen in late 20's and early 30's.
- There are few leads in 30 to 35 age range
- The highest count of lead can be seen in the group of 60 to 63 years.
- The mean age is 46 years which is slightly lower than the median which is 51 years showing the distribution is right skewed.

♦ **Plot: Observation on website_visits:**

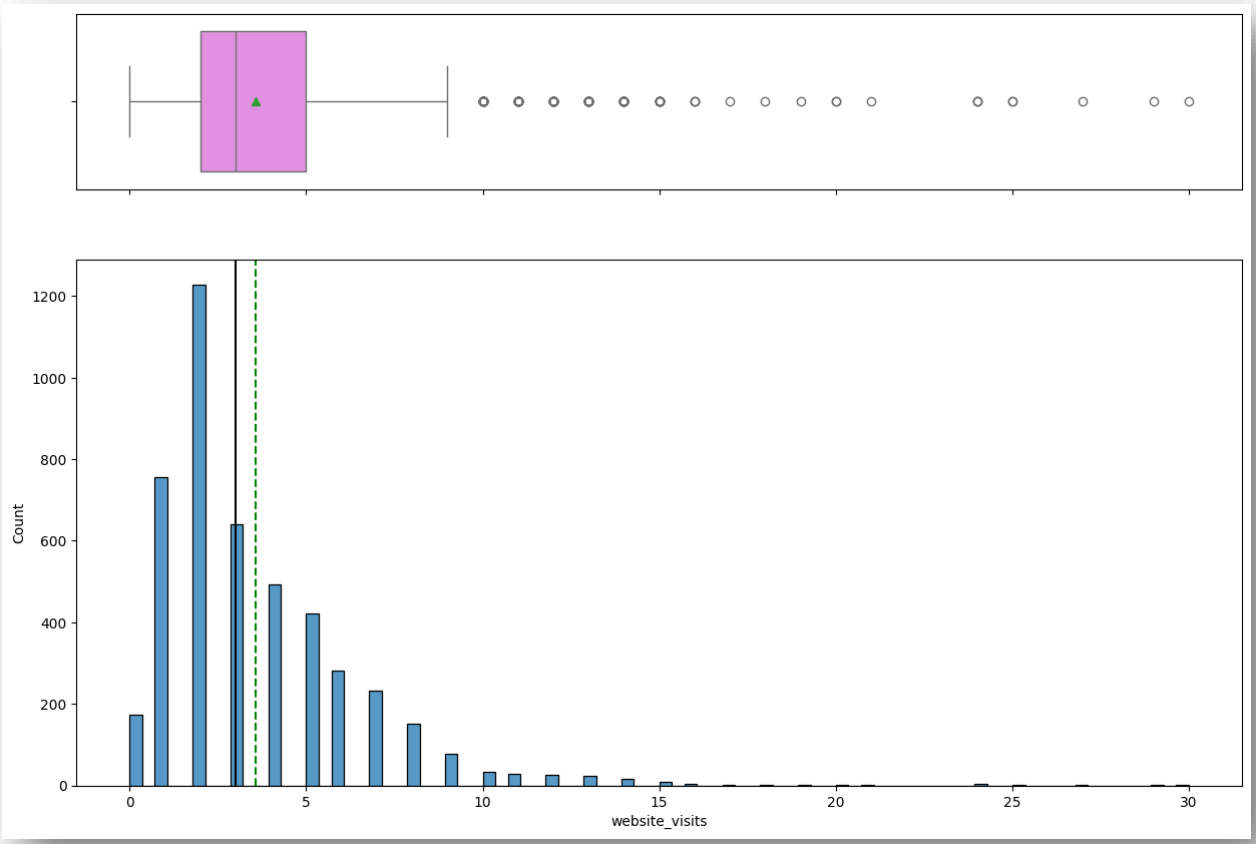


FIGURE 7

Insights on the Box plot:

- There are several outliers above the upper whisker with some leads visiting the website up to 30 times. These outliers
- The whiskers extend from 0 to 9 visits approx. Suggesting the range of data.
- IQR spans from approx. 2 to 5 visits.
- The median number of website visits is around 3 times.

Insights on the Histogram:

- There are few leads with more than 10 visits.
- The distribution is right skewed, with most leads having less visits.
- A sharp drop off can be seen after 5 visits. The high frequency of visits in this range suggests that the initial interactions are very crucial.

♦ Plot: Observation on time_spent_on_website:

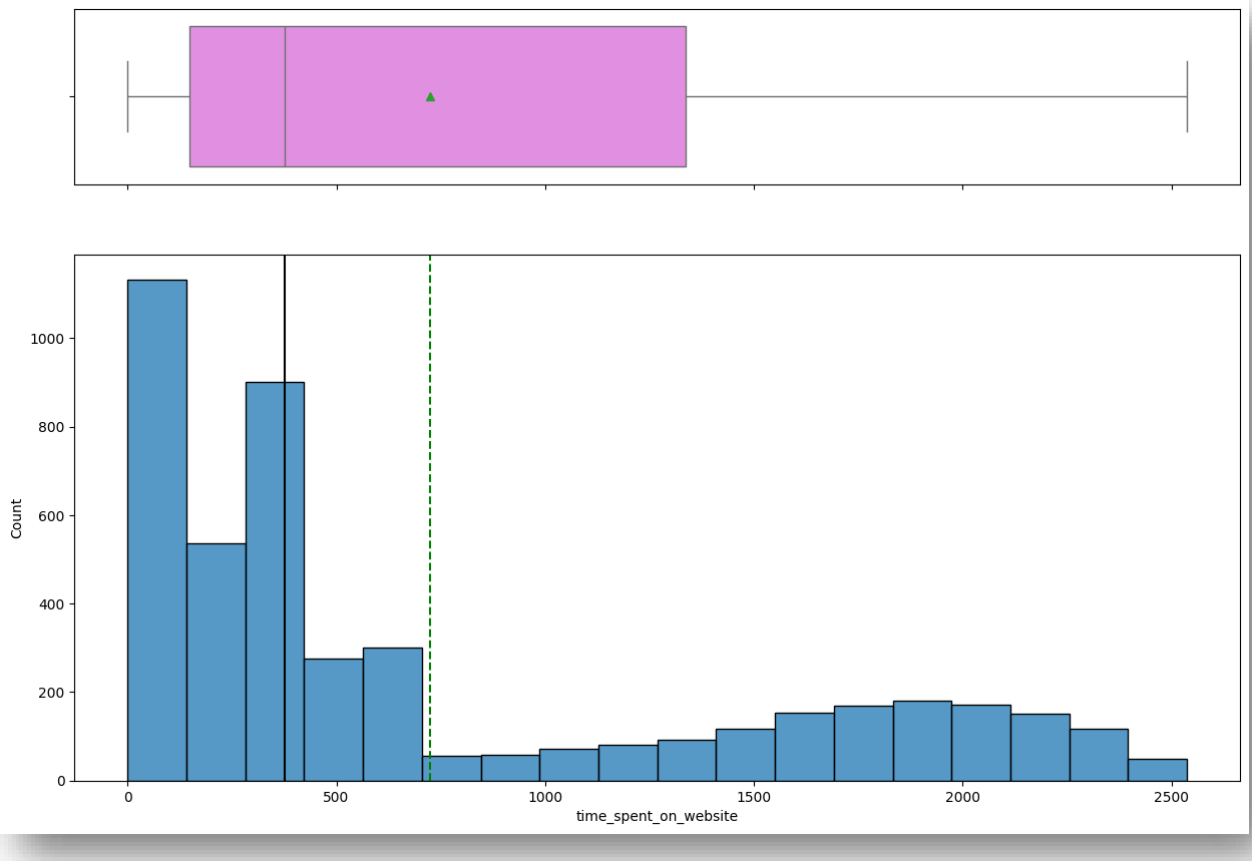


FIGURE 8

Insights based on the Box Plot:

- The median time spent on the website is around 376 seconds.
- The whiskers extend from 0 to 2537 seconds, showing the range of data.
- No significant outliers can be seen in the plot.
- The IQR ranges from approx. 149 to 1337 seconds.

Insights on the Histogram:

- The distribution is right skewed with majority of the leads spending less than 500 seconds on the website.
- Highest frequency is for leads spending 0 to 500 seconds on the website. This high frequency in the lower time range indicates the initial engagement on the website is brief for most of the leads.
- The mean time spent is 724 seconds which is higher than the median of 376 seconds.
- There is a noticeable drop in the frequency after 500 seconds, with small number of leads spending more time on the website.
- The histogram also shows that there is a significant number of leads who spend more than 1000 seconds, showing a deeper level of engagement suggesting that these leads might be more interested.

♦ Plot: Observation on page_per_views_per_visit:

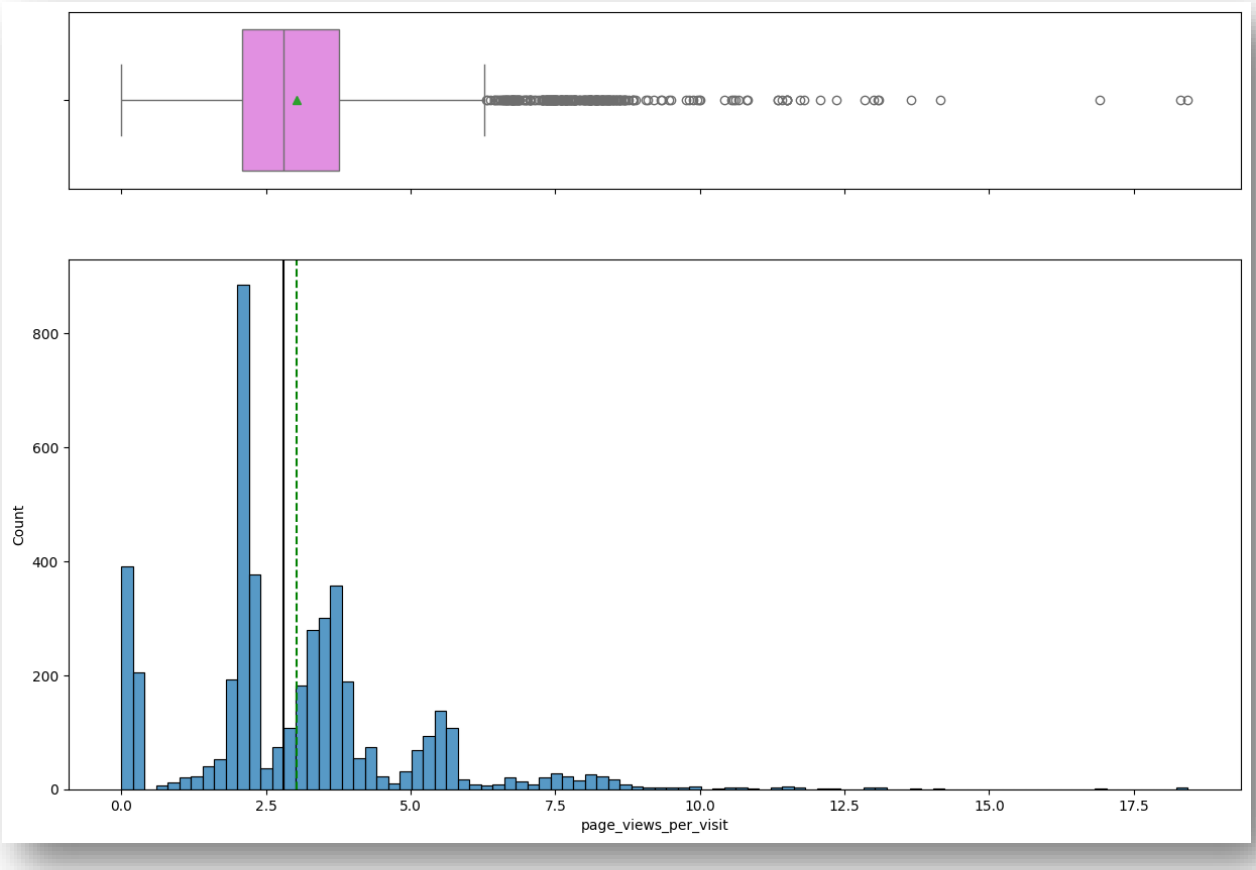


FIGURE 9

Insights based on the Box Plot:

- The median number of page views per visit is approx. 279
- The IQR ranges from 2.08 to 3.76 page views per visit.
- The whiskers extend from 0 to 18.43 pages per visit, showing the range of data.
- The distribution is right skewed with no significant outliers present.

Insights on the Histogram:

- Most leads have 2 to 3 page view per visit indicating that they explore few pages before leaving the website.
- The mean number of page views per visit is 3.03 which is higher than the median.
- The median number of page views per visit is 2.79, indicating slight right skew in the distribution.
- The highest frequency is observed around 2 to 3 page views per visit showing most leads visit a few pages on an average.
- There can be seen a noticeable drop in frequency beyond 5 page views per visit with small number of leads viewing more pages.

♦ **Plot: Observation on number_of_adults:**

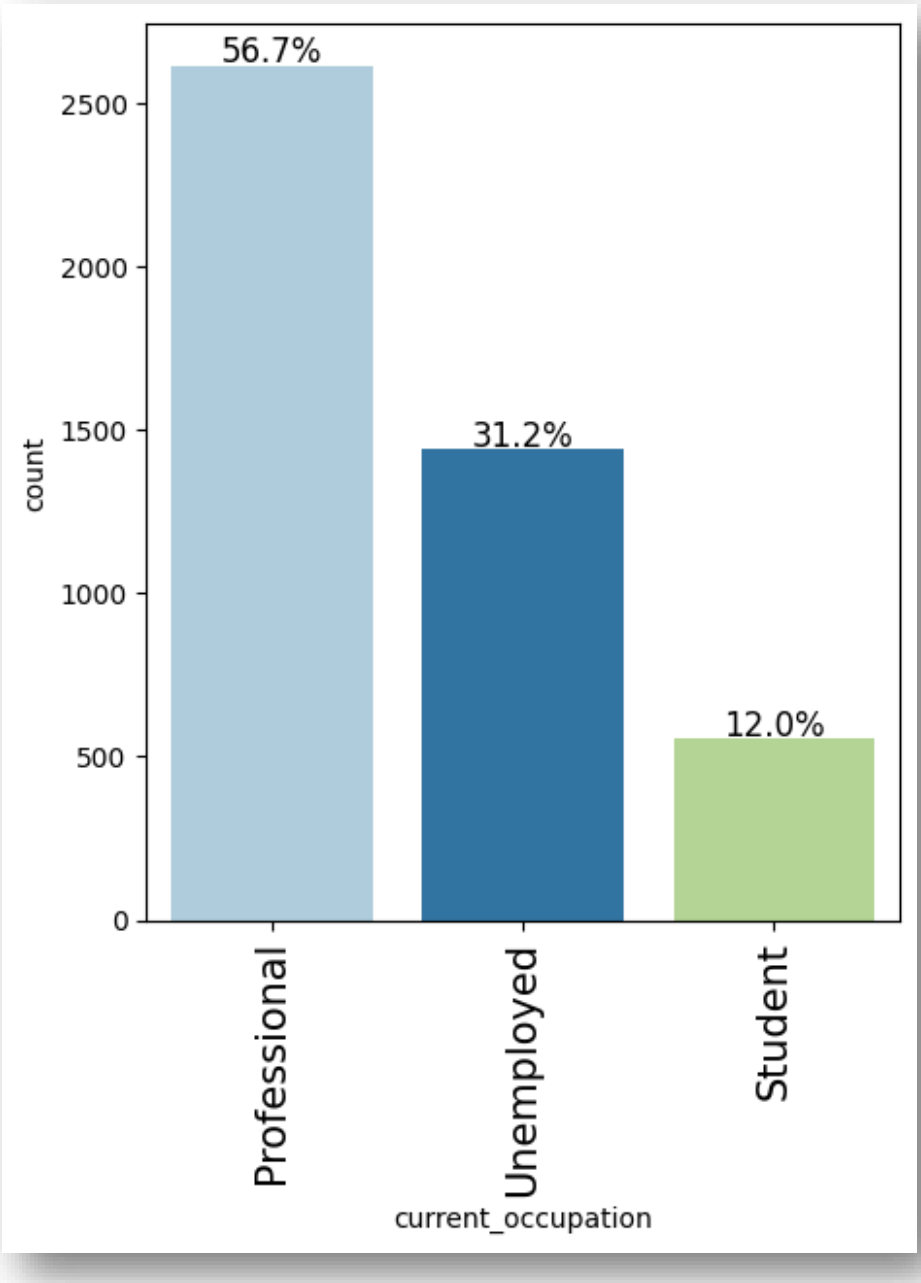


FIGURE 10

Insights based on the Bar plot:

- **Professional:** The majority of users fall under this category representing 56.7% of the total users. This

shows that more than half of the platform's user base consists of professionals.

- **Unemployed:** This category makes up 31.2% of the total users, suggesting that a significant portion of users are not currently employed, which may have connection for services that cater to job seekers.
- **Student:** Students category accounts for 12% of the total users. This smaller percentage shows less focus on student specific content and needs more targeted outreach to this section.

♦ **Plot: Observation on profile_completed:**

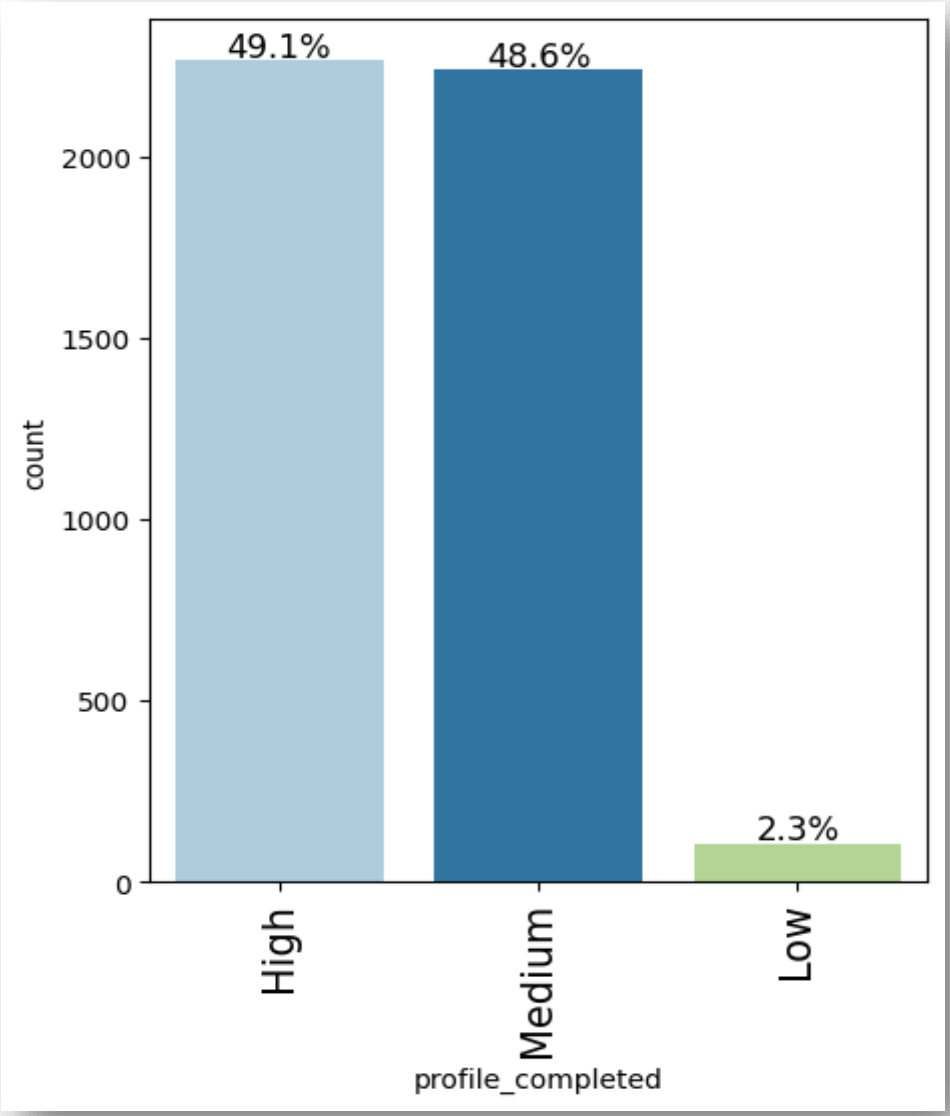


FIGURE 11

Insights based on the Bar plot:

- **High:** Prospects with high profile completion have the highest conversion rate i.e. 41.78%
- **Medium:** These prospects have lower conversion rate of 18.88%

- **Low:** These prospects have the lowest conversion rate of 7.48%
- We can say that having a highly completed profile increases the chances of conversion.

- **Plot: Observation on print_media_type1:**

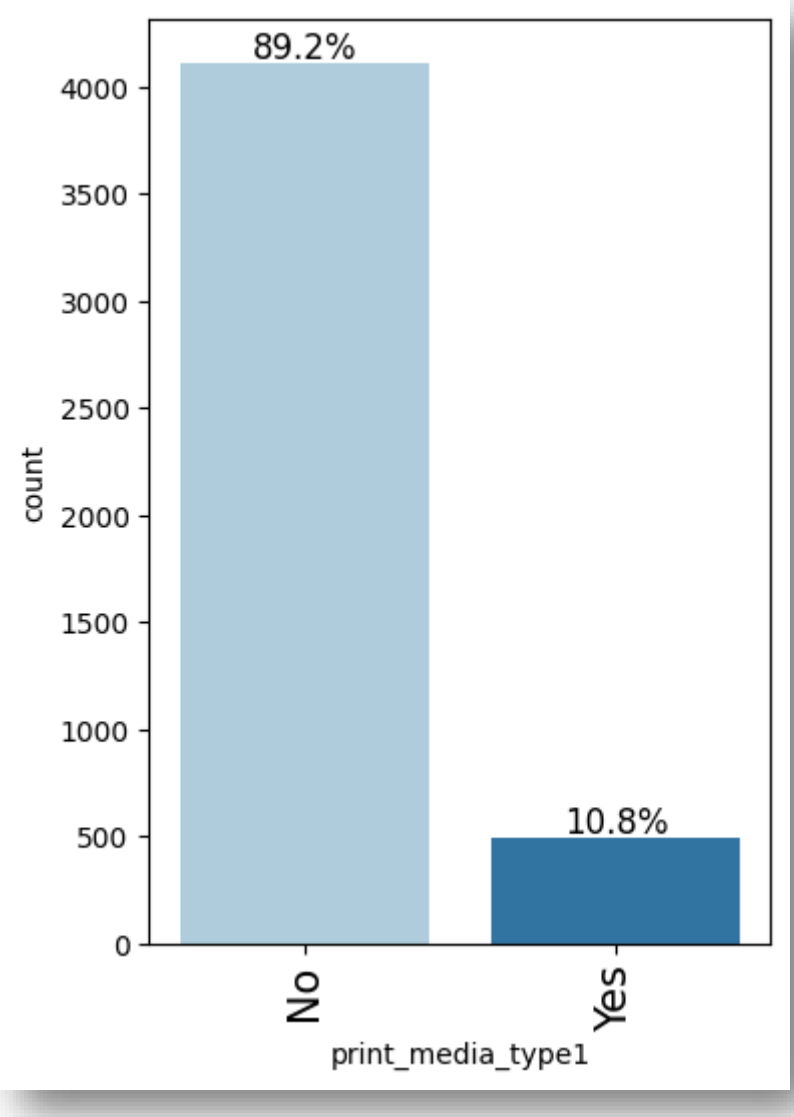


FIGURE 12

Insights based on the Bar plot:

- NO: 89.2% of prospects did not interact through print media type 1
- YES: Only 10.8% of prospects did interaction through print media type 1
- This plot suggests that majority of the leads 89.2% have not interacted with print media type 1, however even though a smaller percentage 10.8% did interact, the conversion rate for those who did interact is slightly higher than those who didn't.
- This indicated that while the reach of print media type 1 is limited, it can still be quite effective in converting leads.

- **Plot: Observation on print_media_type2:**

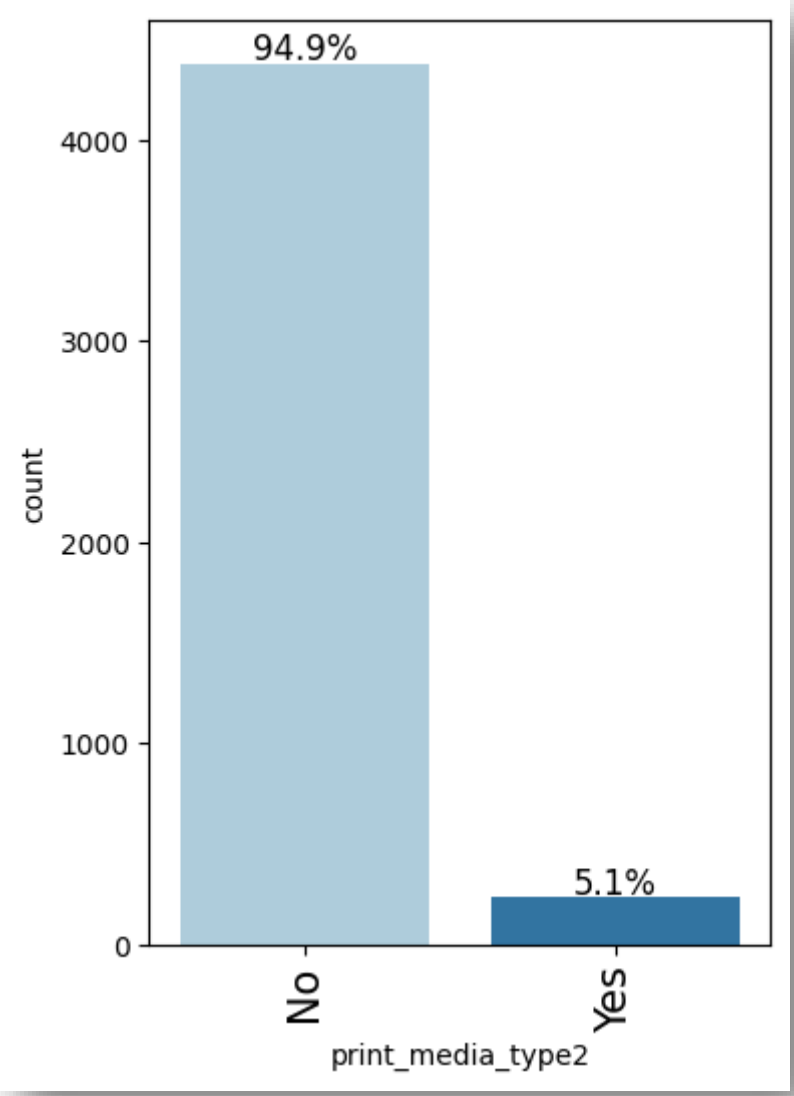


FIGURE 13

Insights based on the Bar plot:

- NO: 94.9% of the prospects did not interact through print media type 2
- YES: Only 5.1% of the prospects did interact with print media type 2
- Most of the prospects did not interact through print media type 2

- Despite lower engagement the conversion rate for those who did interact was higher as compared to those who don't.
- This suggests that print media type 2 is reaching a smaller audience but is relatively more effective in converting those who do engage.

