Extraalearn Project

Supervised Learning-Classification Project

Contents

Business Problem Overview and Solution Approach

The Problem

- The online education market was projected to reach $286.62 billion by the year 2023, with a 10.26% annual growth rate from 2018-2023.
- This market is preferred over traditional education and offers benefits like easy information sharing, personalized learning, and transparent assessments.
- The online education sector keeps attracting new customers, leading to the emergence of many new companies.
- EdTech companies generate leads through digital marketing, social media, websites, and email, then nurture these leads to convert them into paying customers.
- EXTRAALEARN is a new EdTech company interested in knowing which leads are more likely to convert, this will help the company plan their resources properly.
- EXTRAALEARN has supplied data that can be analysed in order to make informed decision.

Solution Approach

- The approach is to draw insight from the data provided, perform exploratory data analysis, make important observations, and build a model that can help predict which lead is more likely to be converted.

Data Background and Content

The data contains the different attributes of leads and their interaction details with ExtraaLearn. The data has 4612 rows and 15 columns. The columns types are integer, object and float (4 int, 10 object and 1 float). The data has the following columns:

- ID: ID of the lead
- age: Age of the lead
- current_occupation
- First_interaction
- Profile_completed
- website_visits
- time_spent_on_website
- page_views_per_visit
- last_activity
- print_media_type1
- print_media_type2
- digital_media
- educational_channels
- referral
- status

There are no duplicates in the dataframe.

EDA Results

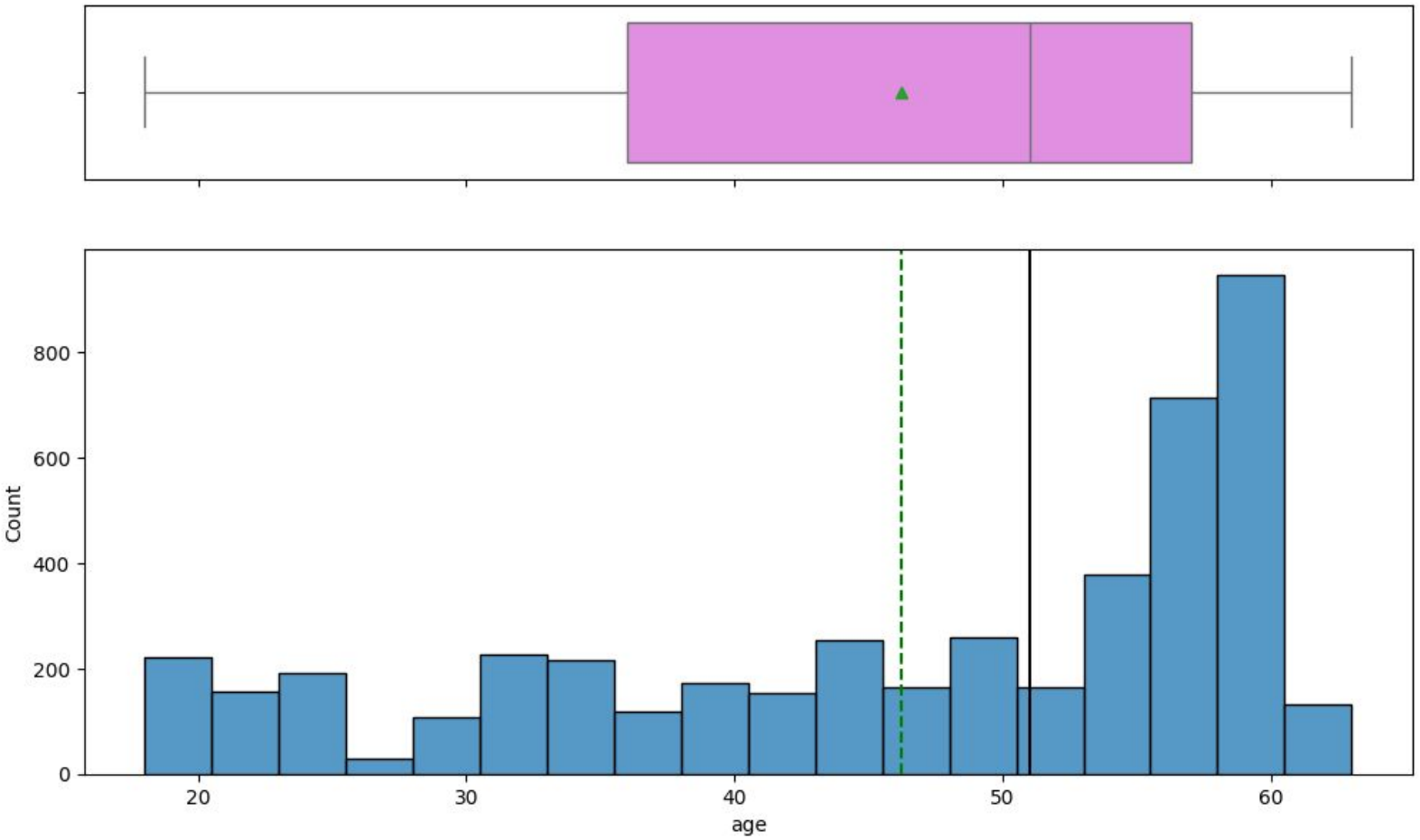
- The average age of leads is about 46.2 years, the minimum age is 18 while the maximum is 63. This sounds reasonable for students and professional. Adults beyond 63 years of age are less likely to want to upskill.
- There are more professional leads than unemployed leads or student lead.
- Most of the leads (2542 of them) first interacted with EXTRAALEARN via the company's website
- Most of the leads (2264 of them) completed at least 75 percent of their profile on EXTRAALEARN's website/mobile app
- The number of times lead visited the website ranges from 0 to 30. It is possible that those that did not visit the website used mobile app.
- The time spent by leads on the website ranges from 0 to 2537 unit time.
- The number of pages visited on the website ranges from 0 to 18.43
- The last interaction between leads and EXTRAALEARN was mostly by Email
- Most leads had not seen the ad of the EXTRAALEARN in the Newspaper
- Most leads had not seen the ad of the EXTRAALEARN in the Magazine
- Most leads had not seen the ad of the EXTRAALEARN in the digital media
- Most leads had not seen the ad of the EXTRAALEARN in educational channels
- Most leads did not hear about EXTRAALEARN through referral
- For the last attribute "status", it is not clear from the data whether being converted was coded as 1, while not being converted was coded as 0. Otherwise, at least 70percent of leads were not converted