♦ **Plot: Observations on digital_media:**

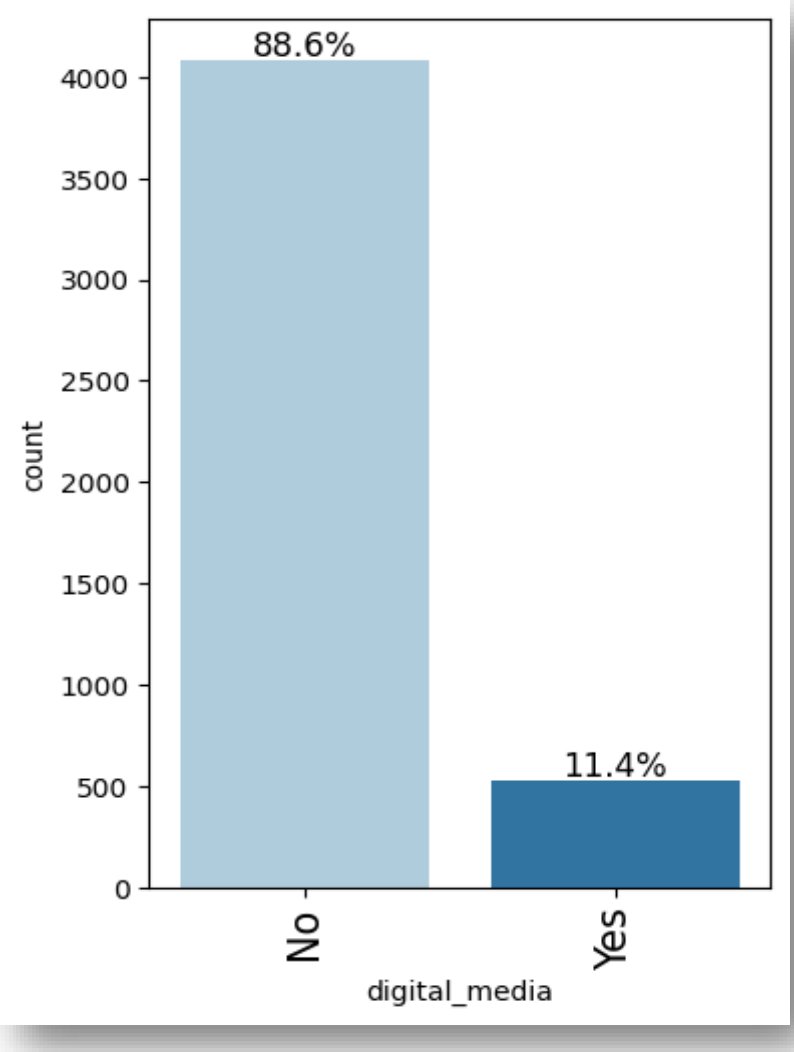


FIGURE 14

Insights based on the Bar plot:

- Most prospects, 88.6% do not use ‘digital media’ for interaction.
- Only 11.4% of the prospects can be seen using ‘digital media’ for interaction.
- Many prospects are not engaging with the digital tools. This indicates lack of access, interest or awareness.

♦ **Plot: Observations on educational_channels:**

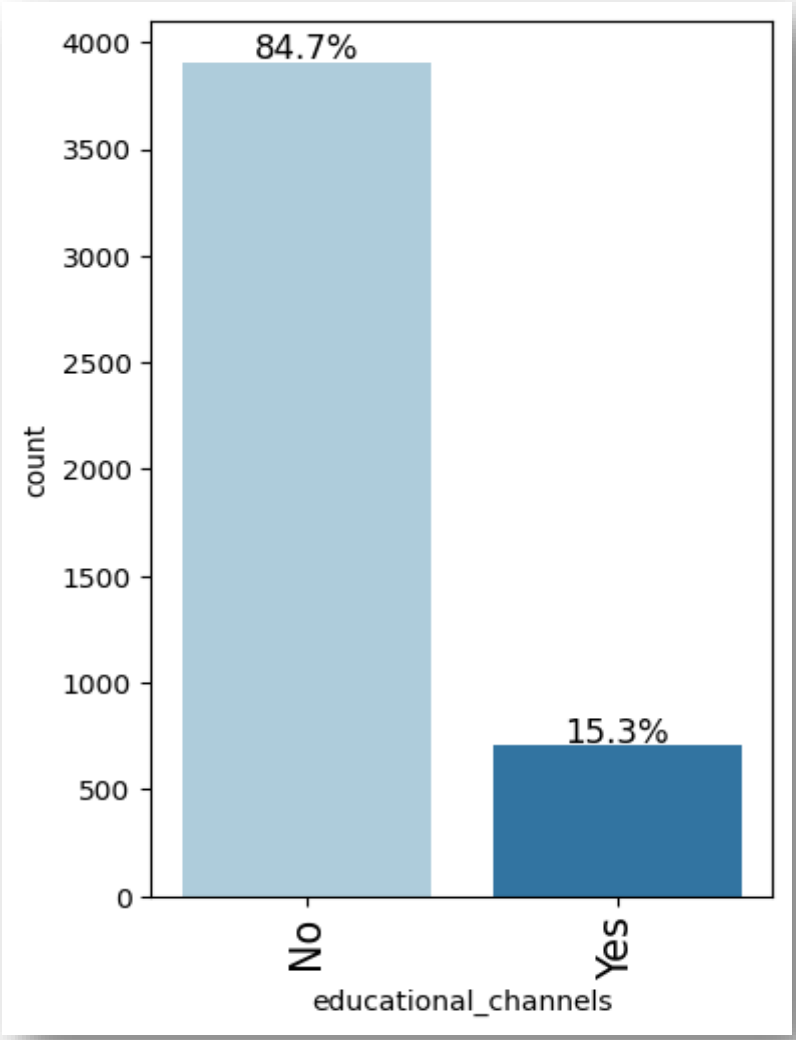


FIGURE 15

Insights based on the Bar plot:

- The bar plot reveals a significant disparity between the user engaging with educational channels and those who do not.
- A large proportion of leads do not use educational channels for the interaction medium.

- Only 13.3% of the leads corresponds to the ‘Yes’ category, suggesting that a relatively small portion of the leads engage with educational channels.
- Most of the data corresponds to ‘No’ category indicating that a significant portion of leads do not engage with education channels.

♦ **Plot: Observations on referral:**

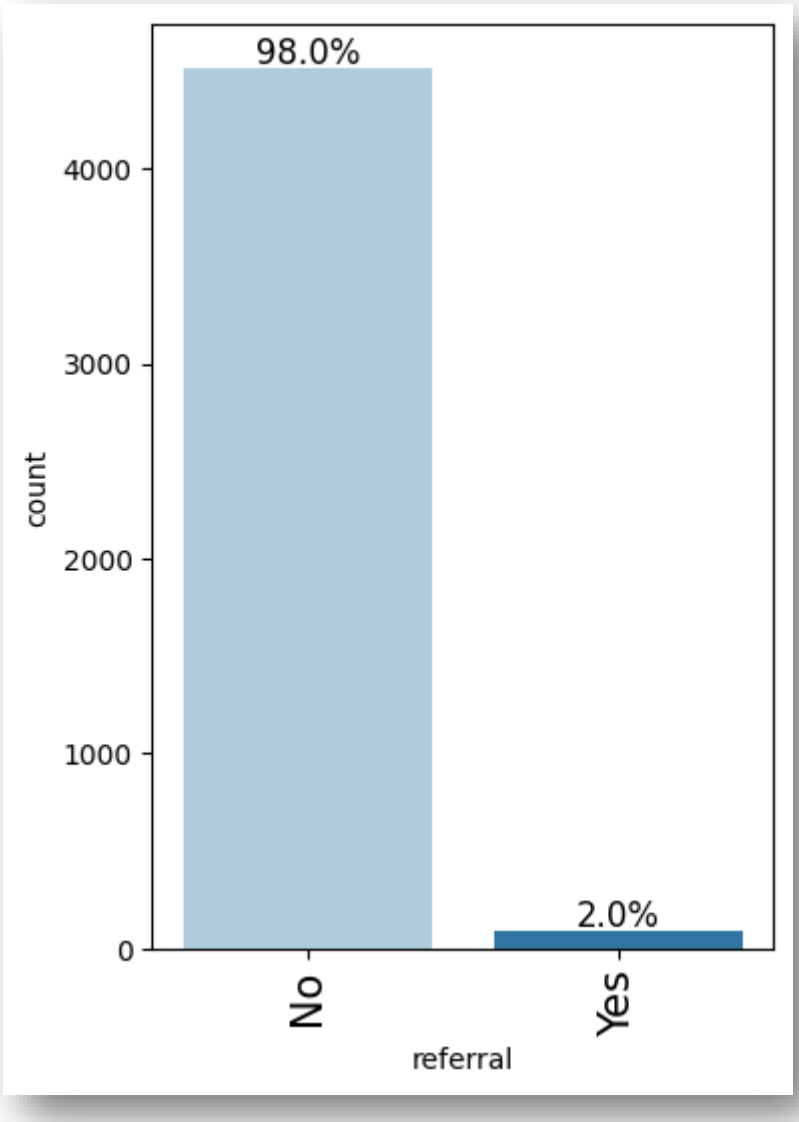


FIGURE 16

Insights based on the Bar plot:

- The plot shows that a vast majority of leads (users) about 98% did not come through a referral, while only 2% were referred by others.
- The extreme low percentage of referrals suggests that the current referral program is not being effectively utilized by the leads.

- The data is highly imbalanced with a significant higher count of non-referrals compared to referrals.

♦ Plot: Observations on Status:

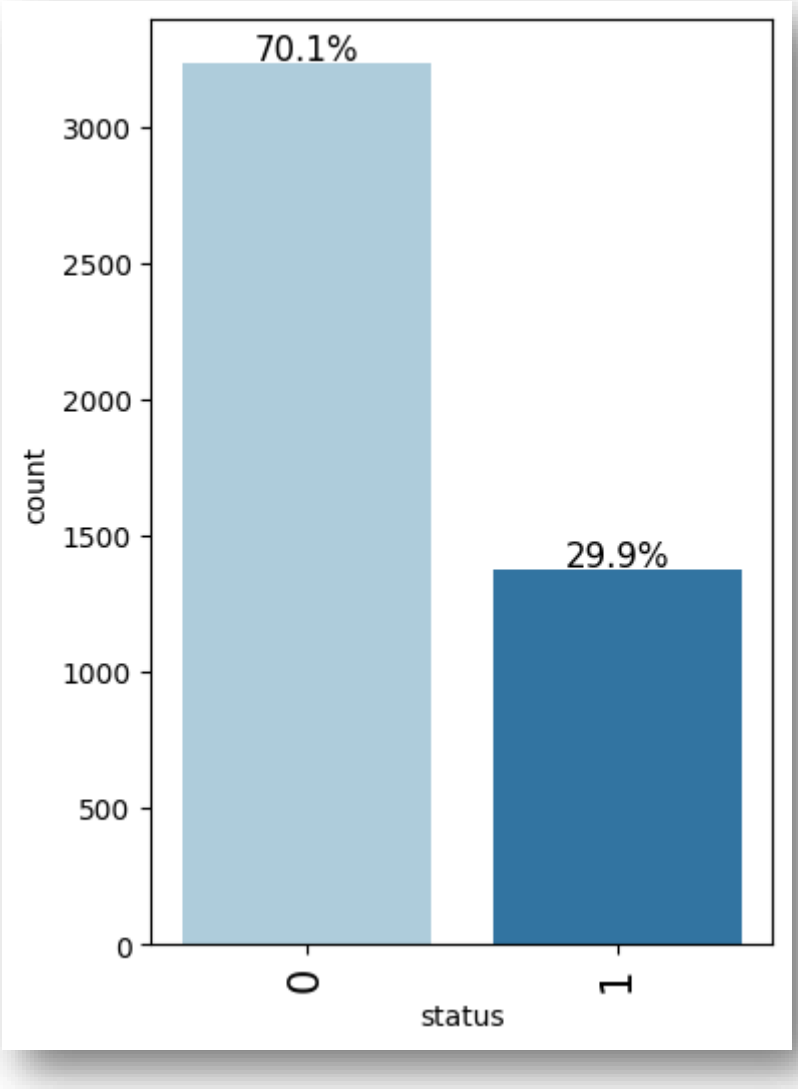


FIGURE 17

Insights based on the Bar plot:

- The majority class (status 0) is more than twice as large as the minority class (status1)
- There is a notable imbalance in the distribution with a higher proportion of observations falling into ‘0’ category as compared to ‘1’ category.

❑ BIVARIATE ANALYSIS

♦ Analysis on the correlation Heatmap:

FIGURE 18

Insights:

- There is a very weak negative correlation between the ‘age’ and the ‘website visits’ and ‘page views per visit’, suggesting that age does not significantly impact these metrics.
- A weak positive correlation exists between the ‘age’ and ‘status’ indicating a slight relationship.
- There is a weak positive correlation between ‘website visits and ‘page views per visit’ indicating that as the website visits increase both time spent, and page views tend to slightly increase.
- A very weak negative correlation can be observed between ‘website visits’ and ‘status’.
- There is a moderate positive correlation between ‘website visits’ and ‘status’ indicating that as the time spent on the website increases ‘status’ tend to increase too.
- Weak positive correlation can be seen between ‘website visits’ and ‘page views per visit’.
- No significant correlation can be seen between ‘page views per visit’ and ‘status’ suggesting that number of page views per visit does not impact status.
- A moderate positive correlation can be observed between ‘status’ and ‘website visits’ indicating that higher time spent on the website is associated with a higher status.

- The moderate correlation between status and time spent on the website shows that users who spend more time on the website are more likely to have a higher status. This shows that engagement on the website is a factor in achieving higher status.
- Age shows negligible correlation with most of the variables suggesting that it does not significantly influence website usage patterns or status.
- Weak correlation between website visits and other variables indicates a slight tendency of high engagement with more visits.

❑ 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.

♦ **The stacked bar plot of Current Occupation vs Status:**

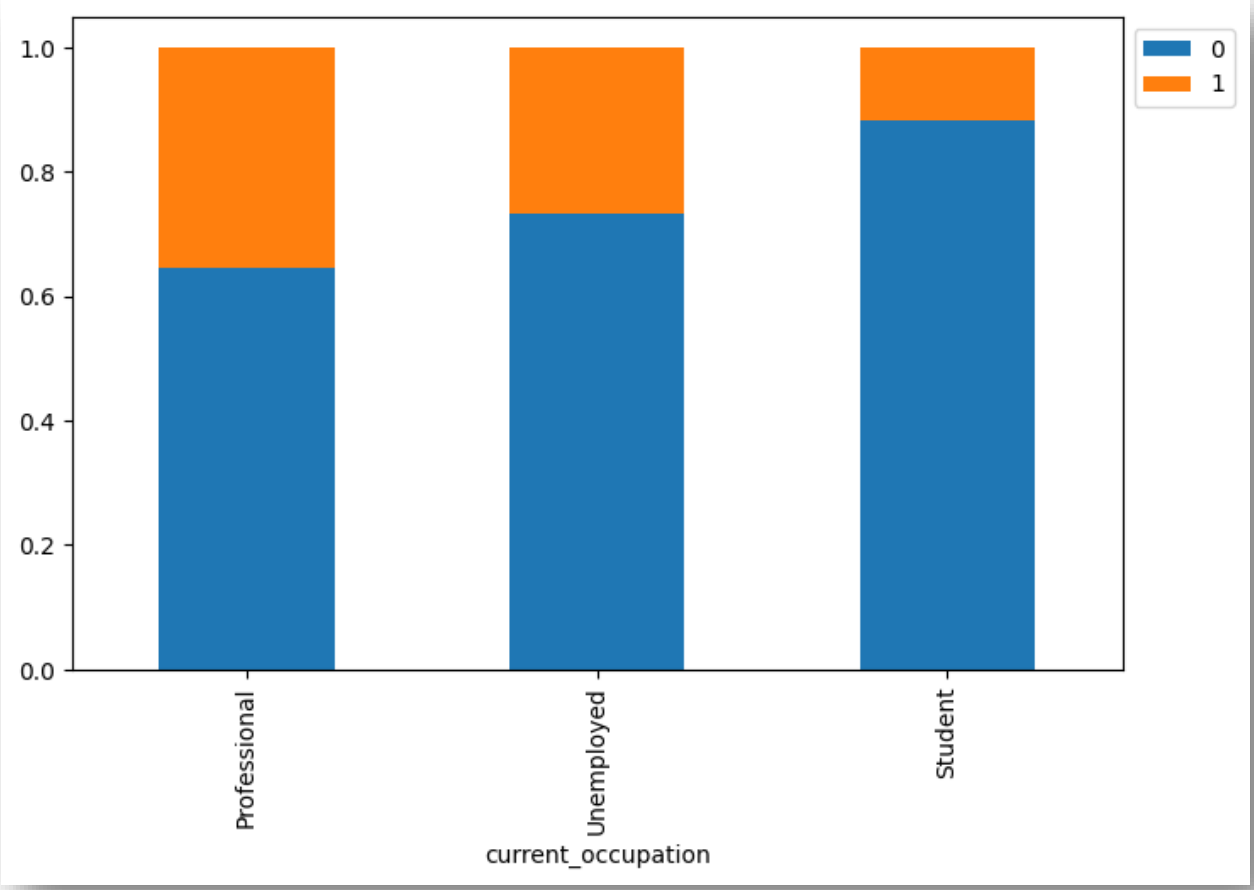


FIGURE 19

Insights:

1. **Professional:**

- Many of the prospects are professional with 2616 individuals.
- Among professionals 64.5% have status 0 while 35.5% have status 1. This suggests a significant proportion of professionals have a higher status 1

2. **Unemployed:**

- There are 1441 unemployed prospects.
- Among the unemployed, 73.4% have status 0

3. Student:

- There are 555 students in the dataset.
- Among students, 88.3% have status 0 and only 11.7% have status 1
- Students have the lowest proportion of individuals with status 1 showing that being student is least associated with higher status.

♦ Box plot of Age by Current Occupation:

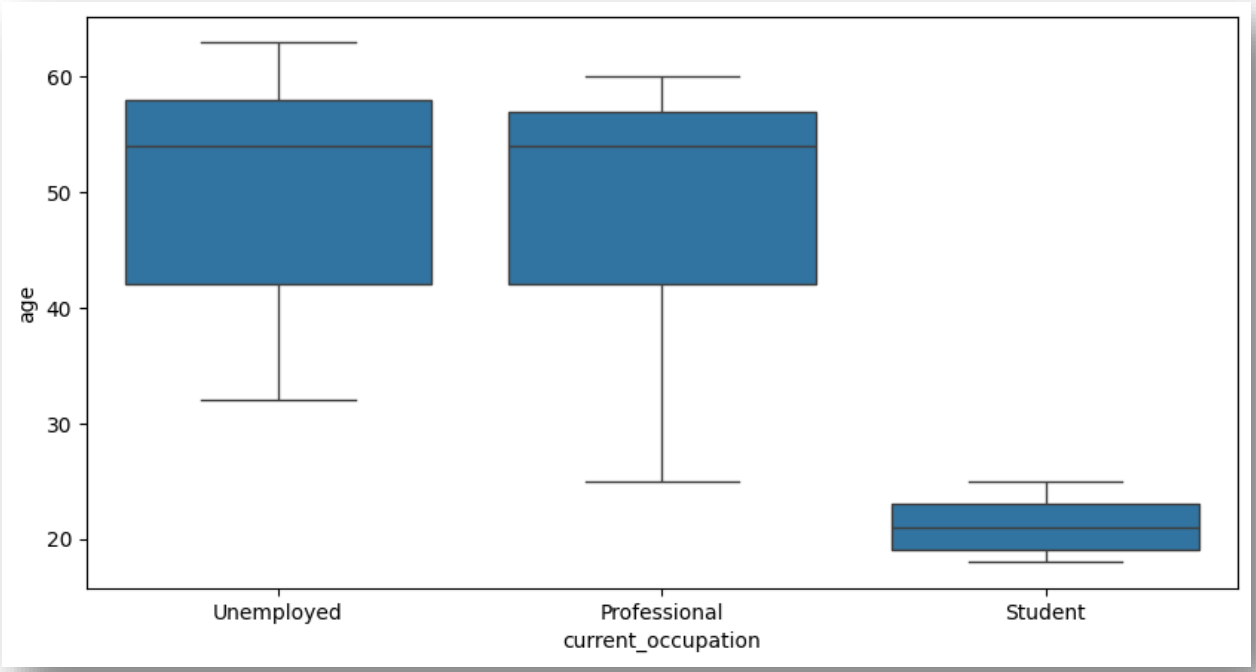


FIGURE 20

Insights based on the Box Plot:

1. Unemployed:

- The plot shows that the median age is around 55 years.
- The IQR ranges between 45 to 58 years.
- The distribution shows some variability with a slight skew towards younger ages.

2. Professional:

- The median age is around 52 years.
- The IQR ranges between 45 – 50 years.
- This group also has a similar spread to the unemployed group but with a slightly lower median.

3. Student:

- It can be seen in the plot that the median age is around 21 years.
- The IQR ranges from 20 – 22 years.
- The age distribution is much tighter and lower as compared to ‘unemployed’ and ‘professional’ groups, indicating that most students are young adults.

♦ Summary Statistics by Current Occupation:

	count	mean	Std. Dev.	Min.	25%	50%	75%	Max.
Current Occupation								
Professional	2616.00	49.34	9.89	25.00	42.00	54.00	57.00	60.00
Student	555.000	21.14	2.00	18.00	19.00	24.00	23.00	25.00
Unemployed	1441.00	50.14	9.99	32.00	42.00	54.00	58.00	63.00

TABLE 2

Insights based on the Table 1:

1. Age Distribution:

- 'Students' are significantly younger with a mean age of 21.14 years and a much lower standard deviation showing a tight age range.
- 'Professional' and 'unemployed' have a similar age distribution with mean ages around 49-50 years and median age of 54 years.

2. Age Variability:

- The 'professional' and 'unemployed' groups have more variability in the ages as compared to the students, as observed from the higher standard deviation (9.89 & 9.99 respectively) compared to students (2.00)

- This shows a more diverse age range among professional and unemployed individuals.

3. Inter Quartile Range (IQR):

- The IQR for professional (42 to 57 years) and unemployed (42 to 58 years) is broader than for students (19 to 23 years)
- This confirms that the students have a narrower age range.

4. Extremes in Age:

- The unemployed group has the highest maximum age (63 years), followed by professionals (60 years) and students (25 years).

♦ Stacked Bar Plot of First Interaction by Target:

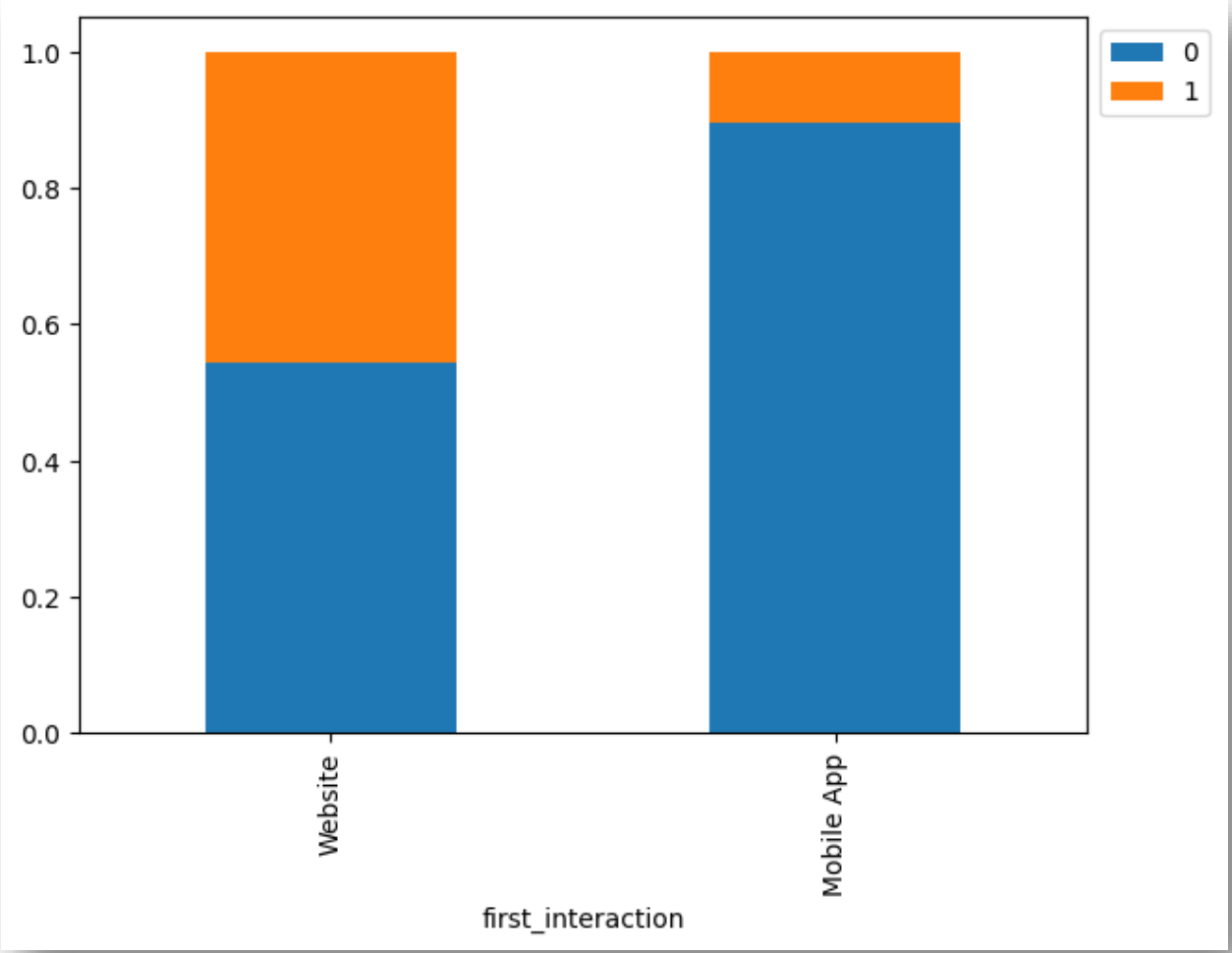


FIGURE 21

Insights based on the Stacked Bar Plot:

1. Website:

- Target 0: Nearly 55%
- Target 1: Nearly 45%
- More leads with their interaction on the website did not get converted.

2. Mobile App:

- Target 0: Nearly 80%
- Target 1: Nearly 20%
- A higher percentage of leads who used mobile app as their first interaction medium did not get converted as compared to those who did.

♦ Distribution & Box Plots of Time spent on website by Target:

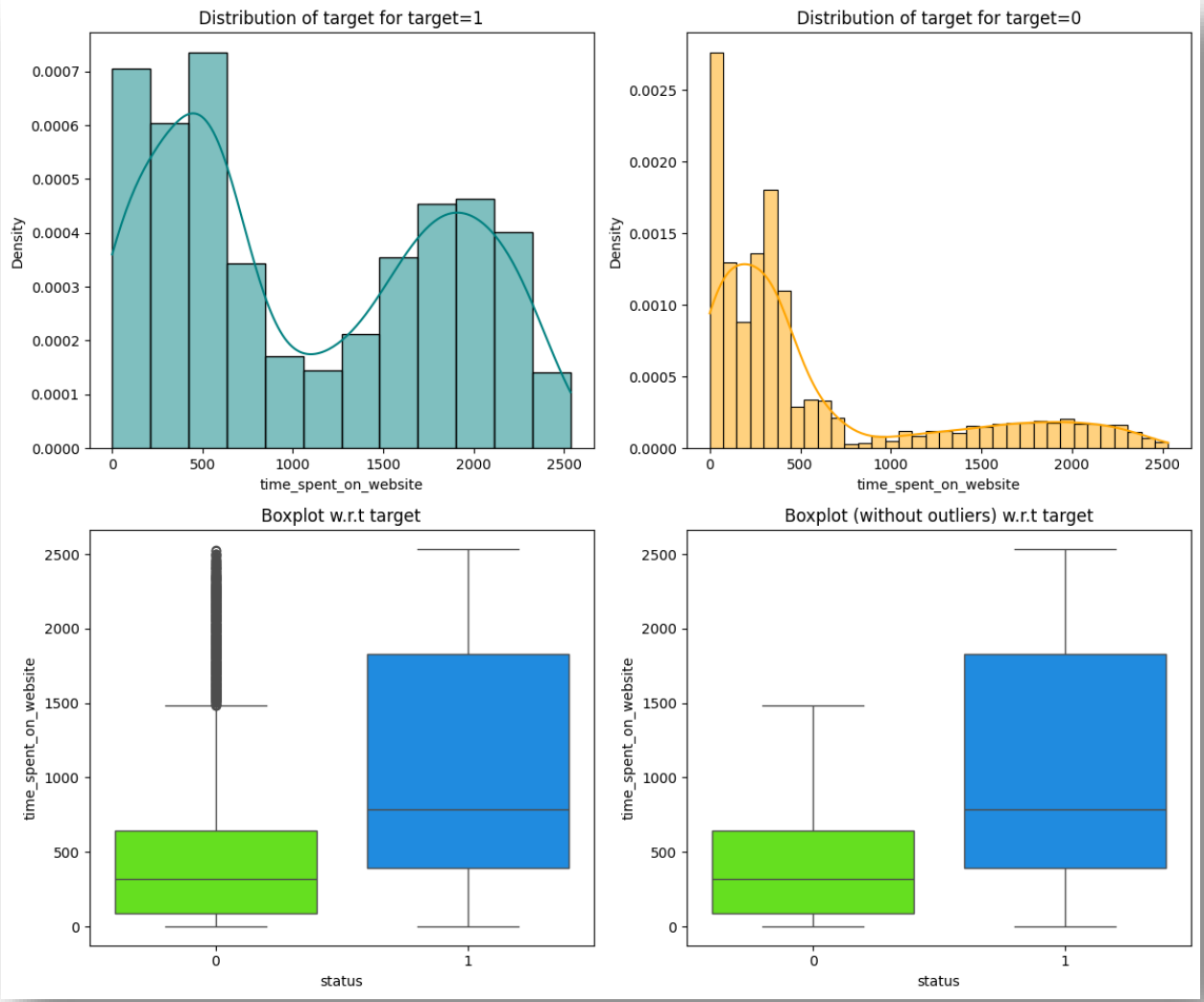


FIGURE 22

Insights based on the Distribution and Box plot:

1. Density Plots:

- Target 1:

a) The distribution shows a few peaks with most users spending between 0 - 500 seconds and 1500 – 2000 seconds on the website.

b) There is a significant downfall between these two peaks.

- Target 0:

a) The distribution is heavily skewed towards the lower end, with most users spending less than 500 seconds on the
b) website.
c) There is a long tail extending towards higher times, but with much lower density.

2. Box Plots:

- With outliers:

a) Target has median time of nearly 200 seconds with a larger spread and many outliers.
b) Target 1 has median time of nearly 1200 seconds with a higher overall spread and fewer outliers.

- Without outliers:

a) Target 0 shows a median time of nearly 200 – 400 seconds with a tight spread.
b) Target 1 shows a higher median time of 1000 – 1800 seconds with a wider spread.
c) A clear comparison can be observed between the two target categories because of the absence of the outliers.

♦ **Distribution plot of Website Visits by Target Status:**

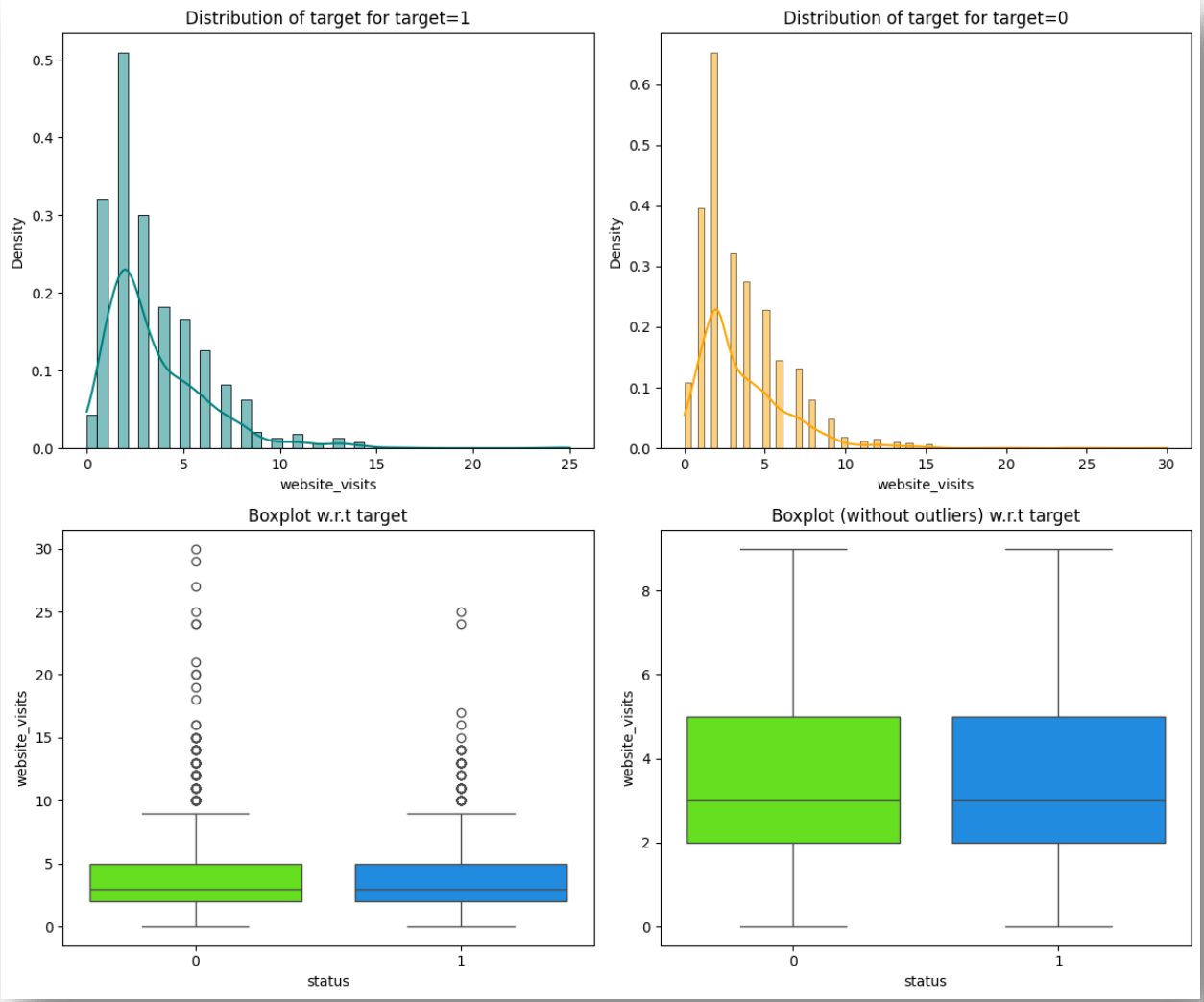


FIGURE 23

Insights:

- Density Plots:

a) The density plot shows that most leads with status 1 (interested) have fewer website visits with a peak at around 2 visits.

b) The distribution has a long tail suggesting some leads have noticeably more visits.

c) Similarly leads with status 0 (not interested) also tend to have fewer visits but with a slightly higher peak at around 1 visit.
d) The long tail is less dense compared to target 1

- Box plots:

a) The box plot shows that the median number of website visits for both target statuses 0 and 1 is around 3 to 4 visits.
b) The box plot also reveals a slightly higher median number of visits for target 1 indicating that leads with target 1 tend to visit the website more frequently.
c) The IQR and spread of the data are similar for both the groups.
d) Several outliers are also present.

♦ Distribution plot of Page views per visit by Target Status:

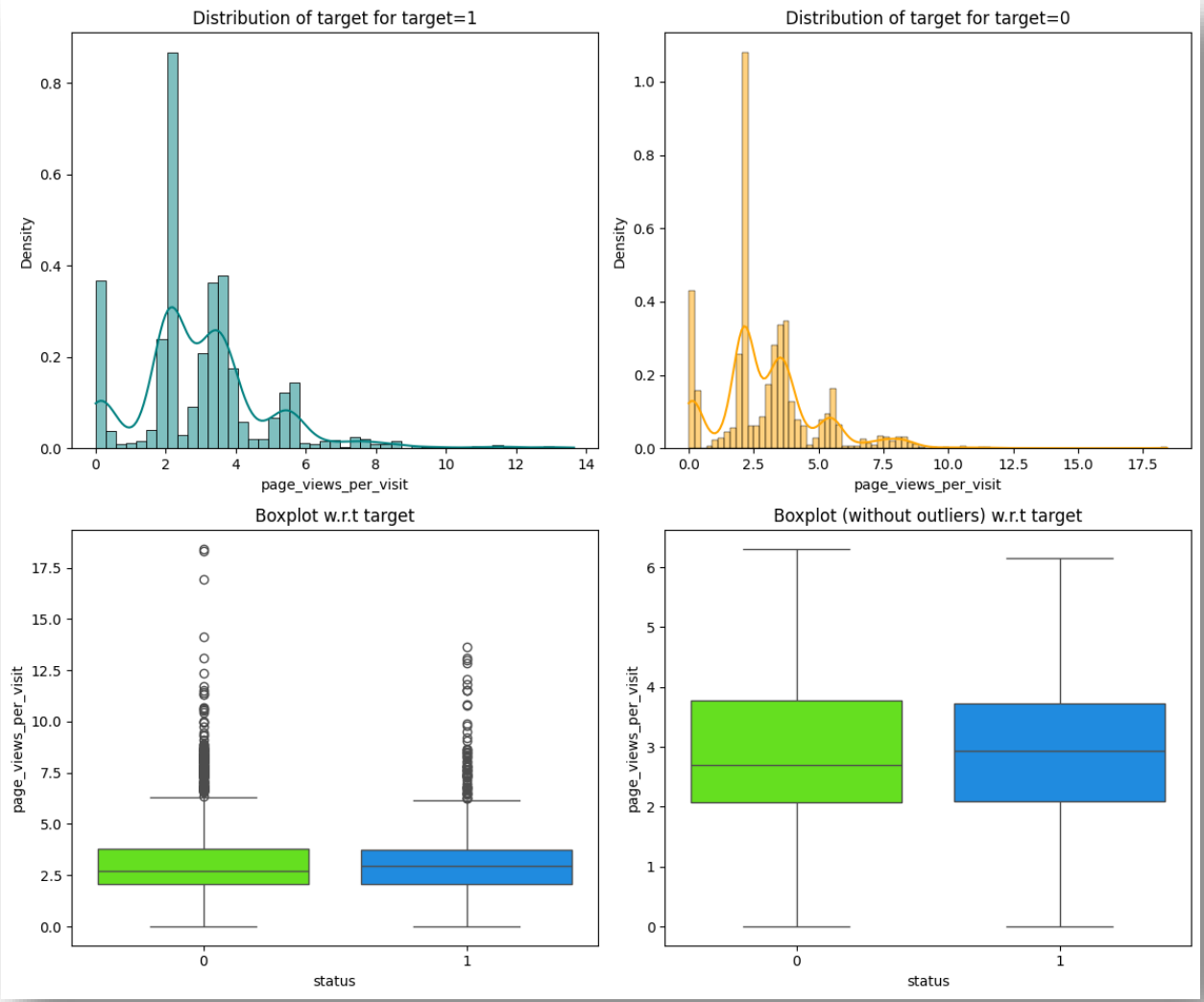


FIGURE 24

Insights:

- Density Plots:

a) The density plot for target 0 shows concentrated page views around 2-3 visits but with a peak at 2 page views per visit.

b) The spread is slightly wider as compared to target 1

c) The density plot for target 1 shows that mostly leads with status 1 have around 2 –3 page views per visit with significant peaks at 2 and around 4
d) There is a long tail suggesting a wide range of page views.
e) Both target 0 and target 1 have a similar distribution of website visits with a high concentration of leads making less visits. However, leads with target 1 have a slightly higher median number of visits, showing they are more engaged and more likely to get converted.

- Box Plots:

a) The box plot shows that the median page views per visit are slightly higher for leads with status 1
b) The IQR is quite similar for both the groups.
c) There are several outliers present in the box plot indicating some leads have significantly higher page views.

♦ Stacked Bar Plot of Profile completed and Status:

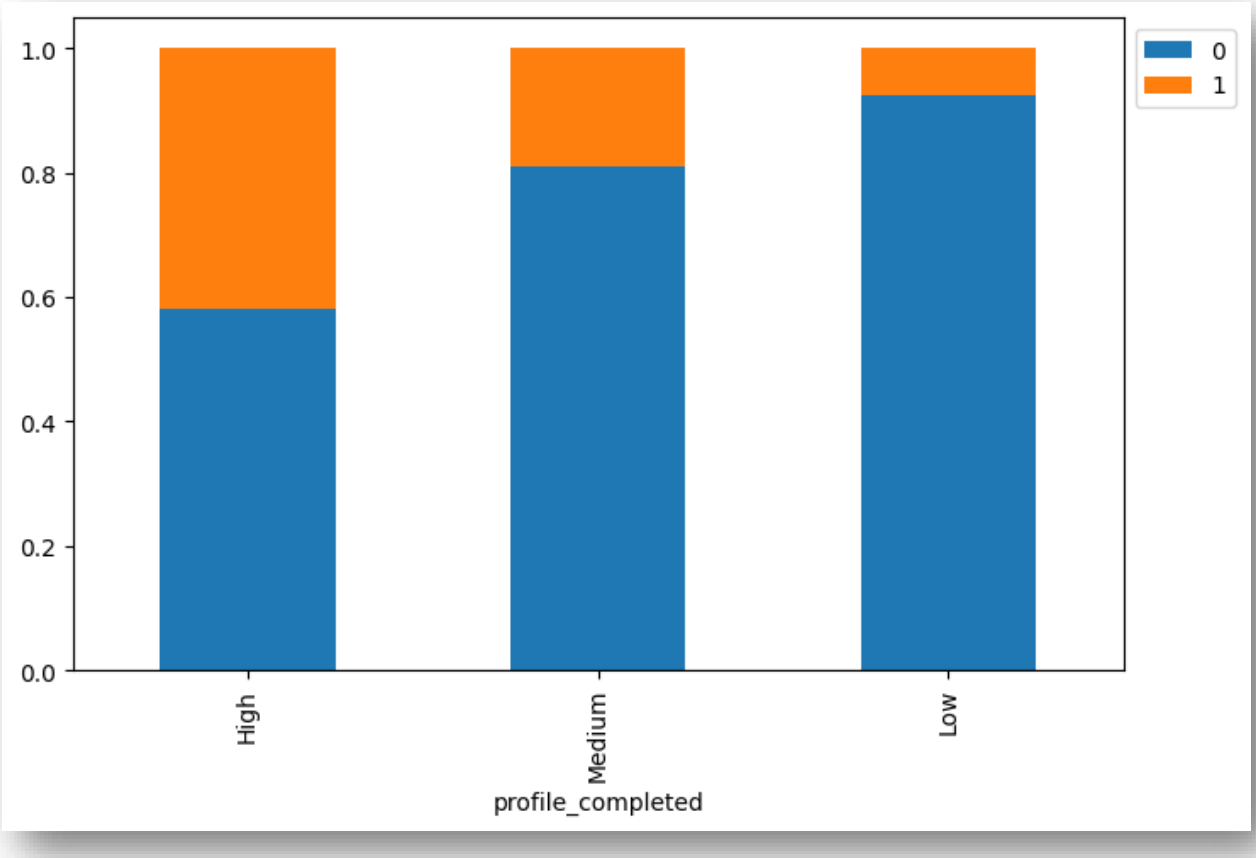


FIGURE 25

Insights:

- The bar plot shows the distribution of profile completion levels (High, medium, low) for both the target status 0 and 1
- Leads with high profile completion are more likely to have status 1 with a nearly equal split between status 0 and status 1
- For medium profile completion, the distribution is more balanced but leads with target 0 is still predominate.
- For low profile completion, most leads have target status 0, and very few having target status of 1.

- There is a clear correlation between profile completion and target status. Leads with high profile completion are more likely to have a target status of 1, suggesting more complete profiles are associated with higher conversion rates.

♦ Stacked Bar Plot of Last Activity and Status:

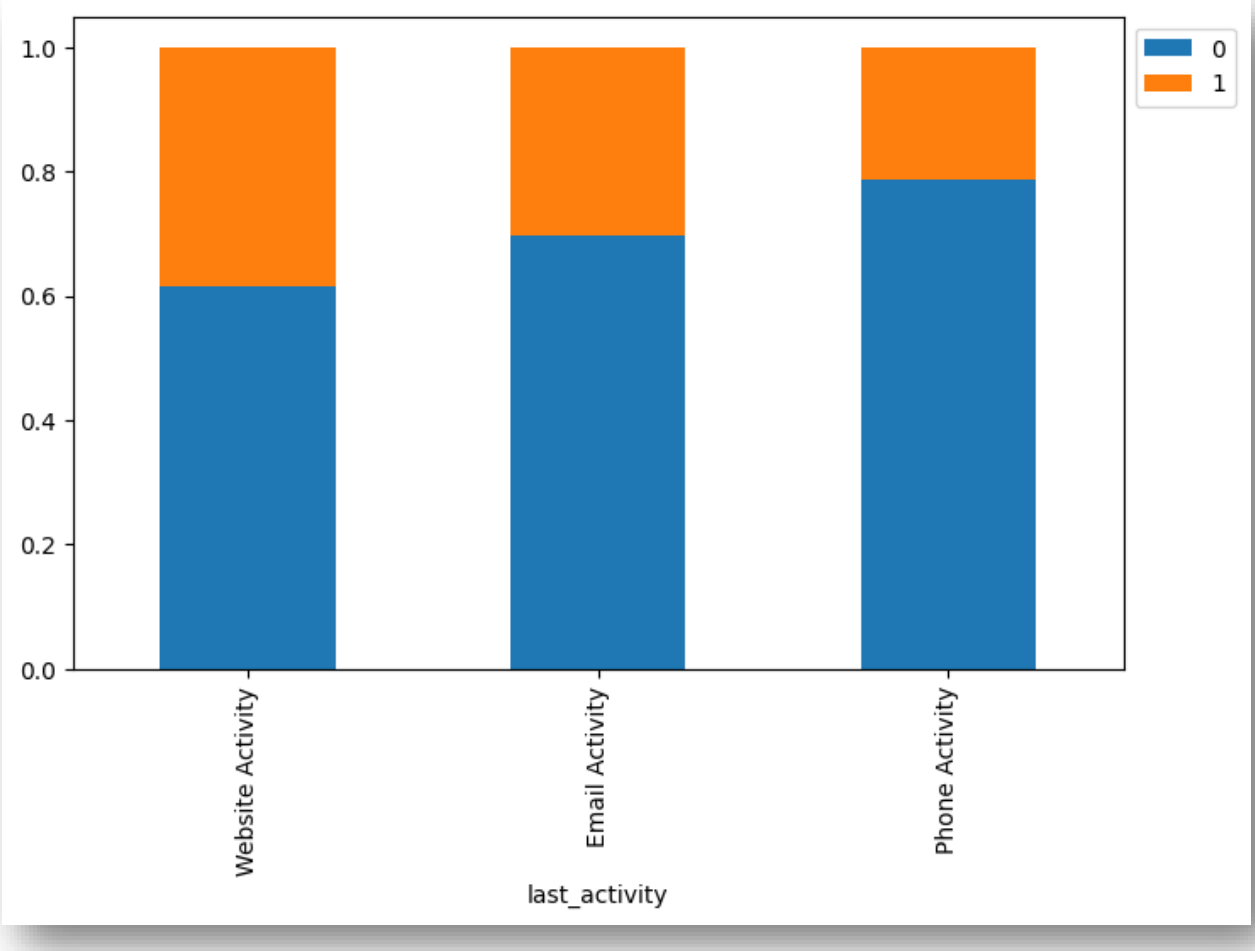


FIGURE 26

Insights:

1. E mail Activity:

a) Leads interaction with e mail activity have a likelihood of being status 0 compared to status 1

b) A significant portion (around 30%) is still status 1

2. Website Activity:

a) Leads interaction through website activity are more likely to be status 0 but the proportion of status 1 is higher as compared to E mail activity.

3. Phone Activity:

a) Leads interacting through phone activity have the highest proportion of status 0 suggesting that phone activity might be less effective in generating interest and are least expected to get converted.