EDA Results

- The number of unique values is 4612, meaning that each column has a unique identity. We can drop the ID column as this gives no special information about the data.

EDA Results

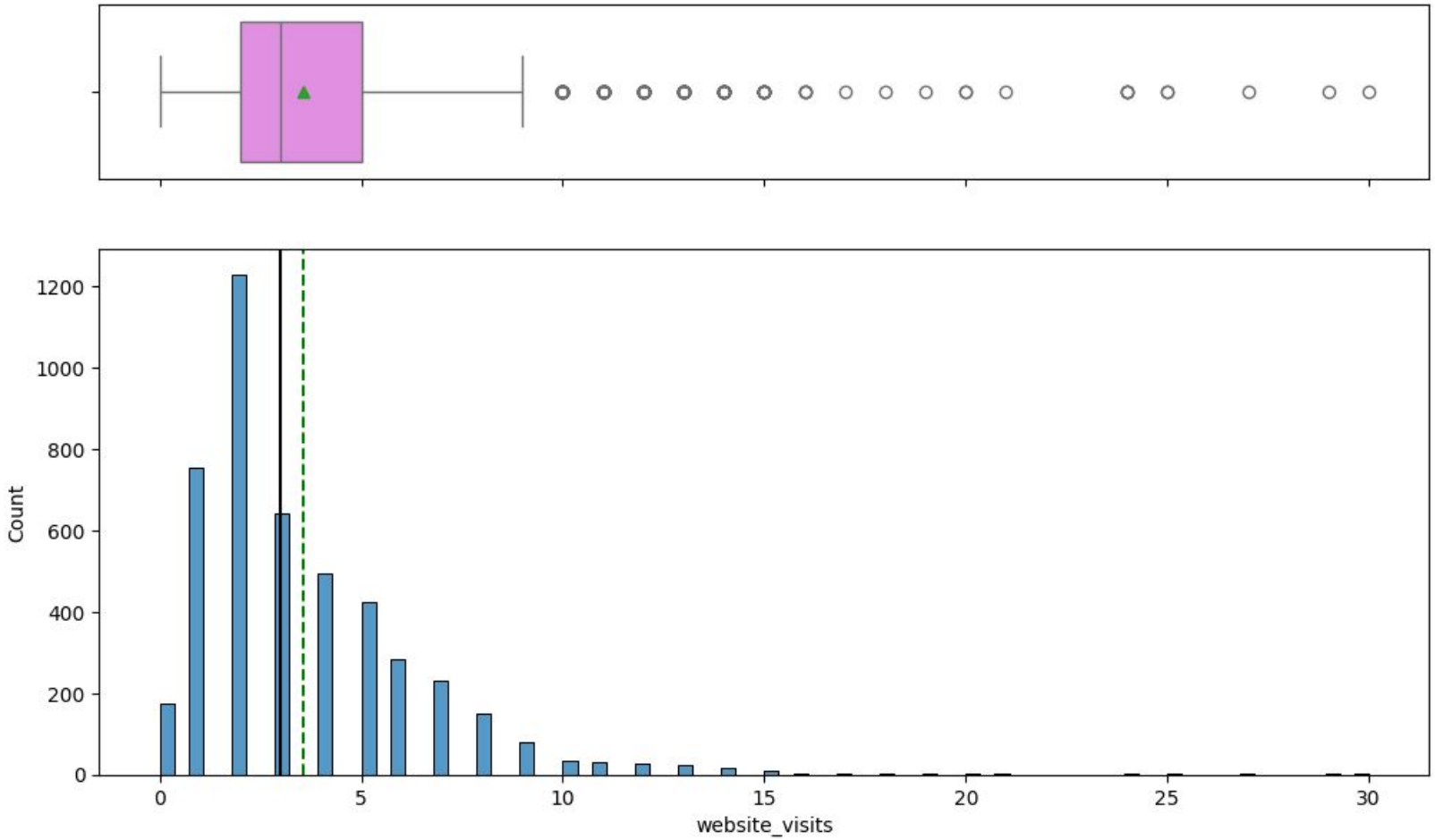
Univariate analysis: Observation on Age



- The distribution is skewed to the left
- There are no outliers as seen in the boxplot
- 50 percent of the leads are below 51 years of age