♦ **Stacked Bar Plot of Print Media Type 1:**

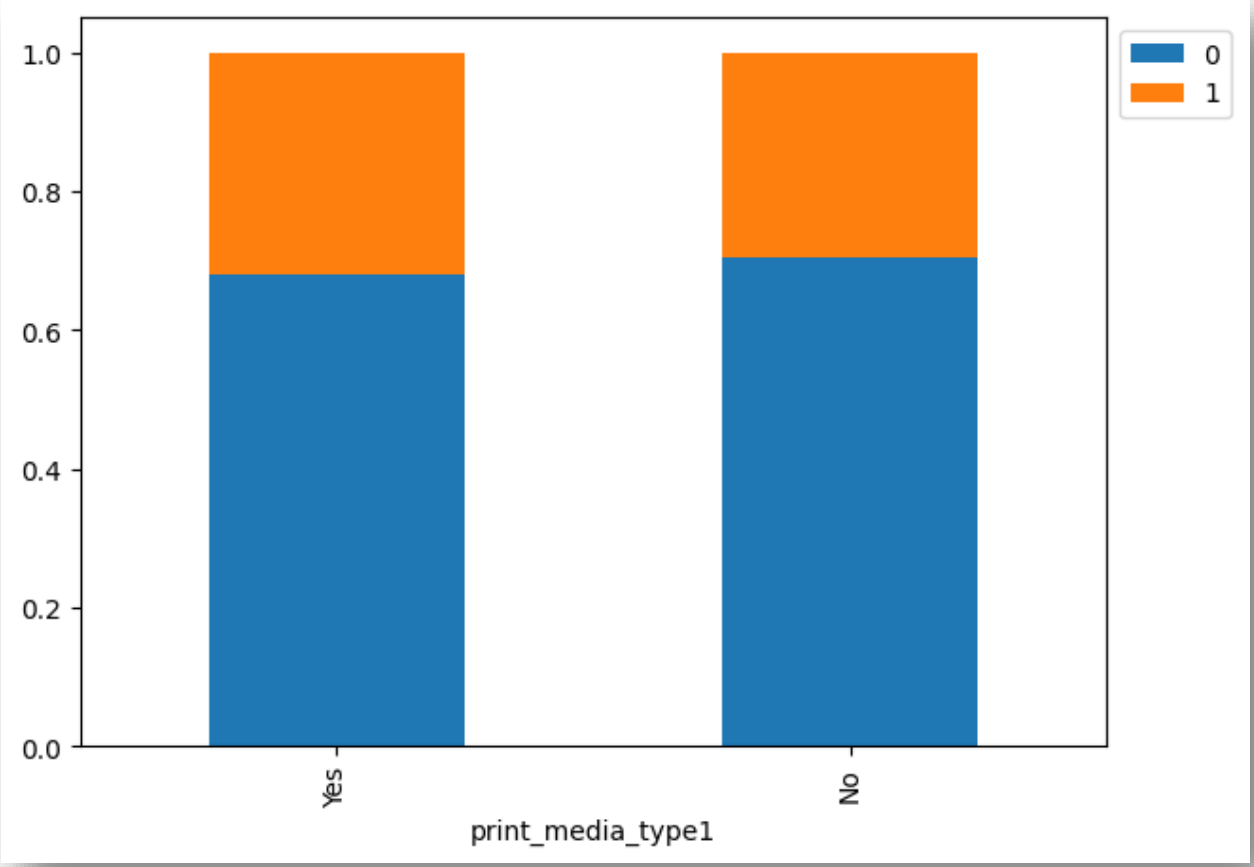


FIGURE 27

Insights:

- Distribution of Leads:

a) The plot shows 2 categories ‘Yes’ and ‘No’ which represents whether the lead has seen the ad in the newspaper or not.

- Conversion Rate Comparison:

a) The ‘Yes’ category has a slightly higher proportion of converted leads compared to the ‘No’ category.

b) This suggests that leads who have seen ad in print media type 1 are more likely to convert to paid customers than those who do not.

- Impact of Print media type 1:

a) The difference in conversion rates between ‘Yes’ and ‘No’ categories suggests that print media type 1 is an effective channel for attracting the potential customers.

- Given the higher conversion rate for leads exposed to print media type 1, ExtraaLearn might consider increasing their investment in this advertising channel.

♦ Stacked Bar Plot of Referral vs Status:

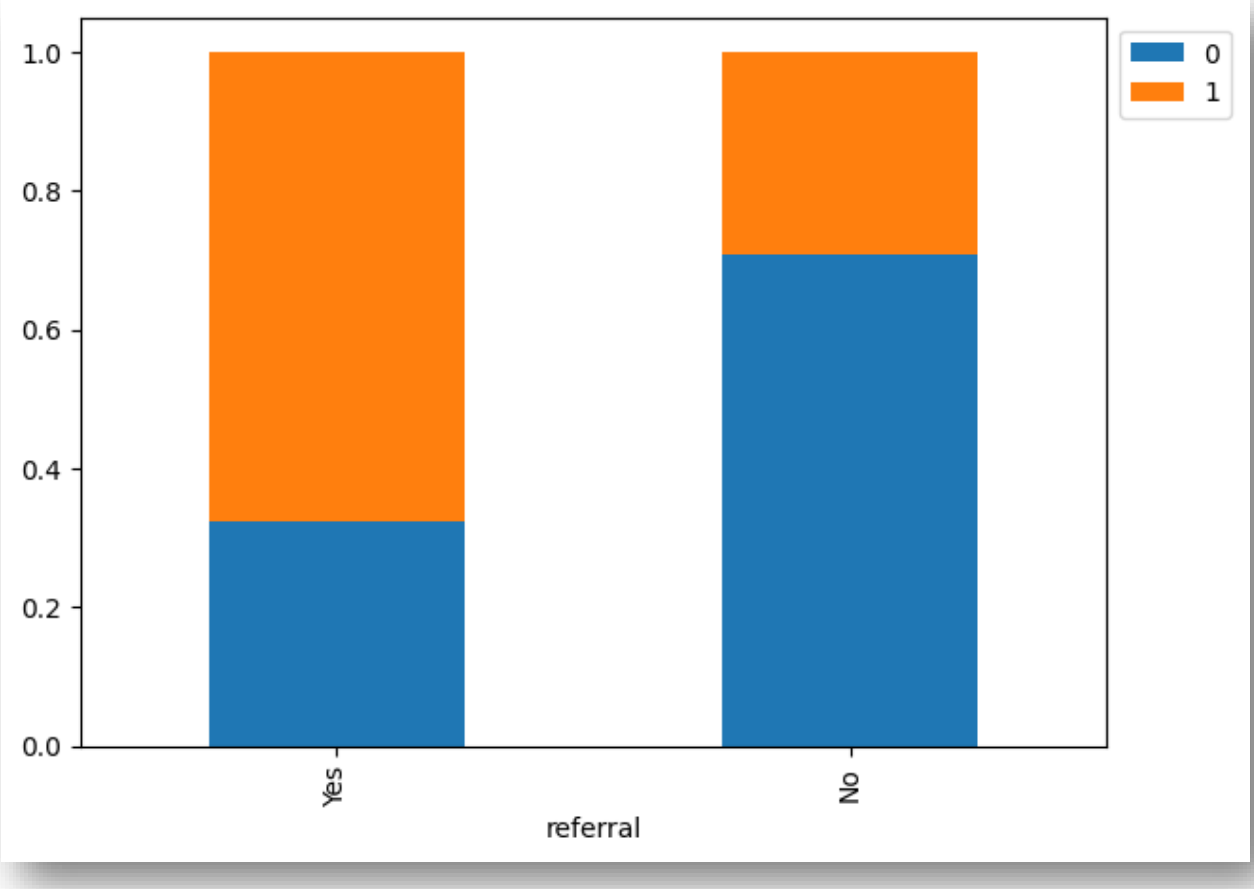


FIGURE 28

Insights:

- Referral: Yes

a) About 40% of the leads with a referral have an outcome of 0

b) About 60% of the leads with a referral have an outcome of 1

- Referral: No

a) About 65% of the leads without a referral have an outcome of 0

b) About 35% of the leads without a referral have an outcome of 1

c) Leads who came from referral are more likely to get converted as compared to those who did not. This indicates that referral might be positively influencing the conversion rate.

♦ Stacked Bar Plot of Digital Media vs Status:

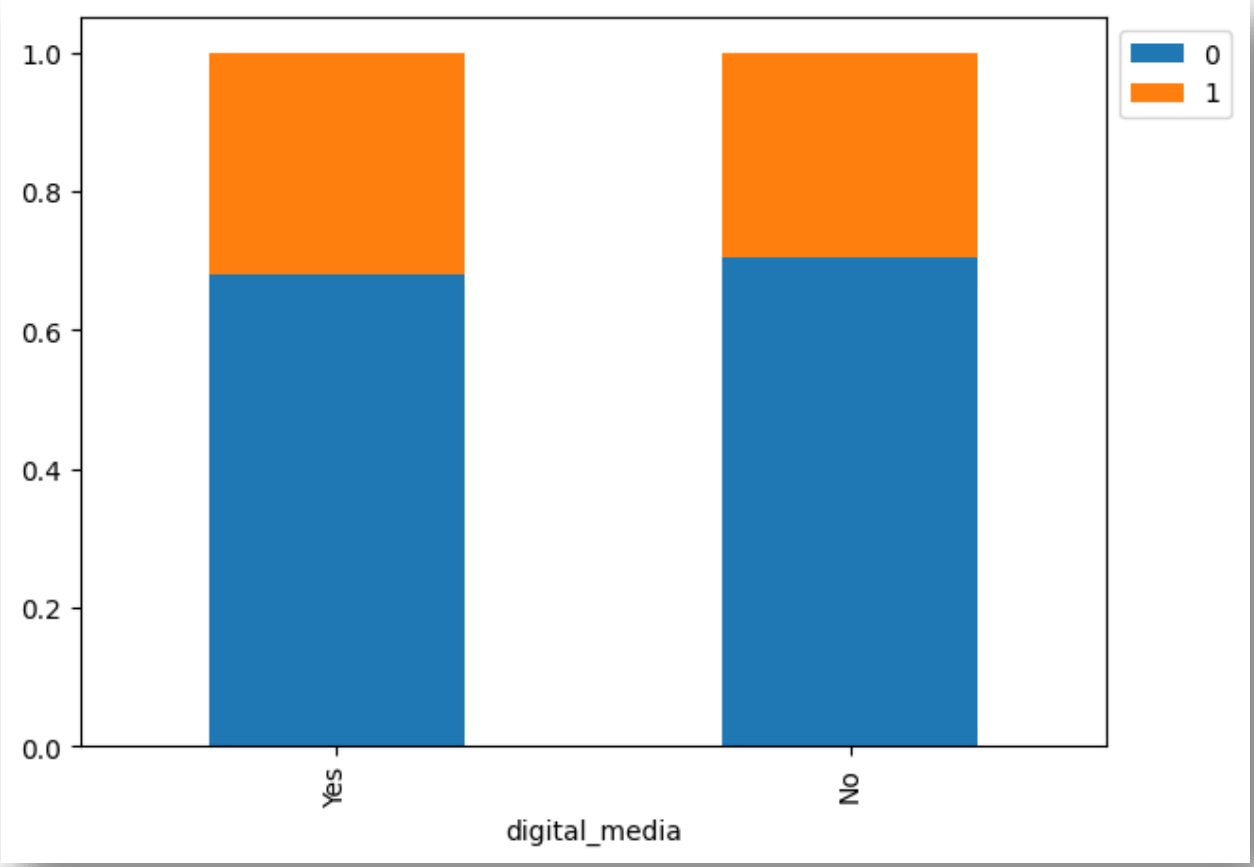


FIGURE 29

Insights:

- Digital Media: Yes

a) About 40% of the leads who used digital media for- interaction have an outcome of 0

b) About 40% of the leads who used digital media for interaction have an outcome of 1

- Digital Media: No

a) About 50% of the leads who did not used digital media for interaction have an outcome of 0

b) About 50% of the leads who did not used digital media for interaction have an outcome of 1

c) A positive correlation can be seen between digital media interaction and status, suggesting that leads who used digital media are more likely to get converted.

❑ DATA PREPROCESSING

♦ Outlier detection using Box Plot:

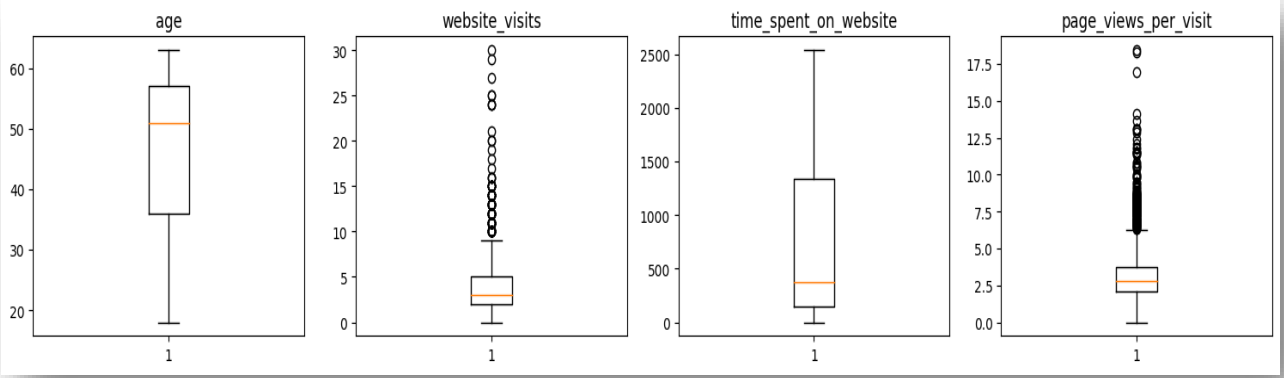


FIGURE 30

Insights based on the above Box Plot of various metrics:

- Age:

a) The median age is around 50 years.

b) IQR is from the range 40 to 55 years.

c) There are no significant outliers in the age distribution.

- Website Visits:

a) The median number of website visits is about 4

b) The IQR ranges from 2 to 6 visits.

c) There are several outliers present in the box plot with visits up to 30

- Time Spent on the Website:

a) Median time spent on the website is about 500 seconds.
b) The IQR ranges from 250 to 1000 seconds.
c) There are several outliers present showing leads spending up to 2500 seconds on the website.

- Page Views Per Visit:

a) The median page views per visit is about 2
b) The IQR ranges from 1.5 to 3 page views per visit.
c) There are outliers present with page views per visit up to 17.5 per visit.
d) The box plot also indicates that most leads are middle aged, typically visit the website few times, spend a moderate amount of time and view a few pages per visit.
e) However, there are few leads with significantly higher engagement, indicated by outliers in the website visits, time spent and page views per visit.

❑ DATA PREPARATION FOR MODELLING

Since we want to predict which lead is more likely to get converted, we will build a model and split the data into train and test data so that evaluation of the model can be done.

➢ Train and test set in the ratio of 75:25, post splitting:

Shape of training set	**(3459, 16)**

Shape of test set	**(1153, 16)**
Shape of training set	**(3459,)**
Shape of test set	**(1153,)**

TABLE 3

➢ **Percentage of classes in training set:**

Status	
0	**0.70107**
1	**0.29893**

TABLE 4

➢ **Percentage of classes in test set:**

Status	
0	**0.70252**
1	**0.29748**

TABLE 5

Insights based on the Table 2, 3, 4 above:

- The class distribution in both the training and test sets are quite similar.
- Around 75% of the samples belong to class 0 and around 25% belong to class 1

- This indicates that the data sets are imbalanced with a noticeable higher proportion of class 0 samples.
- The similar class distribution in both the sets also suggests that the split is done properly, ensuring the model will not be biased towards any particular class due to the training or test data distributions.
- The data exhibits a class imbalance with class 0 being more prevalent than class 1. This imbalance is consistent across both the training and test sets.

❑MODEL BUILDING

Model Evaluation Criterion

♦ **Logistic Regression Results:**

Logit Regression Results

Dep. Variable:	status	No. Observations:	3459
Model:	Logit	Df Residuals:	3442
Method:	MLE	Df Model:	16
Date:	Wed, 17 Jul 2024	Pseudo R-squ.:	0.3553
Time:	06:15:34	Log-Likelihood:	-1360.3
converged:	True	LL-Null:	-2109.8
Covariance Type:	nonrobust	LLR p-value:	7.912e-310

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.4531	0.059	-24.504	0.000	-1.569	-1.337
age	0.0734	0.067	1.103	0.270	-0.057	0.204
website_visits	-0.0226	0.050	-0.457	0.648	-0.120	0.075
time_spent_on_website	0.9652	0.051	18.907	0.000	0.865	1.065
page_views_per_visit	-0.0493	0.050	-0.988	0.323	-0.147	0.048
current_occupation_Student	-0.6201	0.079	-7.887	0.000	-0.774	-0.466
current_occupation_Unemployed	-0.2473	0.049	-5.029	0.000	-0.344	-0.151
first_interaction_Website	1.3399	0.060	22.306	0.000	1.222	1.458
profile_completed_Low	-0.3942	0.076	-5.214	0.000	-0.542	-0.246
profile_completed_Medium	-0.8002	0.052	-15.472	0.000	-0.902	-0.699
last_activity_Phone Activity	-0.3013	0.054	-5.608	0.000	-0.407	-0.196
last_activity_Website Activity	0.2213	0.049	4.492	0.000	0.125	0.318
print_media_type1_Yes	0.0615	0.046	1.331	0.183	-0.029	0.152
print_media_type2_Yes	0.0433	0.046	0.942	0.346	-0.047	0.134
digital_media_Yes	0.0253	0.047	0.542	0.588	-0.066	0.117
educational_channels_Yes	0.0287	0.049	0.591	0.555	-0.067	0.124
referral_Yes	0.1868	0.048	3.879	0.000	0.092	0.281

TABLE 6

Insights:

- Time spent on the website is a highly significant predictor. The positive coefficient indicates that more time spent on the website increases the likelihood of a positive outcome (status=1)
- First interaction by website is also highly significant predictor. A first interaction through website increases the likelihood of getting a lead converted.
- Being a student or unemployed has a negative impact on the likelihood of a positive outcome.

- Lower level of profile completion is significantly associated with a decreased likelihood of being converted.
- Last activity being phone activity has a negative impact while website activity has a positive impact.
- Leads coming from referral has a positive impact on the conversion rate.
- Age, website visits, page views per visit, print media type 1, print media type 2, digital media and educational channels are non-significant predictors in this model.

➢ CHECKING LOGISTIC REGRESSION MODEL PERFORMANCE ON THE TRAINING SET:

	Accuracy	Recall	Precision	F1
0	0.82249	0.65377	0.72532	0.68769

TABLE 7

♦ Plot of Confusion Matrix on Train Set:

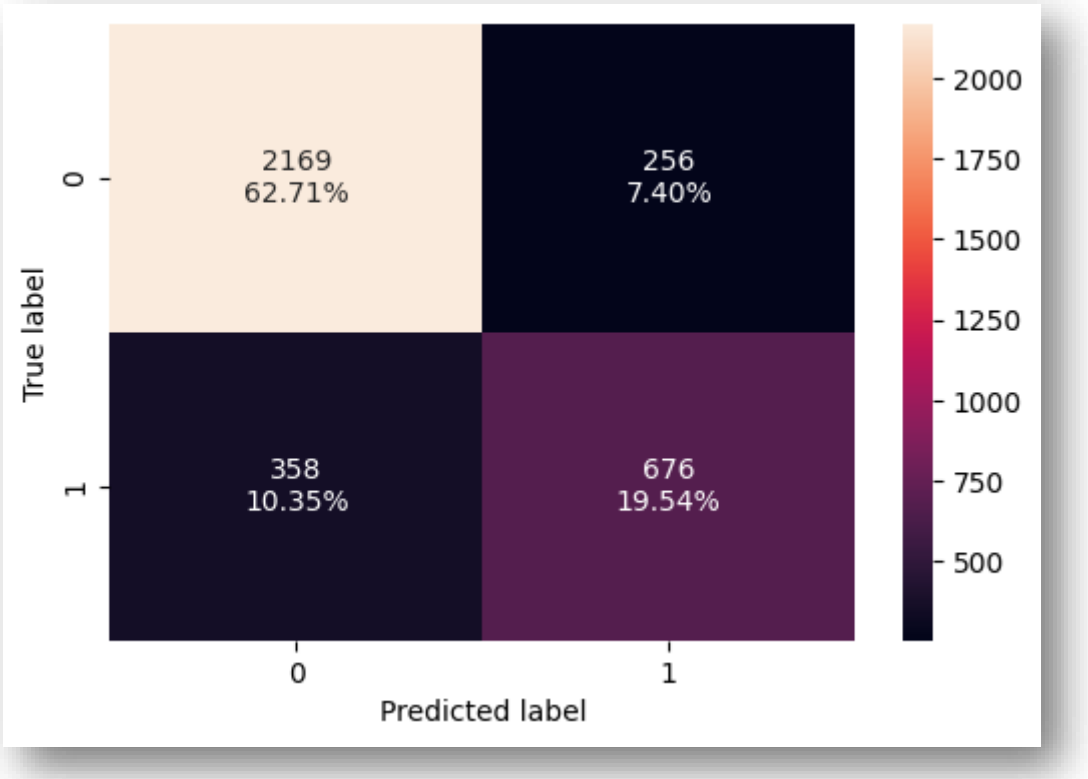


FIGURE 31

Insights:

- True Positive (TP): 676 (19.54%) instances of class 1 are correctly predicted.
- True Negative (TN): 2169 (62.71%) instances of class 0 are correctly predicted.
- False Positive (FP): 256 (7.40%) instances of class 0 are incorrectly predicted as class 1
- False Negative (FN): 358 (10.35%) instances of class 1 are incorrectly predicted as class 0

Calculations:

Based on the confusion matrix, performance matrix can be calculated for the training set.

→ Accuracy:
(TP+TN)/ Total
= (676+2169)/3459 ≈ **82.25%**

→Precision for Class 1:
TP/(TP+FP)
= 676/ (676+256) ≈ **72.53%**

→Recall for Class 1:
TP/(TP+FN)
= 676/ (676+358) ≈ **65.38%**

→F1 score for Class 1:
2*(Precision*Recall)/(Precision+Recall) ≈ **68.77%**

♦ **Plot of Confusion Matrix on Test Set:**

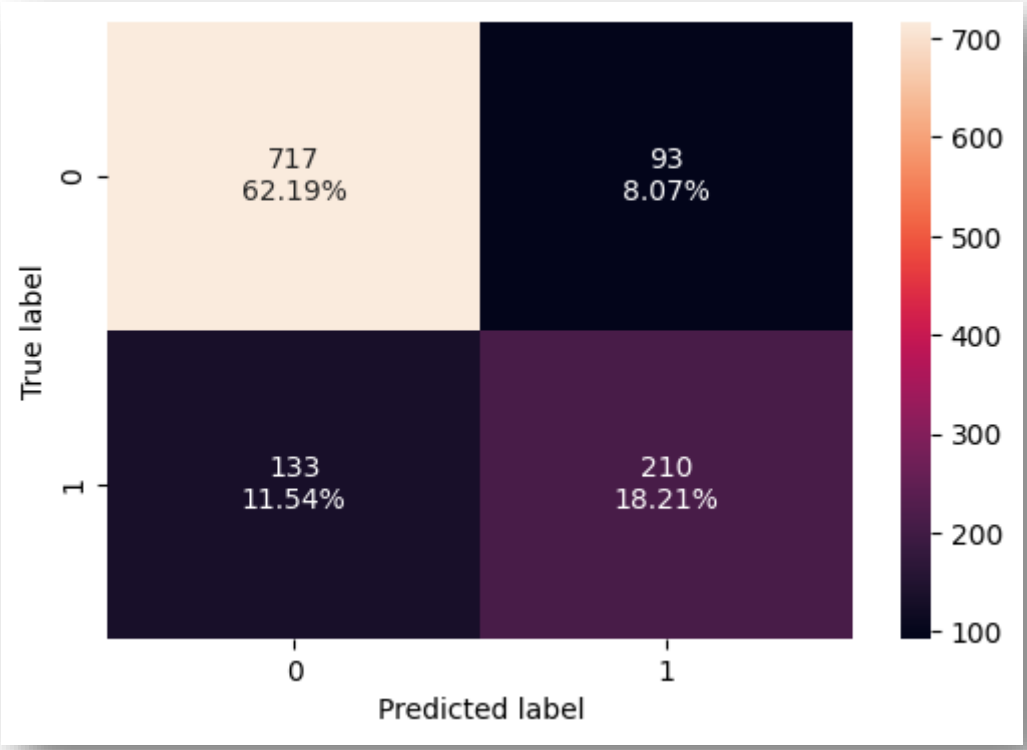


FIGURE 32

Insights:

- True Positive (TP): 210 (18.21%) instances of class 1 are correctly predicted.
- True Negative (TN): 717 (62.19%) instances of class 0 are correctly predicted.
- False Positive (FP): 93 (8.07%) instances of class 0 are incorrectly predicted as class 1
- False Negative (FN): 133 (11.54%) instances of class 1 are incorrectly predicted as class 0

Calculations:

Based on the confusion matrix, performance matrix can be calculated for the training set.

→Accuracy:
(TP+TN)/ Total
= (210+717)/1153 ≈ **80.40%**

→Precision for Class 1:
TP/(TP+FP)
= 210/ (210+93) ≈ **63.90%**

→Recall for Class 1:
TP/(TP+FN)
= 210/ (210+133) ≈ **61.22%**

→F1 score for Class 1:
2*(Precision*Recall)/(Precision+Recall) ≈ **65.00%**

Insights based on the Confusion Matrices of Train and Test Sets:

- The model's performance is consistent across both the training and test sets, with accuracy slightly decreasing from training to test data, which is expected due to overfitting.

- The model performs reasonably well considering the class imbalance, though improvements can be made especially in terms of recall for class 1
- The precision for class 1 is higher than the recall for both the training and test sets, suggesting the model is more conservative in predicting the class 1 and avoids false positive at the cost of missing some true positives.
- The slight drop in performance from the training to test set indicates that the model has generalized well but can slightly be improved.

❑ NAIVE- BAYES CLASSIFIER

After building a Naive Bayes Model, we will be checking the Naive- Bayes Classifier performance on both the training set and testing set and will analyze the results.

➢ Checking Naive- Bayes Classifier Performance on the Training set:

	Accuracy	Recall	Precision	F1
0	0.78925	0.76886	0.61868	0.68564

TABLE 8

♦ Plot of Confusion Matrix on Training Data:

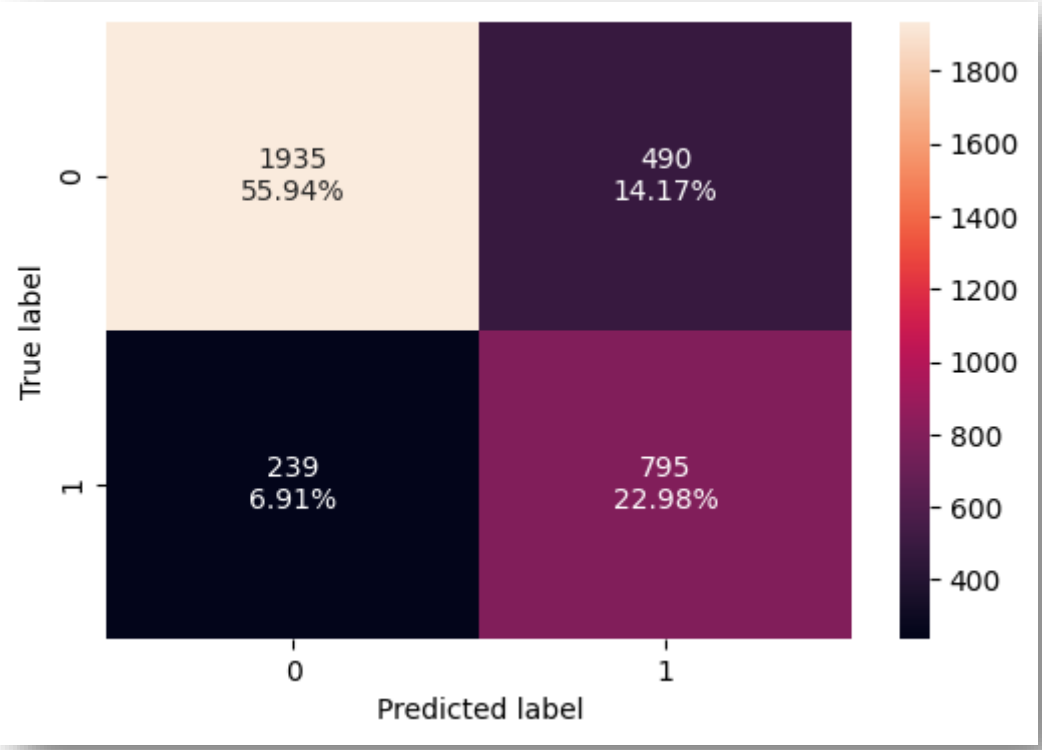


FIGURE 33

Observations:

- True Negative (TN): 1935 (55.94%)
- False Positive (FP): 490 (14.17%)
- False Negative (FN): 239 (6.91%)
- True Positive (TP): 795 (22.98%)

Calculations:

→ Accuracy:

(TP+TN)/ Total

= (795+1935)/ (795+1935+490+239) = 2730/3459 ≈ **78.9%**

→ Precision for Class 1:

TP/(TP+FP)

= 795/ (795+490) = 795/1285 ≈ **61.8%**

→Recall for Class 1:

TP/(TP+FN)

= 795/ (795+239) = 795/1034 ≈ **76.9%**

→F1 Score for Class 1:

2*(Precision+Recall)/(Precision+Recall) ≈ **68.6%**

- **Checking Naive- Bayes Classifier Performance on the Test set:**

	Accuracy	Recall	Precision	F1
0	0.77971	0.75802	0.60325	0.67183

TABLE 9

- **Plot of Confusion Matrix on Test Data:**

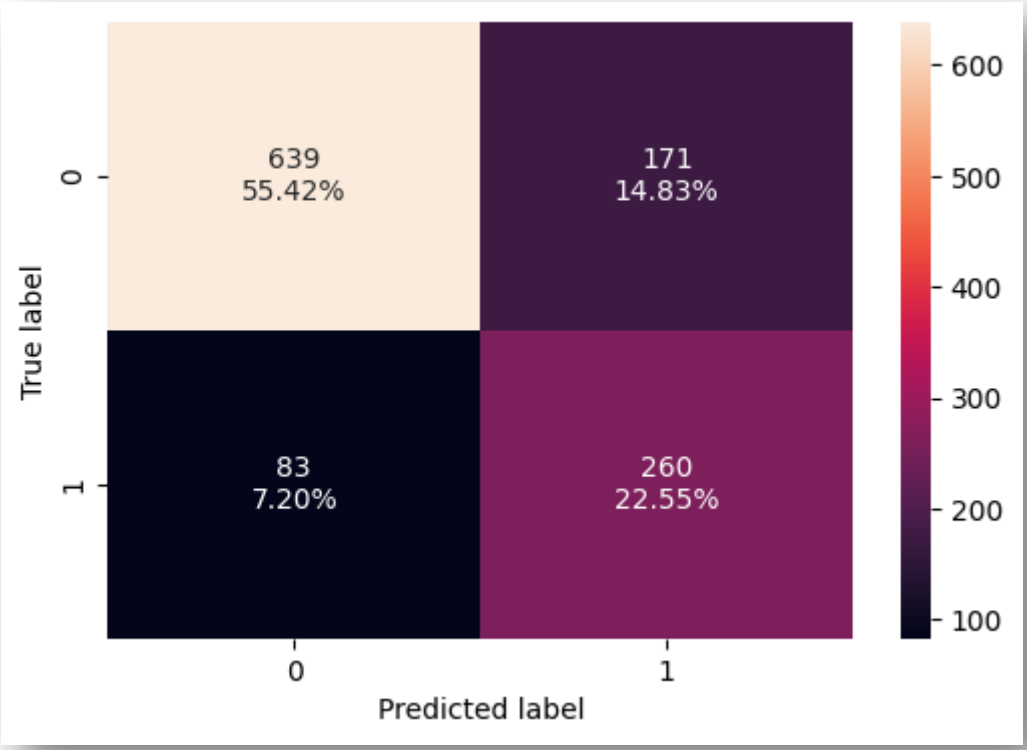


FIGURE 34

Observations:

- True Negative (TN): 639 (55.42%)
- False Positive (FP): 171 (14.83%)
- False Negative (FN): 83 (7.20%)
- True Positive (TP): 260 (22.55%)

Calculations:

→ Accuracy:

(TP+TN)/ Total

= (260+639)/ (260+83+171+639) = 899/1153 ≈ **78.0%**

→ Precision for Class 1:

TP/(TP+FP)
= 260/ (260+171) = 260/431 ≈ **60.3%**

→Recall for Class 1:
TP/(TP+FN)
= 260/ (260+83) = 260/343 ≈ **75.8%**

→F1 Score for Class 1:
2*(Precision+Recall)/(Precision+Recall) ≈ **67.1%**

Insights based on the Confusion Matrices of Training & Test data (FIGURE 31 & 32):

- Accuracy: The train model has a slightly higher accuracy as compared to the test model.
- Precision: The train model has a higher precision as compared to the test model. This indicates that the train model is slightly better at predicting the positive class correctly among the predicted positives.
- Both the models have similar recall values with the train model having a recall of 76.9% and the test model having a recall of 75.8% This suggests that both the models are equally good at capturing the actual positive cases.
- The F1 score for the train model is slightly higher than the test model indicating a better balance between the precision and recall in the train set.
- Both the train and test set show a similar performance in terms of accuracy and recall. However, the train model

slightly outperforms the test model in terms of precision and F1 score.

- If the application prioritizes precision, the train model would be preferable.

❑ KNN CLASSIFIER

After building a KNN Model with k=3

We will check the KNN Classifier performance on the training set & testing set and analyze the results.

➢ Checking KNN Classifier performance on the training set:

	Accuracy	Recall	Precision	F1
0	0.88927	0.77466	0.84227	0.80705

TABLE 10

♦ Plot of Confusion Matrix on the Training set:

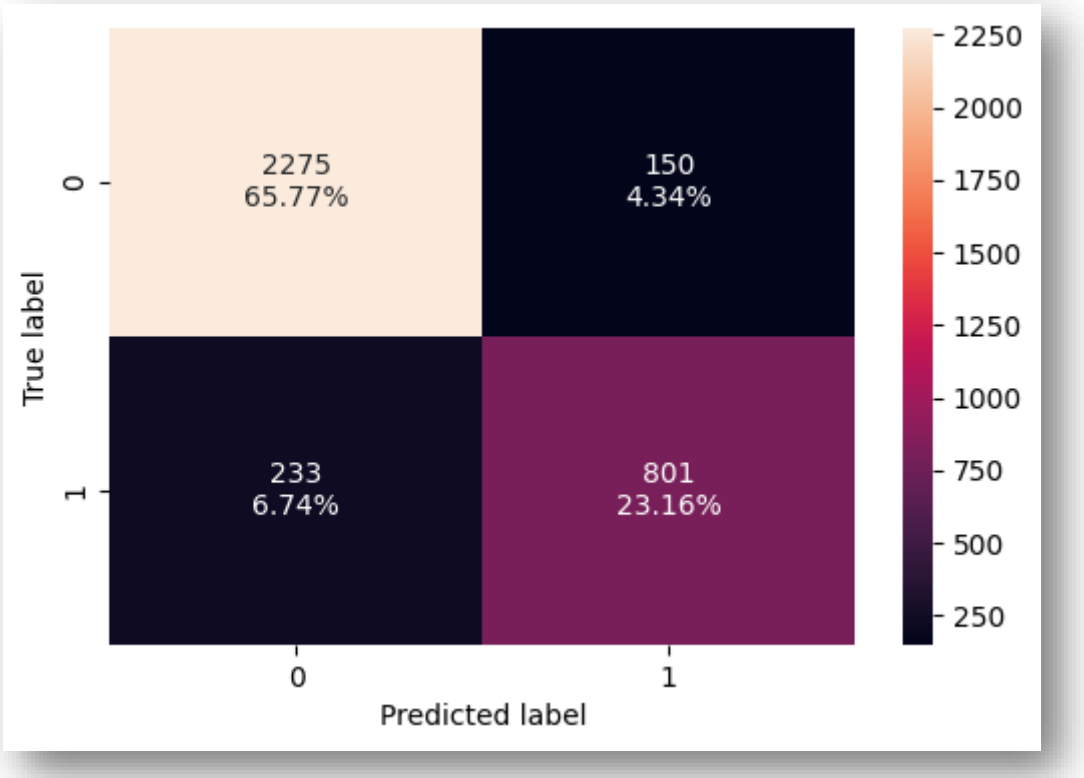


FIGURE 35

Observations:

- True Negative (TN): 2275 (65.77%)
- False Positive (FP): 150 (4.34%)
- False Negative (FN): 233 (6.74%)
- True Positive (TP): 801 (23.16%)

Calculations:

→ Accuracy:
(TP+TN)/ Total
= (801+2275)/ (2275+150+233+801) = 3076/ 3459 ≈ **89.0%**

→ Precision for Class 1:

TP/(TP+FP)

= 801/ (801+150) = 801/951 ≈ **84.2%**

→Recall for Class 1:

TP/(TP+FN)

= 801/ (801+233) = 801/1034 ≈ **77.5%**

→F1 Score for Class 1:

2*(Precision+Recall)/(Precision+Recall) ≈ **80.7%**

→ False Positive Rate:

FP/(FP+TN)

= 150/ (150+2275) = 150/2425 ≈ **61.8%**

→ False Negative Rate:

FN/(FN+TP)

= 233/ (233+801) = 233/ 1034 ≈ **22.5%**

Insights based on the confusion matrix on the training set:

- The model has a high accuracy of 89.0% suggesting that it correctly classifies a high percentage of the total cases.

- With a percentage of 84.2% the model is effective in predicting the positive cases accurately among all cases it predicts as positive.
- The recall of 77.5% shows that the model successfully identifies a large proportion of the actual positive cases but still misses some.
- The F1 score of 80.7% suggests a good balance between precision and recall making the model efficient in handling both the false positives and false negatives.
- The low False Positive Rate (FPR) of 6.2% implies that the model rarely misclassifies actual negatives as positives.
- False Negative Rate (FNR) of 22.5% indicates that the model can be improved more for capturing all the actual positive cases.
- Therefore, we can say that this confusion matrix presents a well performing model, strong in accuracy and precision. However, efforts to reduce the false negative rate could enhance its effectiveness.

➢ Checking KNN Classifier performance on the test set:

	Accuracy	Recall	Precision	F1
0	0.79879	0.59767	0.68562	0.63863

TABLE 11

♦ Plot of Confusion Matrix on the Test set:

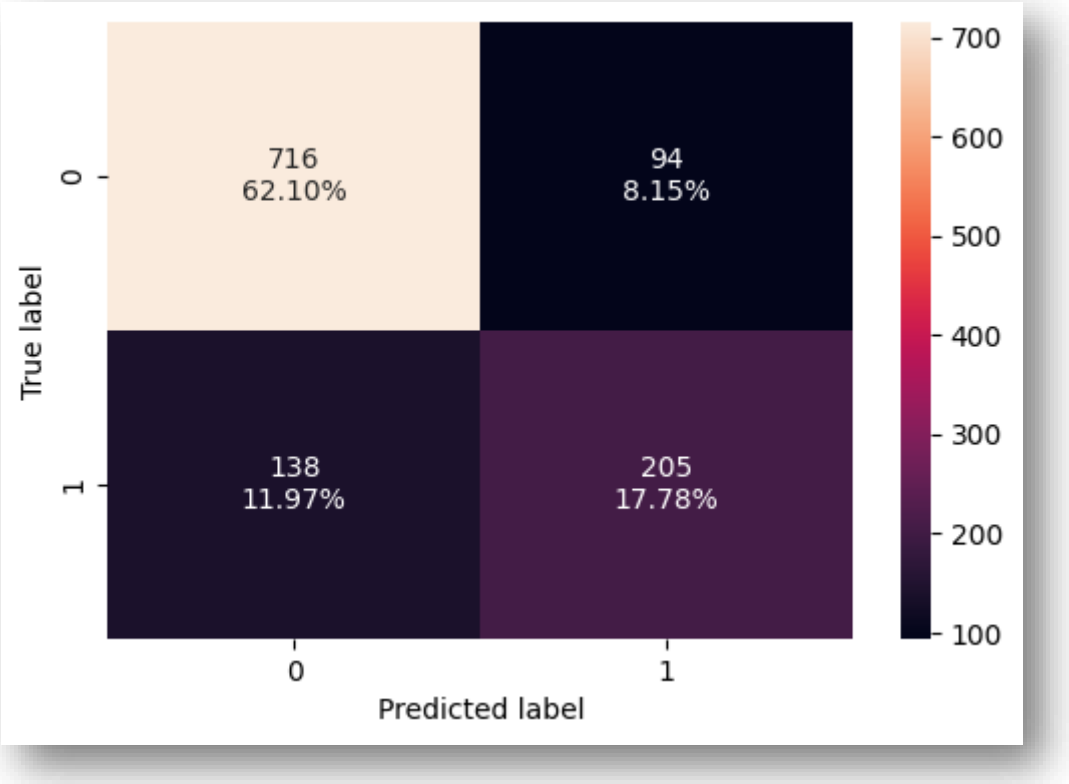


FIGURE 36

Observations:

- True Negative (TN): 716 (62.10%)
- False Positive (FP): 94 (8.15%)
- False Negative (FN): 138 (11.97%)
- True Positive (TP): 205 (17.78%)

Calculations:

→ Accuracy:
(TP+TN)/ Total
= (716+205)/ (716+94+138+205) = 921/1153 ≈ **79.8%**

→ Precision for Class 1:
TP/(TP+FP)
= 205/ (205+94) = 205/299 ≈ **68.6%**

→Recall for Class 1:
TP/(TP+FN)
= 205/ (205+138) = 205/343 ≈ **59.8%**

→F1 Score for Class 1:
2*(Precision+Recall)/(Precision+Recall) ≈ **63.9%**

Insights based on the confusion matrix on the test set:

- The model has relatively high accuracy of approx. 79.97% suggesting that it correctly classifies a significant portion of the test set.
- The precision of 68.56% suggests that the model predicts 68.56% correct. This is moderately good but indicates some room for improvement to reduce false positives.
- The recall of 59.71% suggests that the model identifies about 59.71% of the actual positive cases.

- The F1 score of 63/87% balances precision and recall indicating the overall performance of the model considering both false positives and false negatives.

❑ DECISION TREE CLASSIFIER

➢ Checking Decision Tree Classifier performance on the training set:

	Accuracy	Recall	Precision	F1
0	0.99971	0.99903	1.0000	0.99952

TABLE 12

♦ Plot of Confusion Matrix on the Training set:

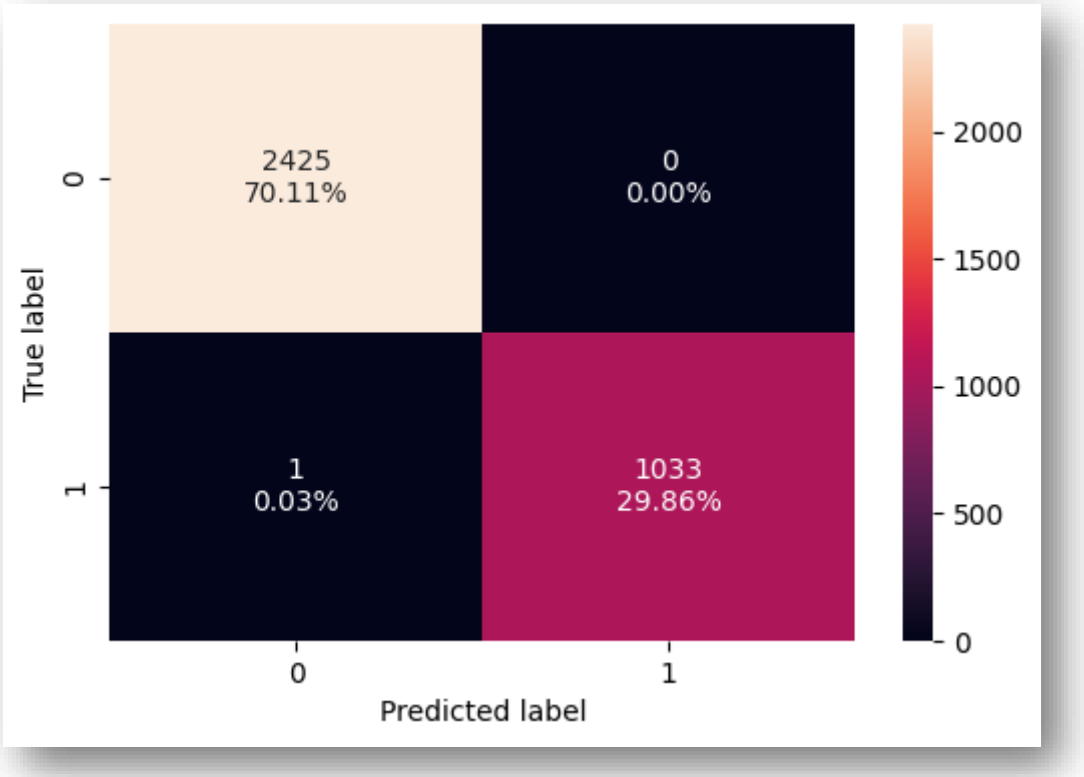


FIGURE 37

Observations:

- True Negative (TN): 2425 (70.11%)
- False Positive (FP): 0 (0.00%)
- False Negative (FN): 1 (0.03%)
- True Positive (TP): 1033 (29.86%)

Calculations:

→ Accuracy:
(TP+TN)/ Total
= (2425+1033)/ (2425+1+0+1033) = 3458/3459 ≈ **99.97%**

→ Precision for Class 1:

TP/(TP+FP)
= 1033/ (1033+0) = 1033/1033 ≈ **100%**

→Recall for Class 1:
TP/(TP+FN)
= 1033/ (1033+1) = 1033/1034 ≈ **99.9%**

→F1 Score for Class 1:
2*(Precision+Recall)/(Precision+Recall) ≈ **99.9%**

Insights based on the confusion matrix on the training set:

- The model has an extremely high accuracy of approx. 99.97% suggesting that its correctly classifies nearly all in the training set.
- The precision of 100% suggests that when the model predicts a positive class it is always correct.
- The recall of 99.90% indicates that the model identifies almost all the actual positive cases with only minimal number of false negatives.
- The F1 score of 99.95% balances the precision and recall, suggesting the overall performance of the model considering both the false positives and false negatives is excellent.

- The model shows almost perfect performance on the training set, with extremely high accuracy, precision, recall and F1 score.
- This level of performance suggests that the model is highly effective at classifying both positives and negative cases.
- This also indicates potential overfitting where the model performs exceptionally well on the training data but may not generalize as good to unseen test set.

➢ Checking Decision Tree Classifier performance on the test set:

	Accuracy	Recall	Precision	F1
0	0.81526	0.67638	0.69461	0.68538

TABLE 13

♦ Plot of Confusion Matrix on the Test set:

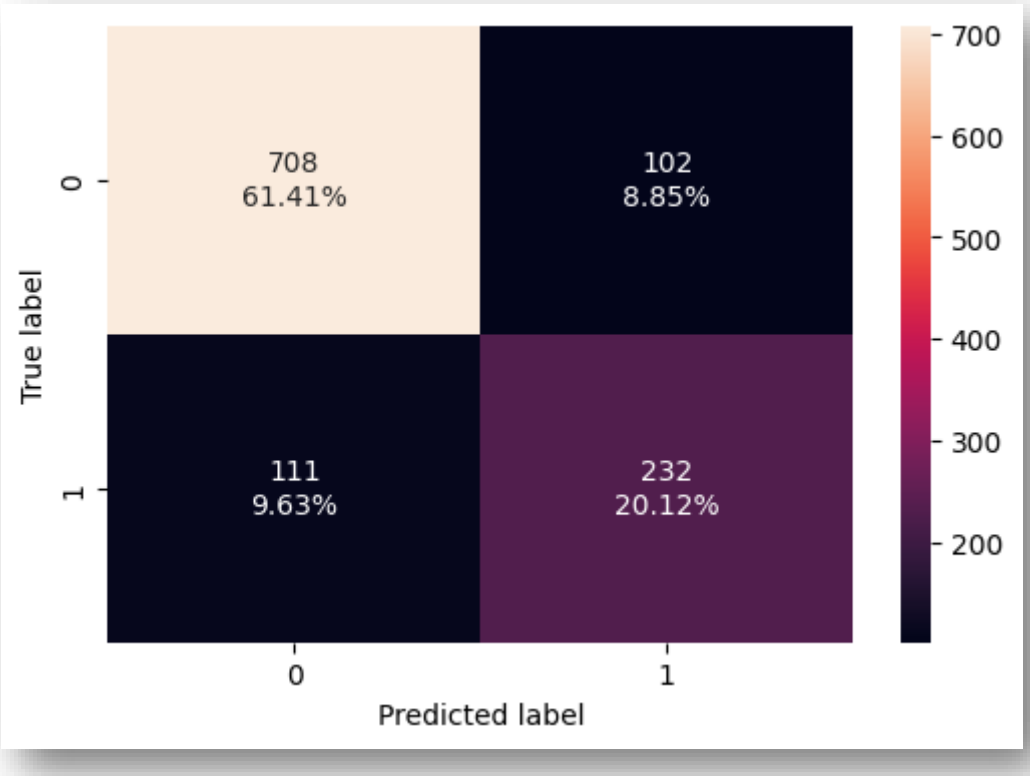


FIGURE 38

Observations:

- True Negative (TN): 708 (61.41%)
- False Positive (FP): 102 (8.85%)
- False Negative (FN): 111 (9.63%)
- True Positive (TP): 232 (20.12%)

Calculations:

→ Accuracy:

(TP+TN)/ Total

= (708+232)/ (708+102+111+232) = 940/1153 ≈ **81.5%**

→ Precision for Class 1:

TP/(TP+FP)

= 232/ (232+102) = 232/334 ≈ **69.5%**

→Recall for Class 1:

TP/(TP+FN)

= 232/ (232+111) = 232/343 ≈ **67.6%**

→F1 Score for Class 1:

2*(Precision+Recall)/(Precision+Recall) ≈ **68.6%**

Insights based on the confusion matrix on the test set:

- The model has a relatively high accuracy of approx. 81.51% suggesting that it correctly classifies a noticeable portion of the test set.
- The precision of 69.46% indicates that when a model predicts a positive class it is correct around 69.46% of the time. This moderately good but shows that there is room for improvement to reduce false positives.
- The recall of 67.67% suggests that the model identifies about 67.67% of the actual positive cases.
- The F1 score of 68.55% balances precision and recall suggesting the overall performance of the model considering both false positives and false negatives.
- The model shows strong performance in identifying negative cases but has a moderate effectiveness in identifying the positive cases.

- This imbalance suggests that while the model is good at avoiding false positive, it may need improvements to better capture true positives and reduce false negatives.

❑ MODEL PERFORMANCE IMPROVEMENT

Logistic Regression: Dealing with Multicollinearity

➢ Variance Inflation Factors (VIF):

Age	**6.96678**
Website_visits	**2.52701**
Time_spent_on_website	**1.96142**
Page_views_per_visit	**3.32519**
Current_occupation_student	**1.34553**
Current_occupation_unemployed	**1.55809**
First_interaction_website	**2.14406**
Profile_completed_low	**1.05209**
Profile_completed_medium	**1.92465**
Last_activity_phoneActivity	**1.51374**
Last_activity_websiteActivity	**1.46403**
Print_media_type1	**1.12686**
Print_media_type2	**1.05689**
Digital_media_yes	**1.12899**
Educational_channels_yes	**1.18695**
Referral_yes	**1.03004**

TABLE 14

Insights based on the VIF Table:

- High Multicollinearity:

a) Age: The VIF for 'age' is 6.96678 which is significantly higher than the common threshold of 5, suggesting a higher degree of multicollinearity with other predictors.

b) This indicates that 'age' is highly correlated with one or more variables in the model.

- Moderate Multicollinearity:

a) Website visits: The VIF is 2.52701 indicating moderate correlation with other variables.

b) Page views per visit: The VIF is 3.32519, which also suggests multicollinearity.

c) First interaction website: The VIF is 2.14406 showing a moderate level of multicollinearity.

- Low Multicollinearity: The remaining variables have VIF values below 2, indicating low multicollinearity.

♦ Histogram of Time Spent on the Website by Student Status:

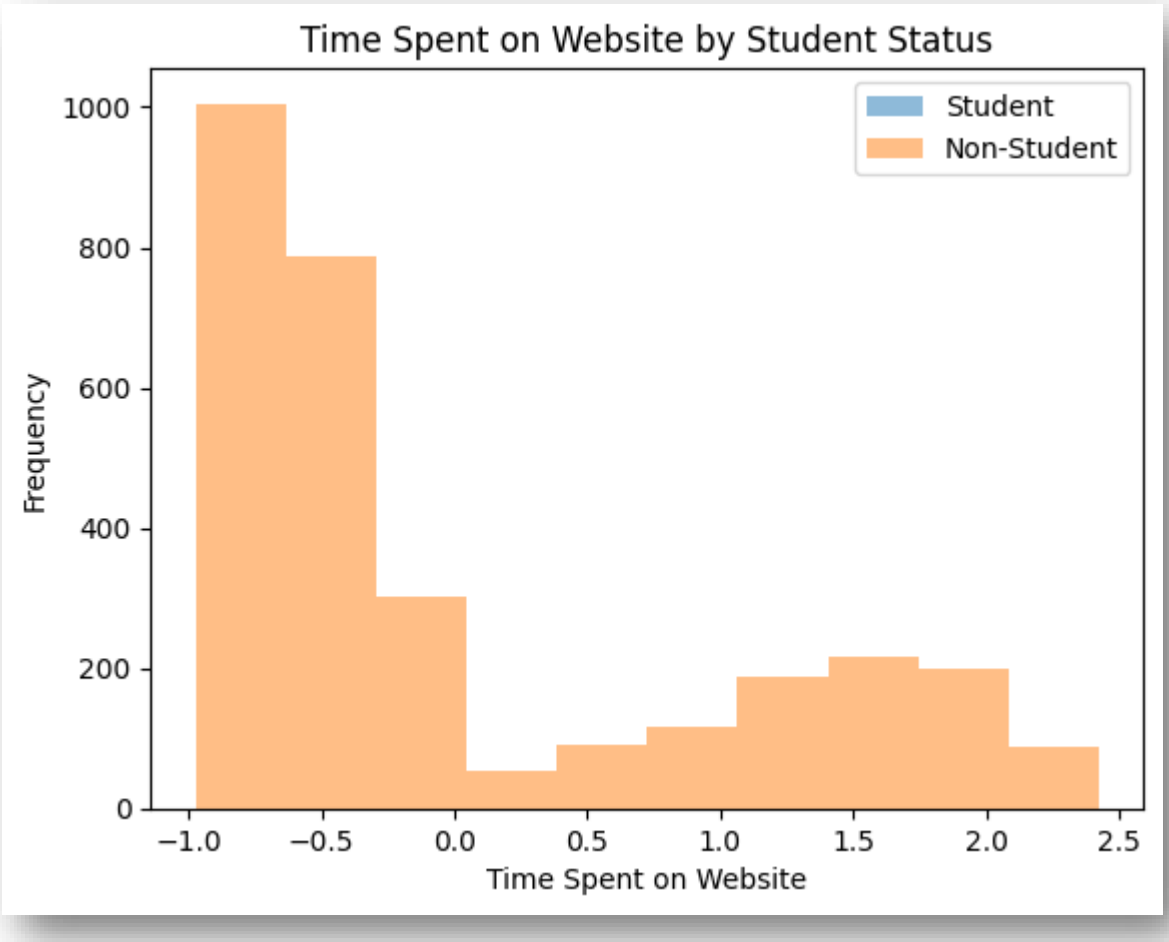


FIGURE 39

Insights:

- The histogram shows the distribution of time spent on the website categorized by student and non-student status.
- The majority of non-students spend between –1 and 0 on the website with a sharp decline in the frequency as the time spend increases.
- This indicates that non students tend to spend less time on the website compared to students.
- Plot also suggests that non students spend less time on the website with a skew towards shorter visits.

♦ **Line plot of Time Spent on the Website:**

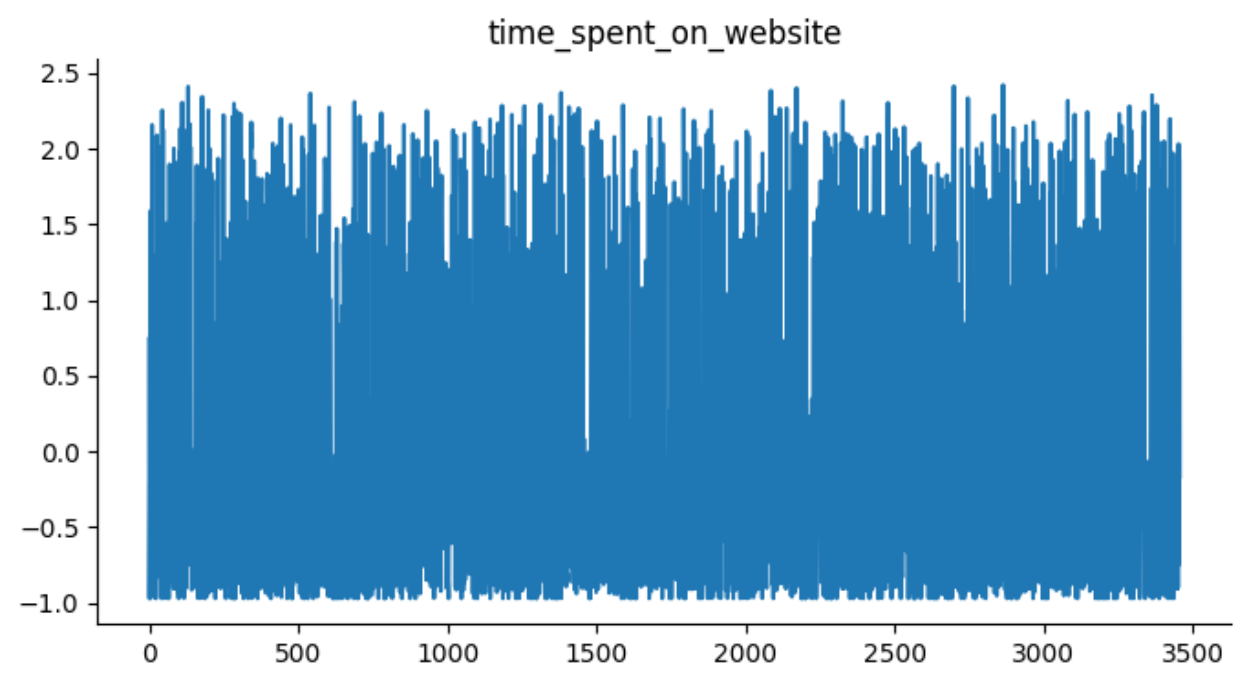


FIGURE 40

Insights:

- The line plot shows the time spent on the website.
- There is a wide range of time spent on the website from –1 to around 2.5
- The plot suggests high variability in the time spent on the website with no clear pattern.
- Some leads show very low time spent on the website which indicates short visits.

♦ **Scatter plot of Current Occupation (Student vs Unemployed):**

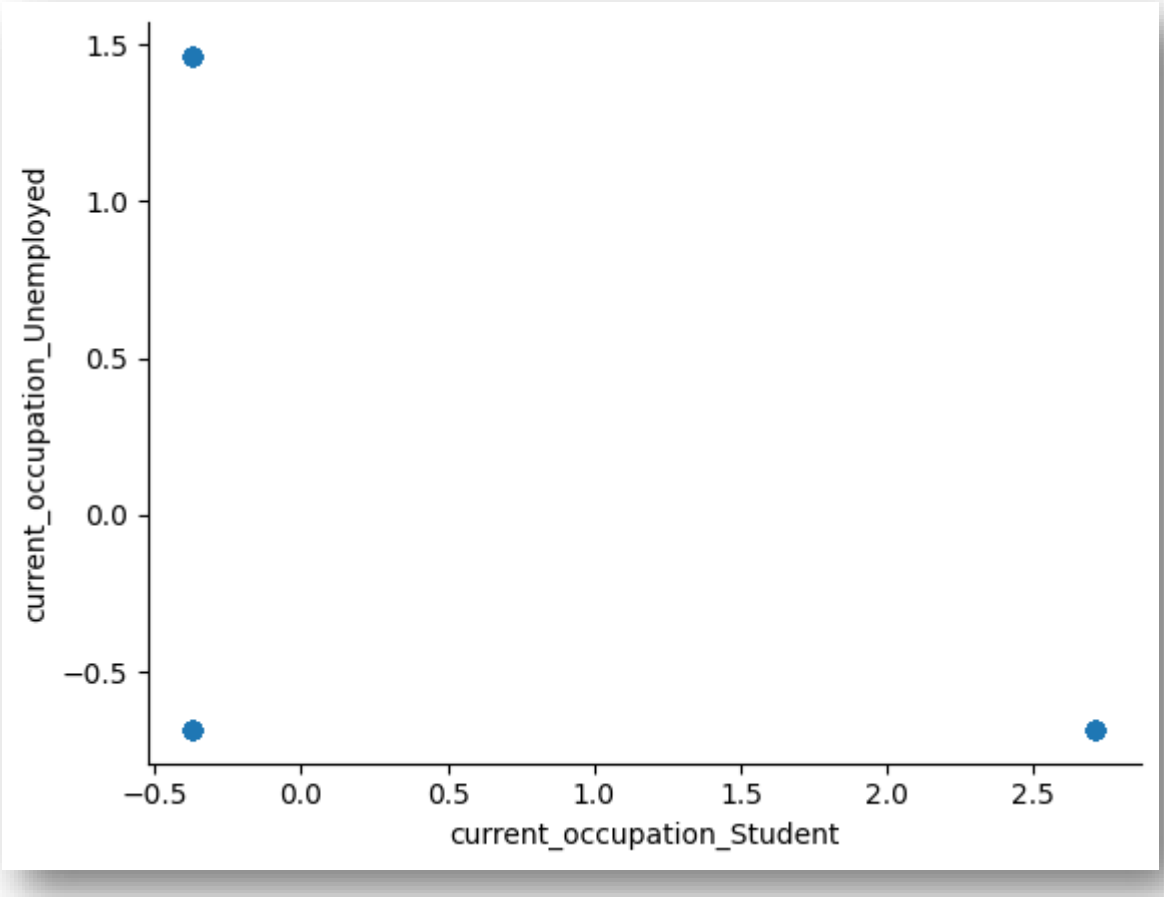


FIGURE 41

Insights:

- The scatterplot compares the current occupation of the leads whether they are ‘student’ or ‘unemployed’.
- The x axis represents the binary variable for students while the y axis represents the binary variable for unemployed.
- There are 2 clusters of point in the plot:

a) One cluster is at (-0.5, -0.5) suggesting users who are neither students nor unemployed.

b) The other cluster is at (2.5, 1.5) indicating users who are both student and unemployed.

- **Scatter plot of Time spent on the website vs current_ occupation_student:**

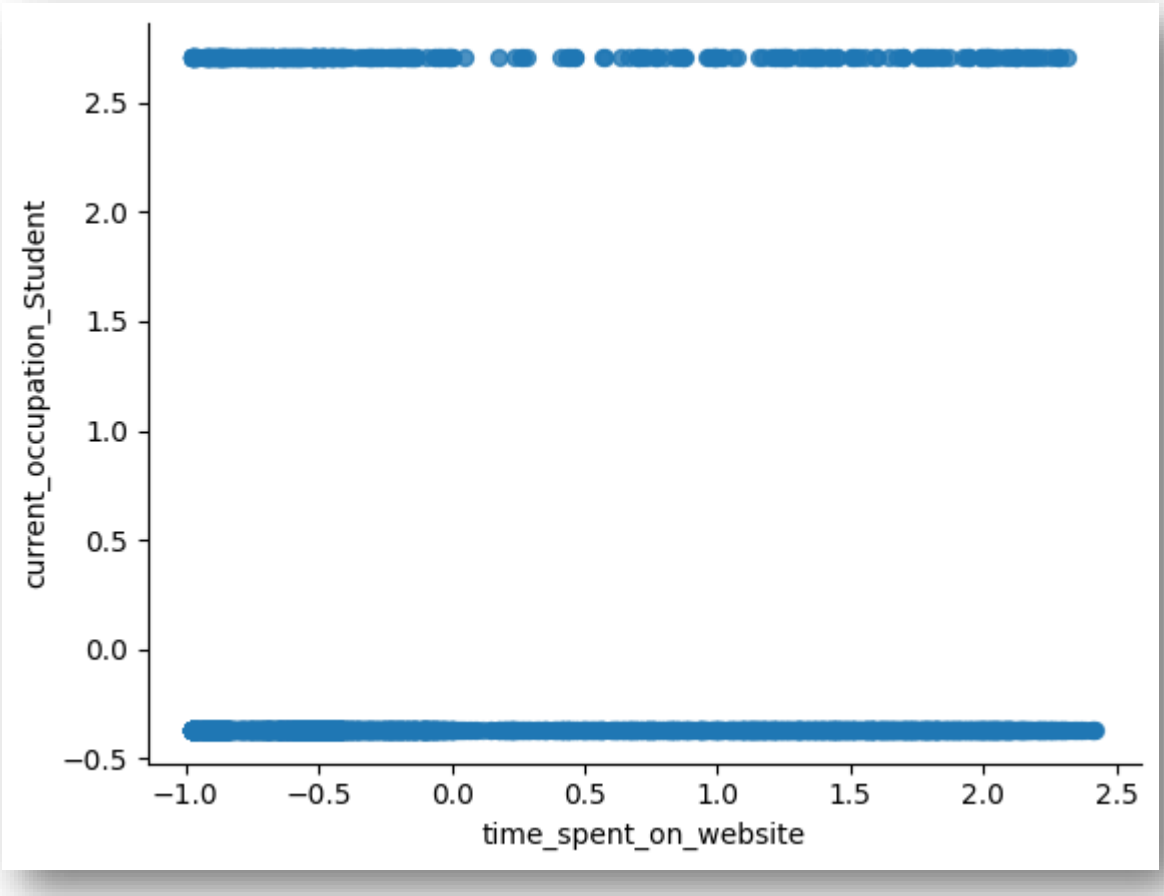


FIGURE 42

Insights:

- The scatterplot shows the relationship between the time spent on the website and current_occupation_student.
- There are two distinct horizontal clusters:

a) Around 0, which likely represents non-students.

b) Around 1, which represents students.

- Both the students and non-students seem to have a similar range of time spent on the website.
- There doesn't appear any significant difference in the time spent between the students and non-students.

♦ Histogram of First_interaction_website:

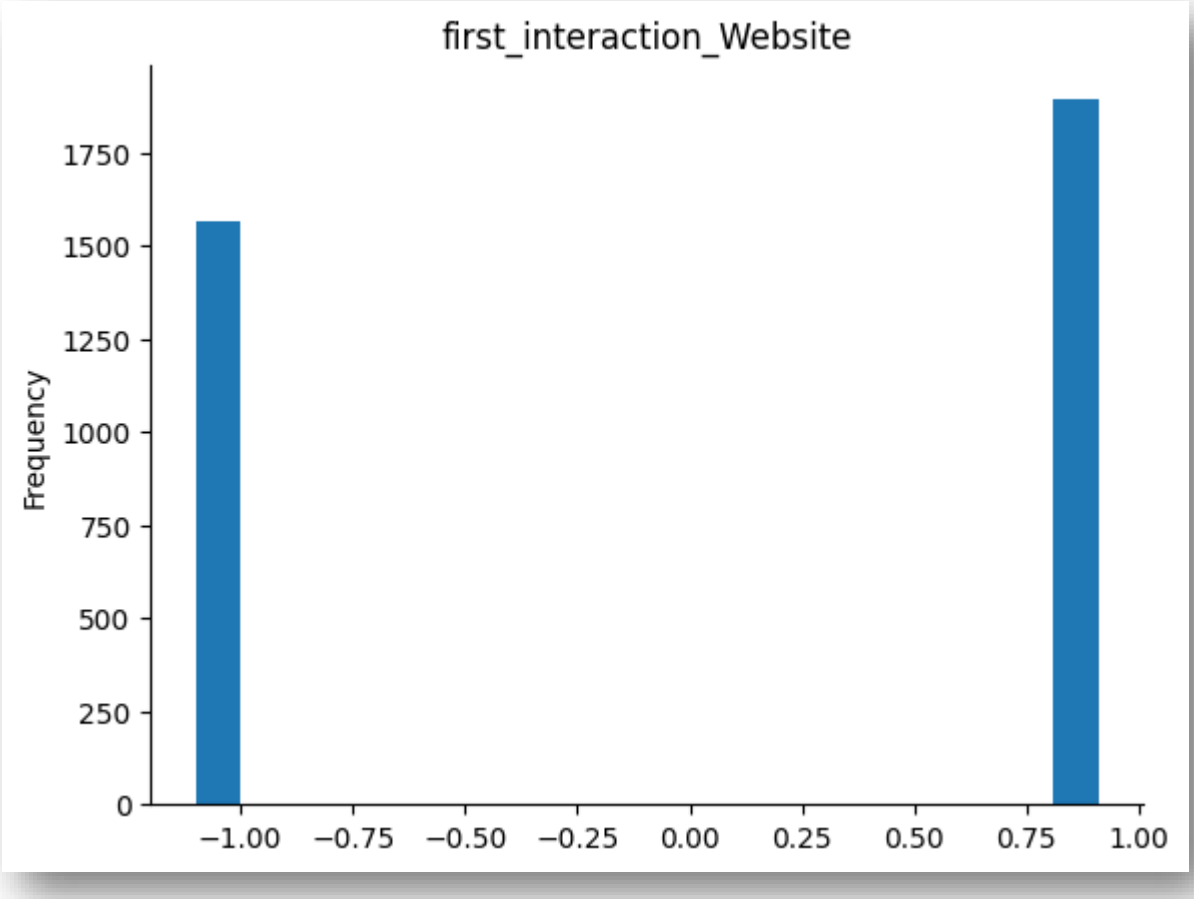


FIGURE 43

Insights:

- The histogram shows the distribution of the variable 'first_interaction_website'.
- The values on the x axis are indicating whether the first interaction was made through website (Yes: 1, No: -1)
- The distribution seems to be bimodal.
- There are two distinct peaks:

a) One at –1, suggesting a substantial number of leads which do not had their first interaction through website.

b) The other at 1, indicating significant number of leads which made their first interaction through website.

- The frequencies of both the peaks are relatively similar indicating a roughly equal split between those whose first interaction was via the website and those whose wasn't.

♦ Histogram of Current_occupation_unemployed:

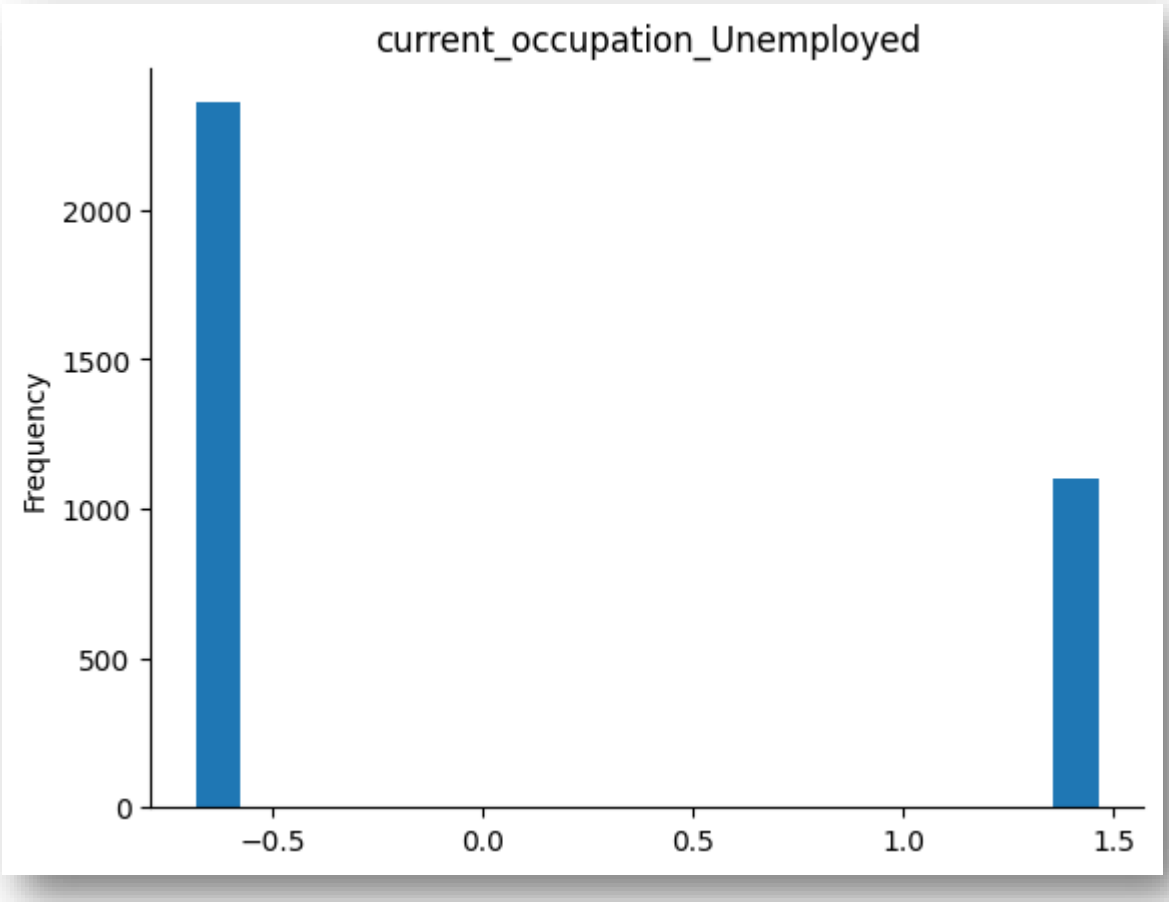


FIGURE 44

Insights:

- The histogram shows the distribution of the variable ‘current_occupation_unemployed’
- The values on the x axis are indicating whether an individual is unemployed (1 for Yes, -1 for No)
- There is a high frequency at –1 indicating a large number of leads are unemployed.
- There is also a significant peak at 1, indicating a large proportion of individuals are unemployed but this group is small.

♦ Histogram of current_occupation_student:

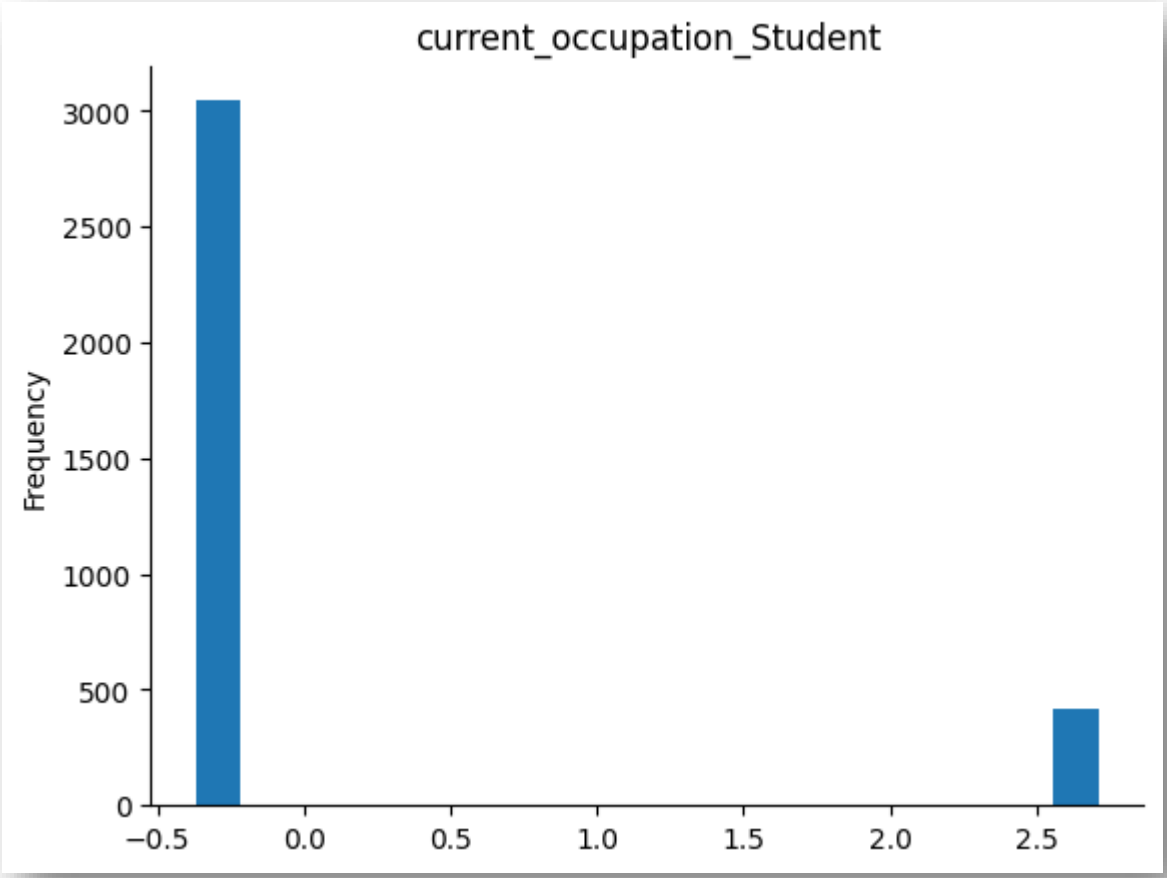


FIGURE 45

Insights:

- The plot shows the frequency distribution of the ‘current_occupation_student’ variable
- There is a noticeable number of data points (over 3,000) corresponding to the category of the students.
- There is a much smaller count for the non-student's category.

♦ **Histogram of time_spent_on_website:**

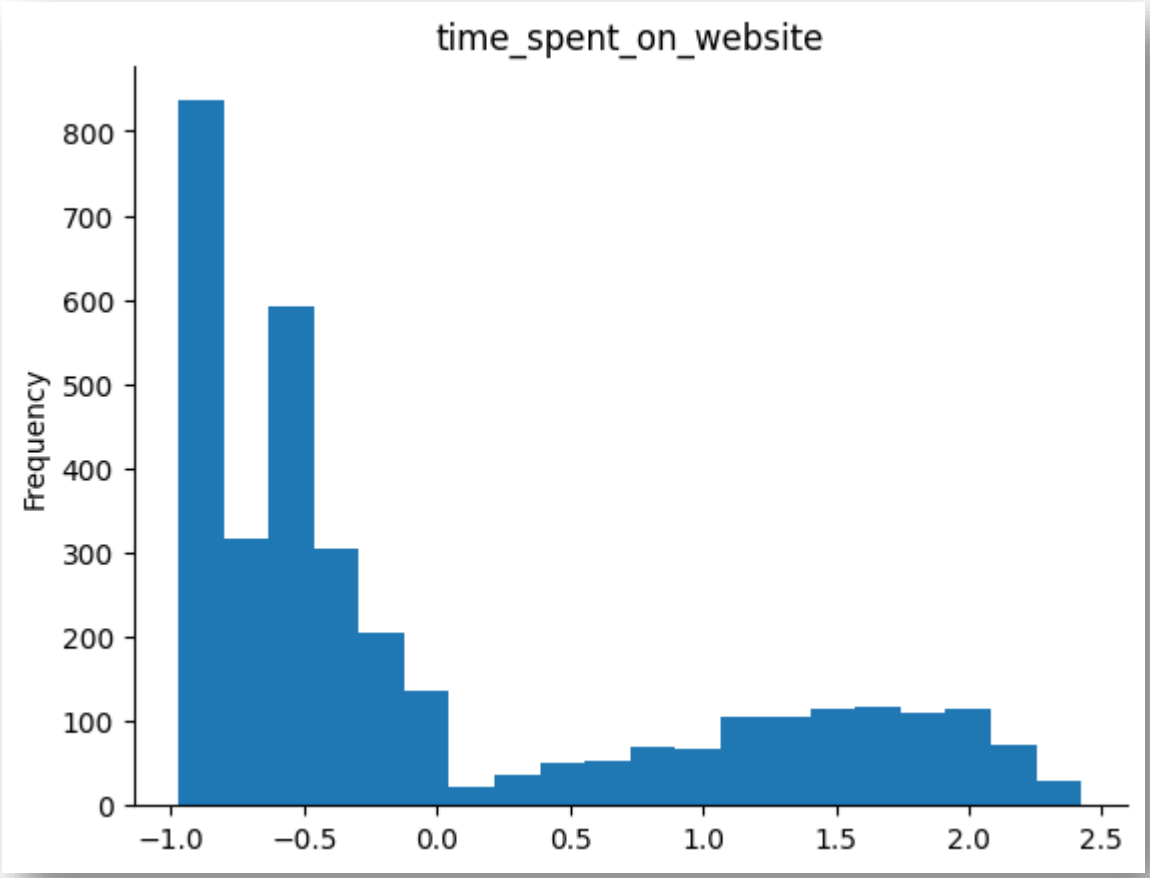


FIGURE 46

Insights:

- The plot shows the frequency distribution of the 'time spent on the website' variable
- The distribution is skewed to the left with a large concentration of the data points around –1 and –0.5
- There is a spread of data points extending to the right, indicating that fewer leads spend more time on the website.

❑ TRAINING THE LOGISTIC REGRESSION MODEL AGAIN

After leaving the rest features and training the logistic regression model again with only the significant features:

♦ Logistic Regression results:

Logit Regression Results

Dep. Variable:	status	No. Observations:	3459
Model:	Logit	Df Residuals:	3449
Method:	MLE	Df Model:	9
Date:	Mon, 29 Jul 2024	Pseudo R-squ.:	0.3539
Time:	05:20:22	Log-Likelihood:	-1363.2
converged:	True	LL-Null:	-2109.8
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.4470	0.059	-24.525	0.000	-1.563	-1.331
time_spent_on_website	0.9609	0.051	18.961	0.000	0.862	1.060
current_occupation_Student	-0.6742	0.063	-10.728	0.000	-0.797	-0.551
current_occupation_Unemployed	-0.2458	0.049	-5.019	0.000	-0.342	-0.150
first_interaction_Website	1.3373	0.060	22.329	0.000	1.220	1.455
profile_completed_Low	-0.3958	0.075	-5.254	0.000	-0.543	-0.248
profile_completed_Medium	-0.8001	0.052	-15.496	0.000	-0.901	-0.699
last_activity_Phone Activity	-0.3022	0.054	-5.639	0.000	-0.407	-0.197
last_activity_Website Activity	0.2186	0.049	4.456	0.000	0.122	0.315
referral_Yes	0.1889	0.048	3.928	0.000	0.095	0.283

TABLE 15

Insights:

- Time_spent_on_website:

a) Positive and significant suggesting that higher time spent on the website increases the chances of the dependent variable being 1.

- Current_occupation_student:

a) Negative and significant, indicating that being a student decreases the chances of the dependent variable being 1.

- Current_occupation_unemployed:

a) Negative and significant, showing that being unemployed reduces the likelihood of the dependent variable being 1.

- First_interaction_website:

a) Positive and significant, suggesting that the first interaction being on the website strongly increases the chances of the dependent variable being 1.

- Profile_completed_low:

a) Negative and significant suggesting that a low level of profile completion decreases the chances of dependent variable being 1.

- Last_activity_PhoneActivity:

a) Negative and significant, indicating that the last activity being phone activity decreases the chances of the dependent variable being 1.

- Profile_completed_medium:

a) Negative and significant, showing that a medium level of profile completion also declines the chances of the dependent variable being 1.

- Last_activity_WebsiteActivity:

a) Positive and significant, suggesting that the last activity being website activity also increases the chance of the dependent variable being 1.

- Referral_Yes:

a) Positive and significant, showing that being referred increases the chance of dependent variable being 1.

Therefore,

→Time spent on the website and first interaction being on the website are strong positive predictors.

→Being student or unemployed negatively impacts the dependent variable.

→Profile completion levels, especially medium have a noticeable negative impact.

→Activities on the website, especially if they are last recorded activity positively influences the dependent variable.

→Referrals have a positive influence on the dependent variable.

♦ Determining the Optimal Threshold using ROC Curve:

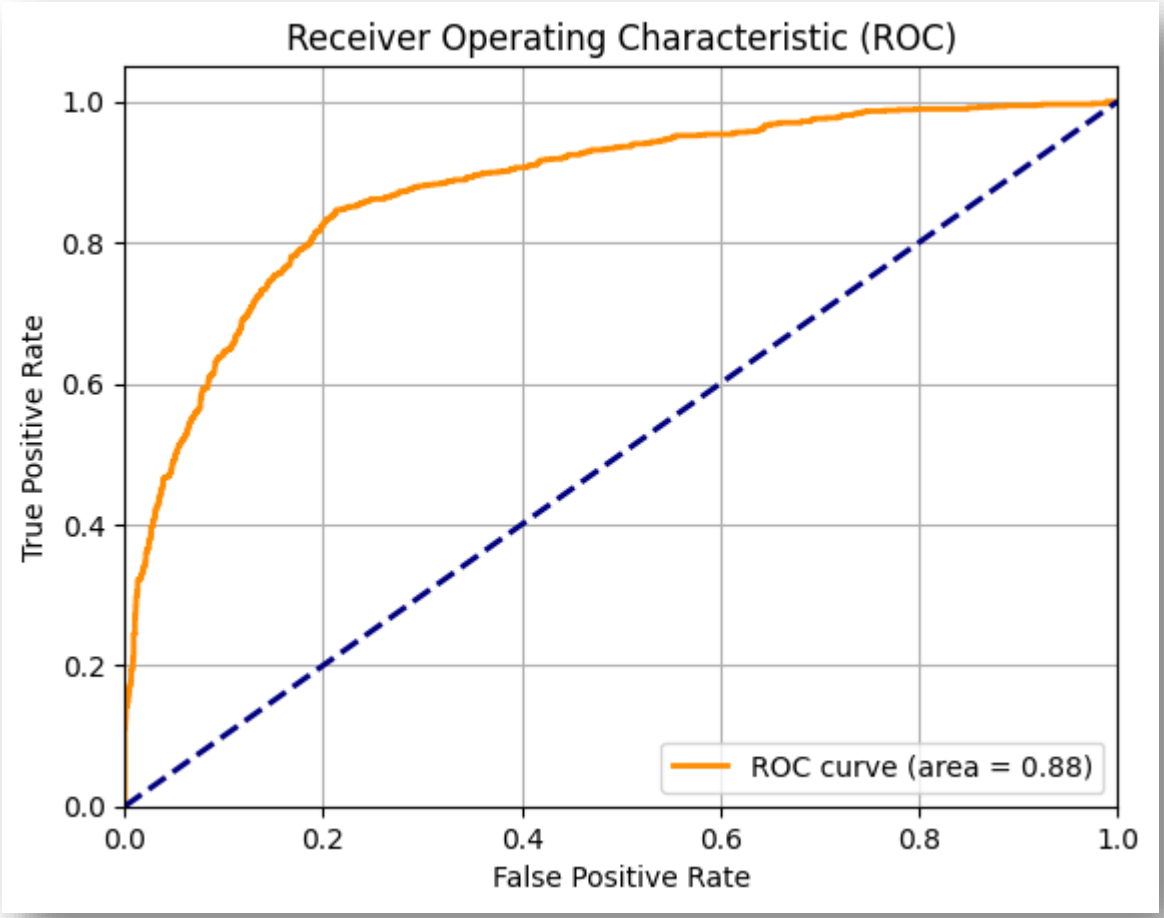


FIGURE 47

Insights:

- True Positive Rate (TPR):

a) Plotted on the y-axis, it is the ratio of correctly predicted positive observations to all the actual positives, also known as recall or sensitivity.

- False Positive Rate (FPR):

a) Plotted on the x-axis, it is the ratio of incorrectly predicted positive observations to all the actual negatives.

- ROC Curve:

a) The ROC Curve represents the performance of the model. The model shows a good performance with AUC of 0.88, suggesting it has a good balance between TPR and FPR.

- Area Under Curve (AUC):

a) The area under the ROC curve is 0.88. This indicates that the model has a good measure of separability and its performing better than a random classifier.

There is an Optimal Threshold of 0.282

➢ **Checking tuned Logistic Regression model performance on the training set:**

	Accuracy	Recall	Precision	F1
0	0.80457	0.84623	0.62859	0.72135

TABLE 16

Observations:

- True Negative (TN): 1908 (55.16%)
- False Positive (FP): 517 (14.95%)
- False Negative (FN): 159 (4.60%)
- True Positive (TP): 875 (25.30%)

Calculations:

→ Accuracy:
(TP+TN)/ Total
= (875+1908)/ (875+1908+517+159) = 2783/3459 ≈ **80.46%**

→ Precision for Class 1:
TP/(TP+FP)
= 875/ (875+517) = 875/1392 ≈ **62.85%**

→Recall for Class 1:
TP/(TP+FN)
= 875/ (875+159) = 875/1034 ≈ **84.62%**

→F1 Score for Class 1:
2*(Precision+Recall)/(Precision+Recall) ≈ **72.0%**

♦ **Plot of Confusion Matrix on the Training set:**

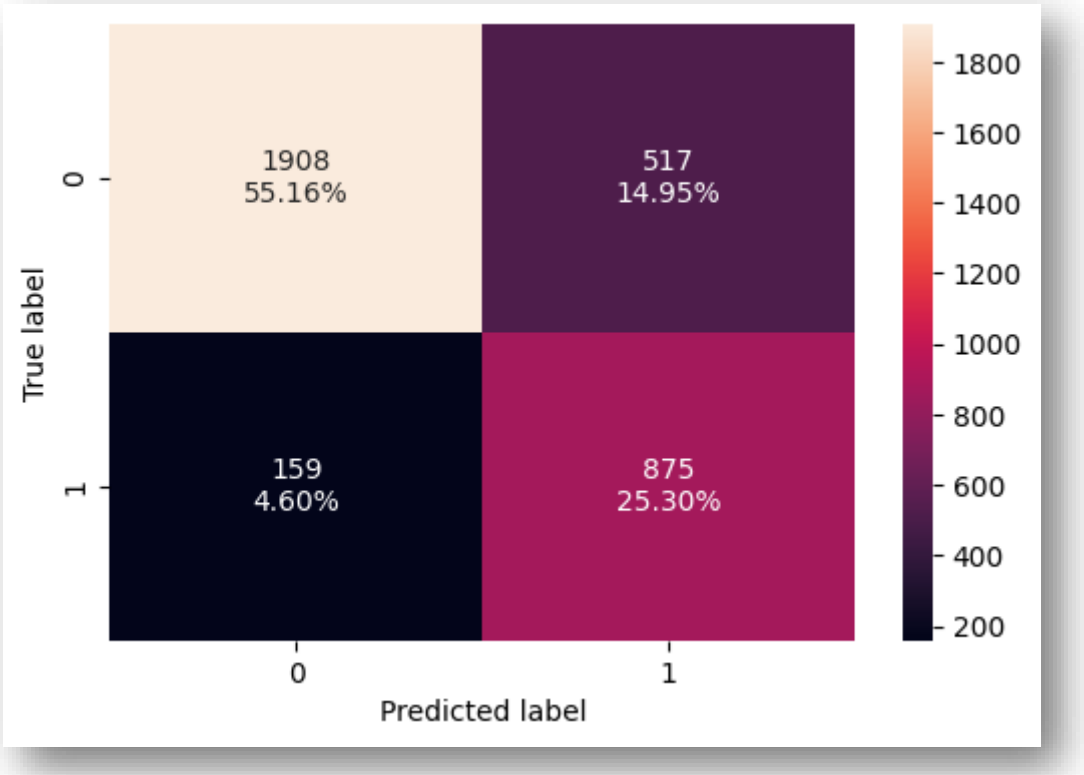


FIGURE 48

Insights:

- The Confusion Matrix shows a good recall but a moderate precision.
- The matrix shows an overall F1 score of about 72%
- The matrix shows that the model is relatively strong in identifying the positive cases (high recall value)
- But has room for improvement in the precision.
- The ROC Curve also supports this by showing a good overall performance.

➢ **Checking tuned Logistic Regression model performance on the test set:**

	Accuracy	Recall	Precision	F1
0	0.79098	0.83090	0.60897	0.70284

TABLE 17

Observations:

- True Negative (TN): 627 (54.38%)
- False Positive (FP): 183 (15.87%)
- False Negative (FN): 58 (5.03%)
- True Positive (TP): 285 (24.72%)

Calculations:

→ Accuracy:
(TP+TN)/ Total
= (285+627)/ (285+183+627+58) = 912/1153 ≈ **79.10%**

→ Precision for Class 1:
TP/(TP+FP)
= 285/ (285+183) = 285/468 ≈ **60.90%**

→Recall for Class 1:

TP/(TP+FN)

= 285/ (285+58) = 285/343 ≈ **83.09%**

→F1 Score for Class 1:

2*(Precision+Recall)/(Precision+Recall) ≈ **70.75%**

♦ Plot of Confusion Matrix on the Test set:

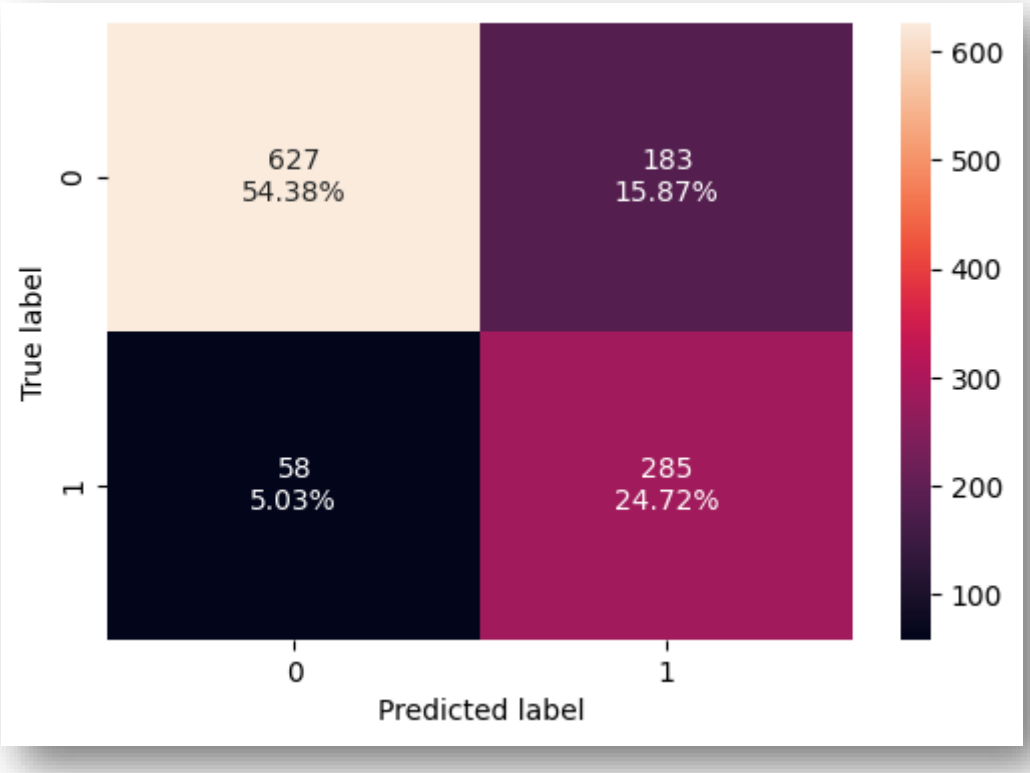


FIGURE 49

Insights:

- The Confusion Matrix on the test set shows a good recall but a slightly lower precision.

- The overall F1 score is about 70.75%
- The matrix indicates that the model is relatively strong in identifying the positive cases.
- This matrix has also a chance of improvement in precision.

❑ KNN CLASSIFIER

KNN CLASSIFIER Performance Improvement using different K values

The range of k values between 2 to 20

The recall values for the different k values in a KNN model indicates how well the model identifies true positives for the various values of k.

```
Recall for k=2: 0.43440233236151604
Recall for k=3: 0.597667638483965
Recall for k=4: 0.5043731778425656
Recall for k=5: 0.6239067055393586
Recall for k=6: 0.5539358600583091
Recall for k=7: 0.6530612244897959
Recall for k=8: 0.5918367346938775
Recall for k=9: 0.6588921282798834
Recall for k=10: 0.5743440233236151
Recall for k=11: 0.641399416909621
Recall for k=12: 0.5685131195335277
Recall for k=13: 0.6209912536443148
Recall for k=14: 0.565597667638484
Recall for k=15: 0.6064139941690962
Recall for k=16: 0.5714285714285714
Recall for k=17: 0.6297376093294461
Recall for k=18: 0.5860058309037901
Recall for k=19: 0.6180758017492711
Recall for k=20: 0.597667638483965

The best value of k is: 9 with a recall of: 0.6588921282798834
```

TABLE 18

Insights:

- Optimal K- value:

a) The highest recall is at k= 9 with a recall value of 0.6588921. This suggests that at k=9, the model is best at identifying true positives.

- High recall values:

a) Other recall values are at k= 5, k= 13 and k= 19. These k values provide decent recall performance but are not optimal.

- Low recall values:

a) The lowest recall value is at k= 2, indicating poor performance in identifying true values, suggesting that using very low k is ineffective.
b) Other low recall values include k= 4, k= 12, k= 14. These k values show relatively poor performance

- Best K:

a) The best K value is k= 9, indicating that the model has ability to correctly identify the positive instances is maximized at this point.

➢ **Checking tuned KNN model performance on the training set:**

	Accuracy	Recall	Precision	F1
0	0.85256	0.68182	0.79571	0.73437

TABLE 19

Observations:

- True Negative (TN): 2244
- False Positive (FP): 181
- False Negative (FN): 329
- True Positive (TP): 705

Calculations:

→ Accuracy:

(TP+TN)/ Total

= (2244+705)/ (2244+181+329+705) = 2949/3459 ≈ **85.23%**

→ Precision for Class 1:

TP/(TP+FP)

= 705/ (705+181) = 705/886 ≈ **79.62%**

→Recall for Class 1:

TP/(TP+FN)

= 705/ (705+329) = 705/1034 ≈ **68.19%**

→F1 Score for Class 1:

2*(Precision+Recall)/(Precision+Recall) ≈ **73.43%**

♦ **Plot of Confusion Matrix on the Training set:**

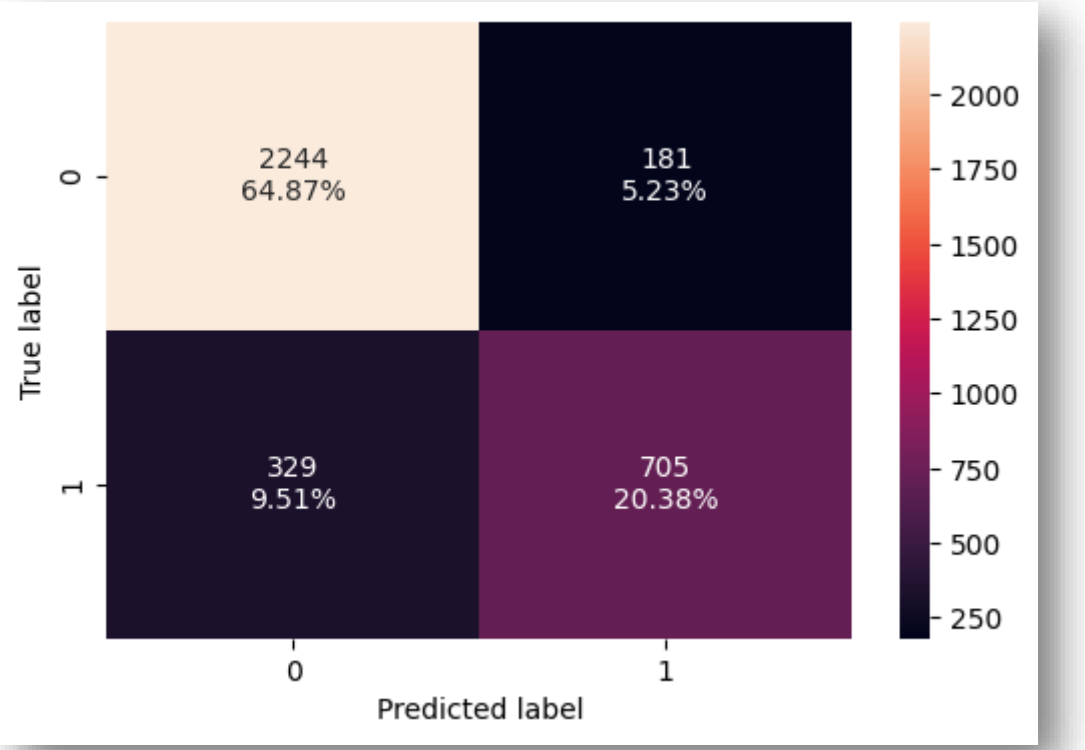


FIGURE 50

Insights:

- The model has a high overall accuracy indicating it generally performs well in predicting both the classes.
- The precision indicates that most of the positive predictors are correct but there is still significant proportion of false positives
- The recall is moderate, implying that the model misses about 31.81% of the actual positive cases.

➢ **Checking tuned KNN model performance on the test set:**

	Accuracy	Recall	Precision	F1
0	0.83695	0.65889	0.76094	0.70625

TABLE 20

Observations:

- True Negative (TN): 2244
- False Positive (FP): 181
- False Negative (FN): 329
- True Positive (TP): 705

Calculations:

→ Accuracy:
(TP+TN)/ Total
= (2244+705)/ (2244+181+329+705) = 2949/3459 ≈ **85.2%**

→ Precision for Class 1:
TP/(TP+FP)
= 705/ (705+181) = 705/886 ≈ **79.62%**

→Recall for Class 1:
TP/(TP+FN)

= 705/ (705+329) = 705/1034 ≈ **68.2%**

→F1 Score for Class 1:

2*(Precision+Recall)/(Precision+Recall) ≈ **73.5%**

→False Positive Rate (FPR):

=181/ (181+2244) = 181/2425 ≈ **7.4%**

→False Negative Rate (FNR):

=329/ (329+705) = 329/1034 ≈ **31.8%**

♦ Plot of Confusion Matrix on the Test set:

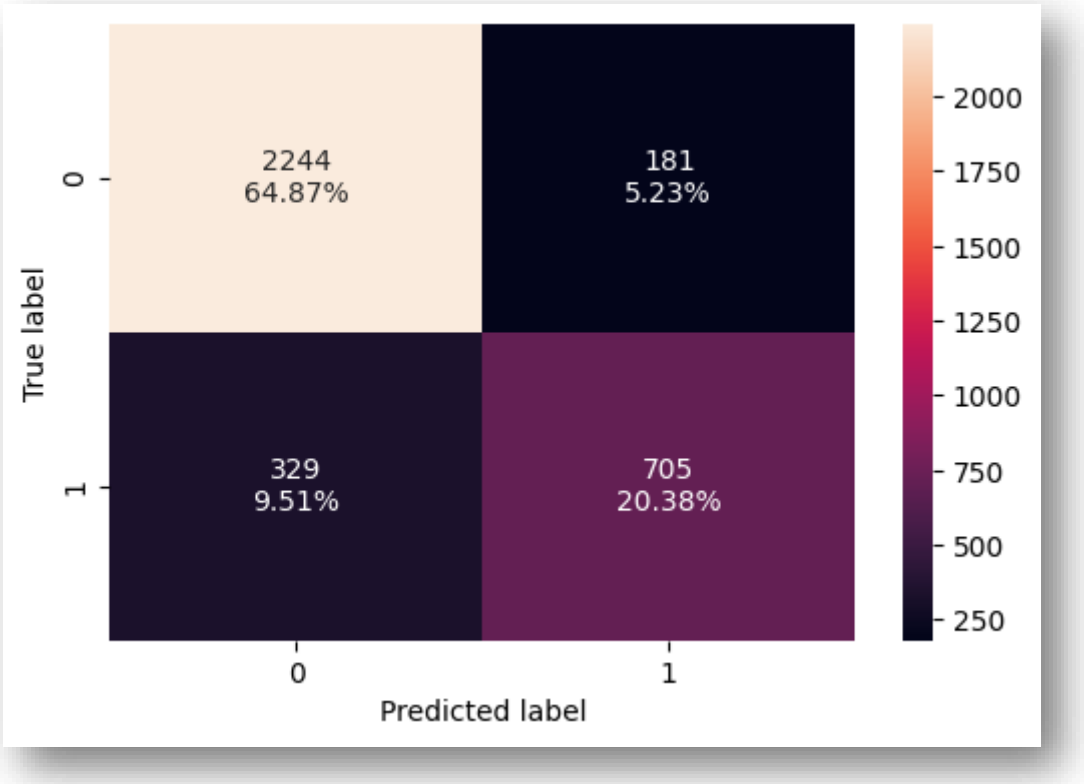


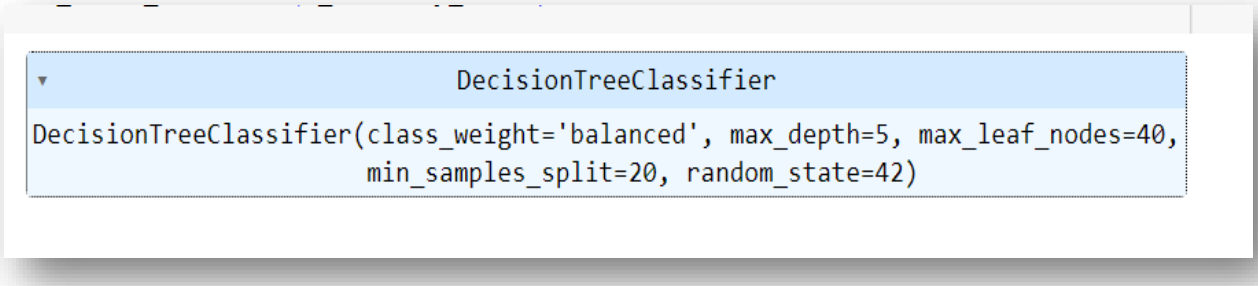
FIGURE 51

Insights:

- The model has a good accuracy and precision suggesting that it performs well in predicting both the classes.
- However, the recall is relatively lower meaning it misses a significant number of positive cases.
- The F1 score which balances the precision and recall, shows a reasonably good performance, though there is room for improvement.

❑ DECISION TREE CLASSIFIER (PRE- PRUNING)

Pre pruning the tree:



➢ Checking tuned Decision Tree Classifier performance on the training set:

	Accuracy	Recall	Precision	F1
0	0.83463	0.88685	0.66837	0.76226

TABLE 21

Observations:

- True Negative (TN): 1970
- False Positive (FP): 455
- False Negative (FN): 117
- True Positive (TP): 917

Calculations:

→ Accuracy:
(TP+TN)/ Total
= (917+1970)/ (1970+455+117+917) = 2887/3459 ≈ **83.5%**

→ Precision for Class 1:
TP/(TP+FP)
= 917/ (917+455) = 917/1372 ≈ **88.7%**

→Recall for Class 1:
TP/(TP+FN)
= 917/ (917+117) = 917/1034 ≈ **88.7%**

→F1 Score for Class 1:
2*(Precision+Recall)/(Precision+Recall) ≈ **76.1%**

→False Positive Rate (FPR):
= 455/ (455+1970) = 455/2425 ≈ **18.7%**

→False Negative Rate (FNR):
= 117/ (117+917) = 117/1034 ≈ **11.3%**

♦ Plot of Confusion Matrix on the Training set:

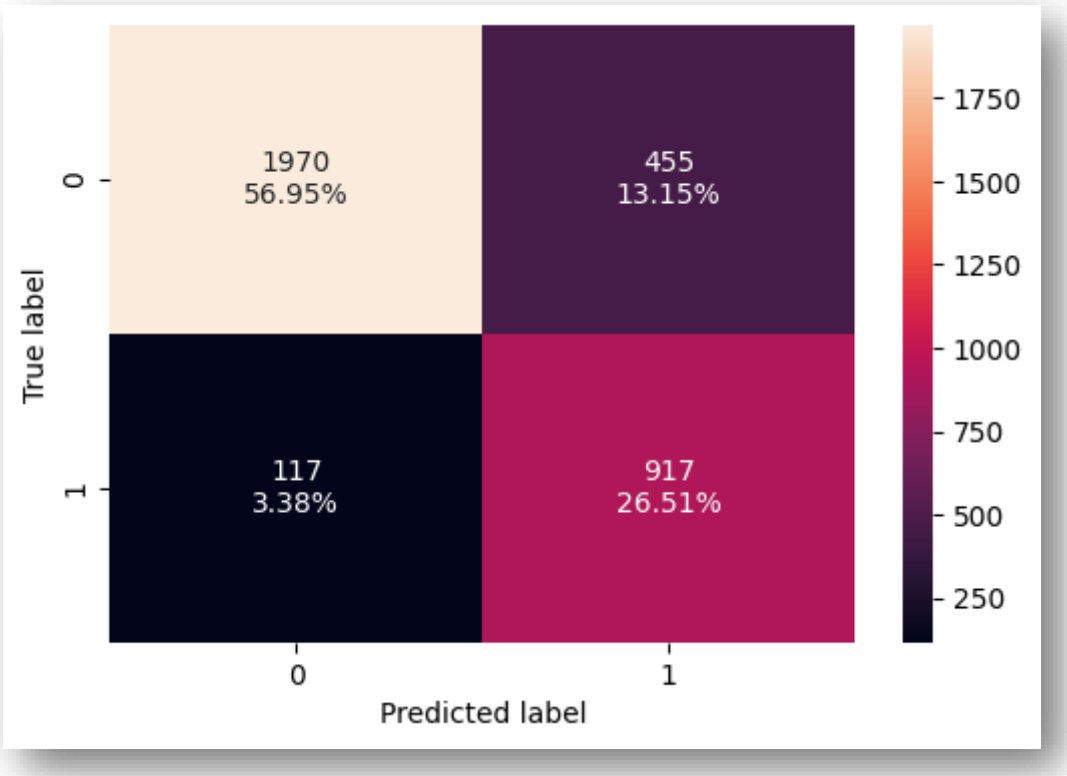


FIGURE 52

Insights:

- The model has a good accuracy and recall indicating it performs well in identifying the positive cases.
- However, the precision is relatively lower, showing significant number of positive predictors are incorrect.
- The F1 score which balances the precision and recall shows a good performance but still there is a scope for improvement especially in reducing the false positives.

➢ **Checking tuned Decision Tree Classifier performance on the test set:**

	Accuracy	Recall	Precision	F1
0	0.83955	0.89213	0.67401	0.76788

TABLE 22

Observations:

- True Negative (TN): 662
- False Positive (FP): 148
- False Negative (FN): 37
- True Positive (TP): 306

Calculations:

→ Accuracy:

(TP+TN)/ Total

= (662+306)/ (662+148+37+306) = 968/1153 ≈ **84.0%**

→ Precision for Class 1:

TP/(TP+FP)

= 306/ (306+148) = 306/454 ≈ **67.4%**

→Recall for Class 1:

TP/(TP+FN)

= 306/ (306+37) = 306/343 ≈ **89.2%**

→F1 Score for Class 1:

2*(Precision+Recall)/(Precision+Recall) ≈ **76.8%**

♦ **Plot of Confusion Matrix on the Test set:**

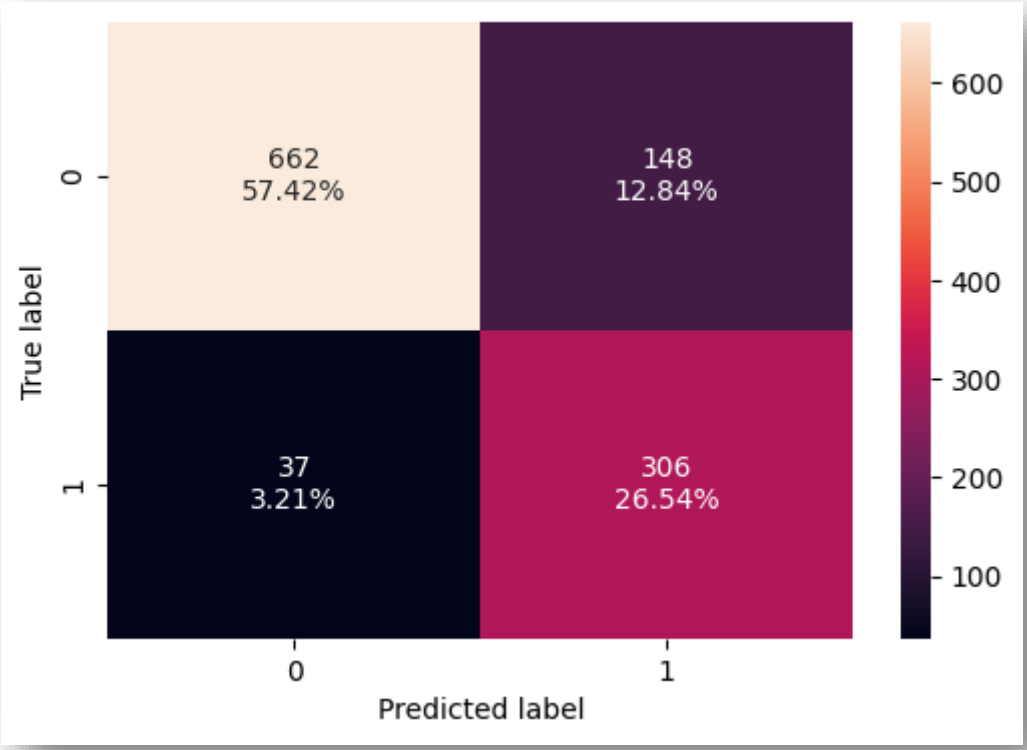


FIGURE 53

Insights:

- The model has an accuracy of approx. 84.0% indicating that it correctly classifies majority of the instances on the test set.
- Precision of 67.4% suggests that among the instances predicted as positive 67.4% are actually positive. This shows that there are some false positives but majority of the predicted are correct.

- Recall of 89.2% indicates that the model successfully indicates 82.9% of the actual positive instances which is quite high.
- There are 148 instances where the model predicted positive, but the actual label was negative.
- And there are 37 instances where the model predicted negative, but the actual label was positive.
- The confusion matrix shows a strong model performance on the test set with a high accuracy, good precision and an excellent recall.

❑ VISUALIZING THE DECISION TREE

The Decision Tree visualization represents a classification of models' decision process, displaying the splits based on different features to predict the target variable.

♦ Decision Tree:

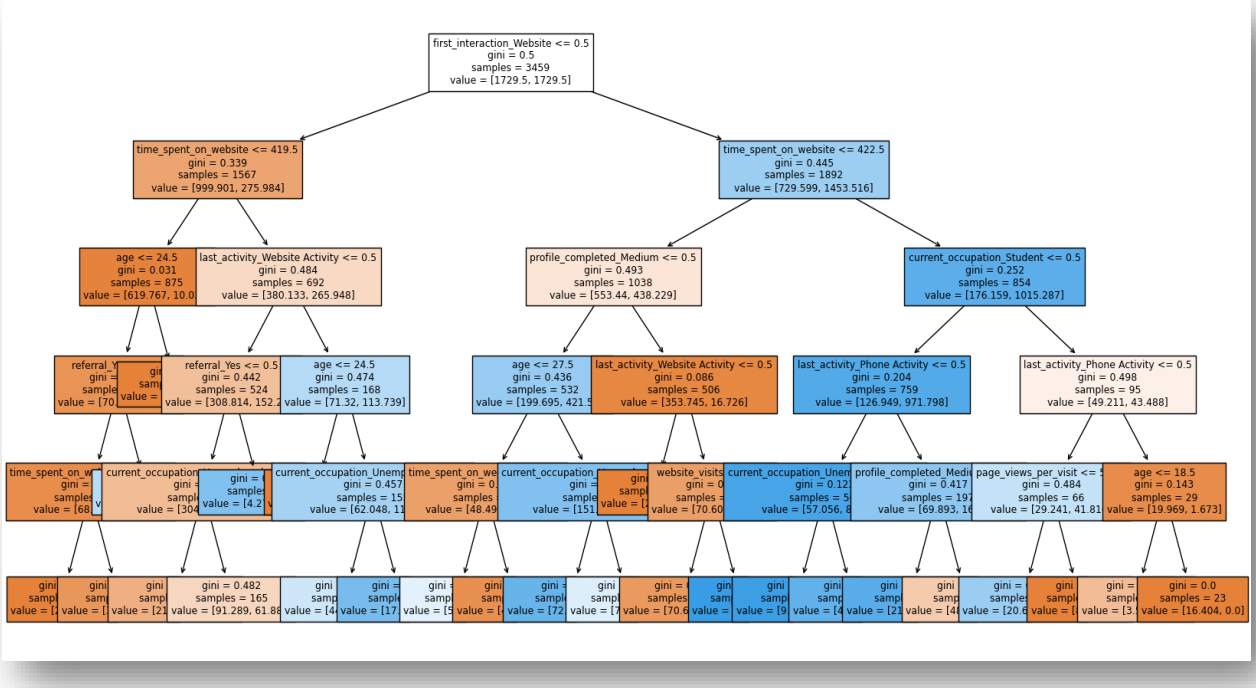


FIGURE 54

Insights:

- **Root node:**

a) Feature - ‘first_interaction_website’: This split suggests whether leads first interaction was via the website is the most important feature in the dataset for the initial split.

- **First level splits:**

a) Left split- ‘time_spent_on_website’

b) Right split- ‘first_interaction_website’: This indicates that after determining the first interaction, the amount of time spent on the website is critical for classifying the data.

- **Second level splits:**

a) Left split of split- ‘Age’:
b) Right split of left split- ‘last_activity_WebsiteActivity
c) Left split of right split- ‘profile_completed_medium’
d) Right split of right split- current_occupation_student':
These splits show the importance of age, last activity on the website, profile completion, and current occupation in further classifying the leads.

- **Deeper splits:**

a) Various branches further splits based on the additional features such as referral, current_occupation_unemployed, website_visits, last_activity_PhoneActivity, profile_completed_medium.

Insights:

- First interaction and time spent: The initial interaction method and time spent on the website are significant factors in predicting the leads behavior.
- Younger leads (Age<= 24.5) with significant website activity are more likely to show distinct behavior.
- Leads with medium profile completion and specific occupation (student or unemployed) have different behavior patterns.
- The type of last activity significantly influences the classification suggesting different engagement levels.
- Further splits on features like referral, website visits, page views per visit shows that leads engagement metrics shows small insights into user classification.

❑ MODEL PERFORMANCE COMPARISON & FINAL SELECTION

The table shows the performance metrics (accuracy, precision, recall, F1) for several classification models, both base and tuned versions on a dataset.

➢ PERFORMANCE COMPARISON OF MODELS ON TRAINING SET:

	Logistic regression base	Logistic regression tuned	Naive Bayes base	KNN base	KNN tuned	Decision Tree base	Decision Tree tuned
Accuracy	0.82249	0.80457	0.78925	0.88927	0.85256	0.99971	0.83463
Recall	0.65377	0.84623	0.76886	0.77466	0.68182	0.99903	0.88685
Precision	0.72532	0.62859	0.61868	0.84227	0.79571	1.00000	0.66837
F1	0.68769	0.72135	0.68564	0.80705	0.73437	0.99952	0.76226

TABLE 23

Insights:

1. **Accuracy:**

- Highest Accuracy: The Decision Tree Base model has the highest accuracy at 0.99971, suggesting it almost perfectly predicts the outcomes from the training data. This shows the potential overfitting.
- Lowest Accuracy: The Naive Bayes Base model has the lowest accuracy at 0.78925 suggesting it correctly predicts outcomes for about 78.93% of the training data.
- KNN Base vs KNN Tuned: The accuracy of the KNN Model decreases slightly from 0.88927 to 0.85256

2. **Recall:**

- Highest Recall: The Decision Tree Base model achieves the highest recall at 0.99903, suggesting it captures nearly all the actual positive cases suggesting overfitting.
- Lowest Recall: The KNN Tuned model has the lowest recall at 0.68182, missing significant number of actual positive cases.
- The recall for logistic regression improved significantly with tuning increasing from 0.65377 to 0.84623

3. **Precision:**

- Highest Precision: The Decision Tree Base model has 1.0000 suggesting all positive predictors are correct indicating overfitting.
- Lowest Precision: The logistic regression tuned model has the lowest precision at 0.62859 showing it has a higher rate of false positives.

4. **F1 Score:**

- Highest F1 score: The Decision Tree Base model has the highest F1 score at 0.99952, balancing precision and recall.
- Lowest F1 score: The Naive Bayes Base model has the lowest F1 score at 0.68564, showing lower performance in both precision and recall.
- The F1 score for logistic regression improved with tuning, going from 0.68769 to 0.72135 suggesting overall better performance.

Overall Insights:

1. **Decision Tree Base Model:**
   - Performs the best on the training set with the highest accuracy, recall, precision and F1.
   - These metrics show severe overfitting, where the model memorizes the training data and may not work well with the new unseen data.
2. **KNN Models:**
   - KNN Base shows the highest accuracy and a balanced recall and precision suggesting a good fit on the training data.
   - In KNN Tuned, there is a slight decrease in the accuracy and precision.
3. **Logistic Regression:**
   - Base shows moderate performance with a F1 score of 0.68769

- Tuned: Improvement in recall and F1 score but a decrease in precision, showing model has become sensitive and produces more false positives.

4. **Naive Bayes:**

- Base shows the weakest performance in all the metrics compared to other models, with F1 score of 0.68564 suggesting it is less effective on the training set.

5. **Base vs Tuned Model:**

- Tuning improves the recall and F1 score of the models but may decrease precision.
- The Decision Tree model when tuned, shows significant reduction in overfitting.

❑ PERFORMANCE COMPARISON OF MODELS ON TEST SET:

The table shows the performance metrics (accuracy, precision, recall, F1) for several classification models, both base and tuned versions on a dataset.

	Logistic regression base	Logistic regression tuned	Naive Bayes base	KNN base	KNN tuned	Decision Tree base	Decision Tree tuned
Accuracy	0.82249	0.80457	0.77971	0.79879	0.83695	0.81526	0.83955
Recall	0.65377	0.84623	0.75802	0.59767	0.65889	0.67638	0.89213

Precision	0.72532	0.62859	0.60325	0.68562	0.76094	0.69461	0.67401
F1	0.68769	0.72135	0.67183	0.63863	0.70625	0.68538	0.76788

TABLE 24

Insights based on the comparison:

1. **Accuracy:**

- Highest Accuracy: The decision Tree tuned model has the highest accuracy at 0.83955, suggesting that it correctly predicts the outcomes for approx. 83.96% of the test data.
- Lowest Accuracy: The Naive Bayes Base model has the lowest accuracy at 0.77971, indicating it correctly predicts the outcome for about 77.99% of the test data.
- KNN Tuned vs KNN Base: Tuning the KNN model improves its accuracy from 0.79879 to 0.83695

2. Recall:

- Highest Recall: The Decision Tree Tuned model achieves the highest recall at 0.89213, suggesting it effectively captures a high proportion of the actual positive cases.
- The KNN Base model has the lowest recall at 0.59767, showing that it misses a significant number of the actual positive cases.
- The recall for logistic regression improved with tuning, increasing from 0.65377 to 0.84623

3. Precision:

- Highest Precision: The KNN Tuned model has the highest precision at 0.76094, suggesting that a high proportion of the predictors are correct.
- Lowest Precision: The Naive Bayes Base model has the lowest precision at 0.60325, implying it has a higher rate of false positives.
- Impact of Tuning: Tuning the logistic regression model decreased its precision from 0.72532 to 0.62859, which indicates a tradeoff between recall and precision.

4. F1 Score:

- Highest F1 score: The Decision Tree Tuned model has the highest F1 score at 0.76788 balancing the precision and recall effectively.
- Lowest F1 score: The KNN Base Model has the lowest F1 score at 0.63863, showing lower performance in both precision and recall.
- The F1 score for logistic regression improved with tuning, moving from 0.68769 to 0.72135, indicating better overall performance.

Overall Insights:

1. Decision Tree Tuned Model:

- Performs the best with highest accuracy, recall and F1 score.
- Precision is relatively good, making it a well-balanced model.

2. KNN Tuned Model:

- Shows a noticeable improvement over its base version in all the metrics.

- Has the highest precision among all the models making it effective in reducing false positives.
- Recall improved but it is still relatively low compared to the Decision Tree Tuned.

3. **Logistic Regression:**

- Tuning significantly improved the recall and F1 score.
- However, precision decreased, suggesting more false positives.

4. **Naive Bayes:**

- It shows the weakest performance in all the metrics compared to other models.
- Precision and recall are moderate leading to an overall lower F1 score.

5. **Base vs Tuned Models:**

- Decision Tree and KNN model benefit the most from tuning, with substantial improvements in performance metrics.
- Tuning generally improves the recall, accuracy and F1 score of the models but may affect precision negatively.

❑ ACTIONABLE INSIGHTS & RECOMMENDATIONS:

Based on the data set and various performance metrics from the model, here are the actionable insights and recommendations for ExtraaLearn:

Actionable Insights:

1. First Interaction Channels:

- Leads who first interacted through the website tend to convert at a higher rate compared to those who first started their interaction through the mobile app.
- The average number of website visits is approx. 357 with a standard deviation of 2.83
- The average time spent on the website is around 724 seconds. (approx. 12 minutes)
- The average number of page views per visit is 3.03
- This indicates that the website experience might be more effective in engaging the prospects initially.

2. Age and Occupation Impact:

- Professionals and unemployed individuals show higher engagement and conversion rates. They may be more motivated to enhance their skills for career advancement or re- entry into the job market.
- Younger leads (users) particularly students are less likely to get converted compared to the older users. This might be due to a higher tendency for students to explore but not commit.
- The average age of prospects is 46 years, though the prospect age range is from 18 to 63 years.

3. Profile Completion:

- Profile completion levels vary, but a significant portion has ‘high’ or ‘medium’ completion levels.

- High and medium profile completion rates are associated with higher conversion rates. Encouraging users to complete their profiles could be beneficial.

4. Website Engagement:

- Time spent on the website and page views per visit are significant predictors of conversion. Users who spend more time and view more pages are more likely to get converted.
- Last activity type also influences conversion with 'website activity' being more indicative of the conversion compared to other activities like 'email activity'.

5. Referral and Media Channels:

- Print media does not show a strong correlation with conversion suggesting it may be less effective for this target audience.
- Users/ leads referred by others are exposed to digital media campaigns and tend to have a higher conversion rate.

6. Conversion Status:

a) The average status score is approx. 0.30 suggesting a lower conversion rate.

Recommendations:

1. Enhanced Website Experience:

- Implement features that encourage longer sessions and more page views such as interactive content, personalized recommendations and engaging visuals to attract.
- Optimize the website user experience to ensure it is engaging as well as informative, since it is a critical channel for initial interactions.

2. Profile Completion Incentives:

- Simplify the profile completion process to make it quicker and more user friendly.
- Develop incentives to encourage users to complete their profiles, such as offering discounts, free trial periods or access to premium content.

3. Referral Programs:

- Use gamification elements to make the referral process fun and engaging.
- Strengthen referral programs by offering rewards to users who refer friends or colleagues.
- Provide easy-to-use sharing options on the website and mobile app.

4. Leverage Digital Media:

- Track and analyze the performance of different digital channels to allocate budget effectively.
- Invest more in digital media campaigns as they have shown to be more effective in gaining conversions.

5. Monitor and Improve Last Activity Engagement:

- Track users last activity and maintain follow-up communications based on these actions, and adjust strategies based on these insights.
- For users whose last activity was ‘E mail activity’ consider a personalized Email follow- ups to re-engage them and guide them back to the website.
- Establish a feedback loop with users to gather insights on their preferences and areas for improvement.

6. Utilize Educational Channels:

- Offer more educational content such as webinars, tutorials and how to guides to attract and retain users.
- Clearly communicate the benefits of engaging with the educational content to the prospects.

✓ Based on these insights, focusing on improving the user engagement on the website, optimizing the initial interaction experience enhancing targeted marketing efforts can significantly increase user engagement,

improve profile completion rates and boost overall conversion rates for ExtraaLearn.

- ✓ Additionally, leveraging the strengths of the tuned decision tree models can help in accurately identifying the potential converts and customizing the marketing strategies accordingly.