EDA Results

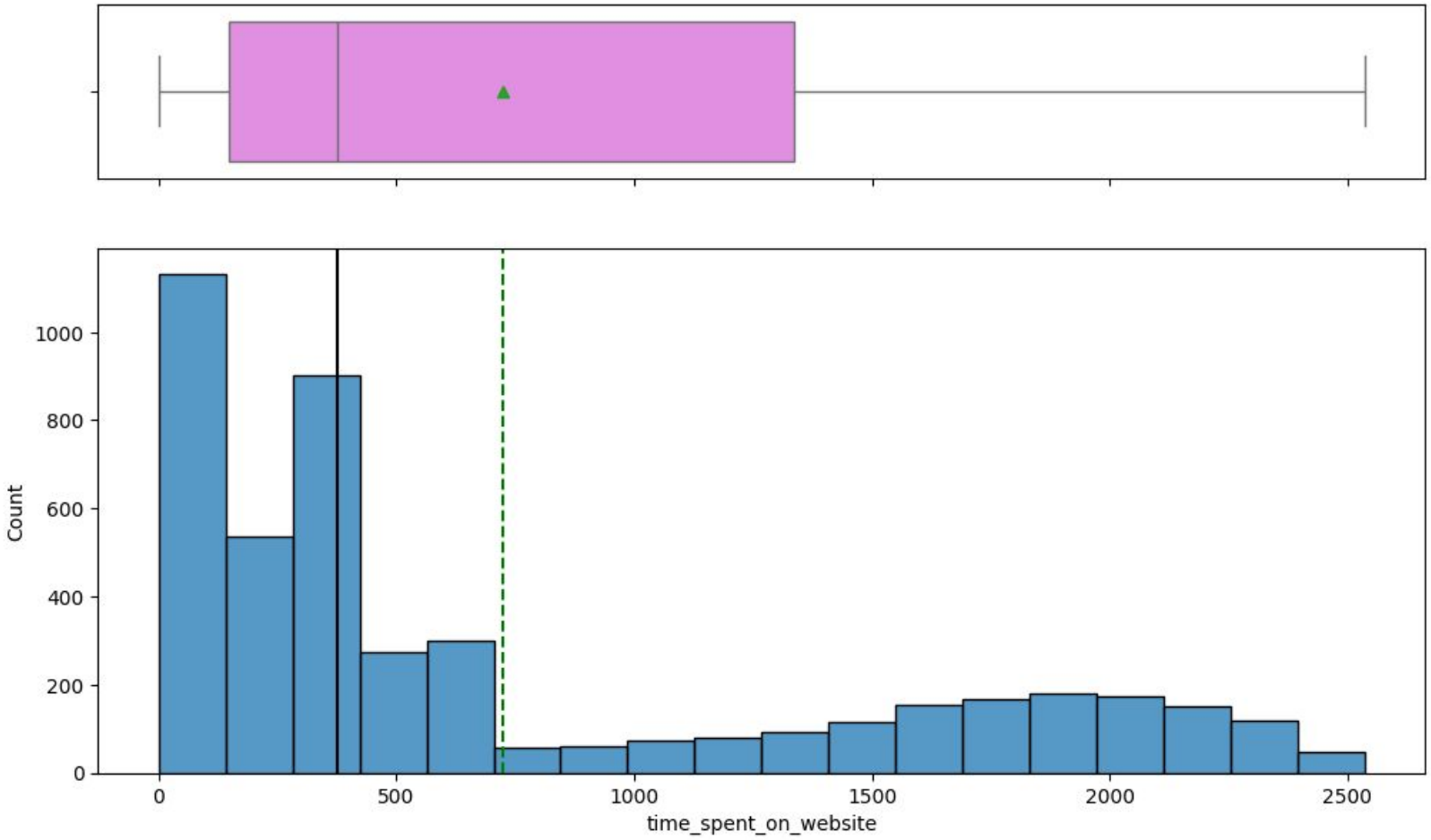
Univariate analysis: Observation on Website visit



- The distribution is skewed to the right
- There are outliers. This is evident from the box plot and the histogram. Some leads seem to visit the website way too often
- Around 1200 leads visited the website twice, about 790 leads visited the website once, and the number of leads seem to decrease as the frequency of website visit increases.
- 174 leads did not visit the website

EDA Results

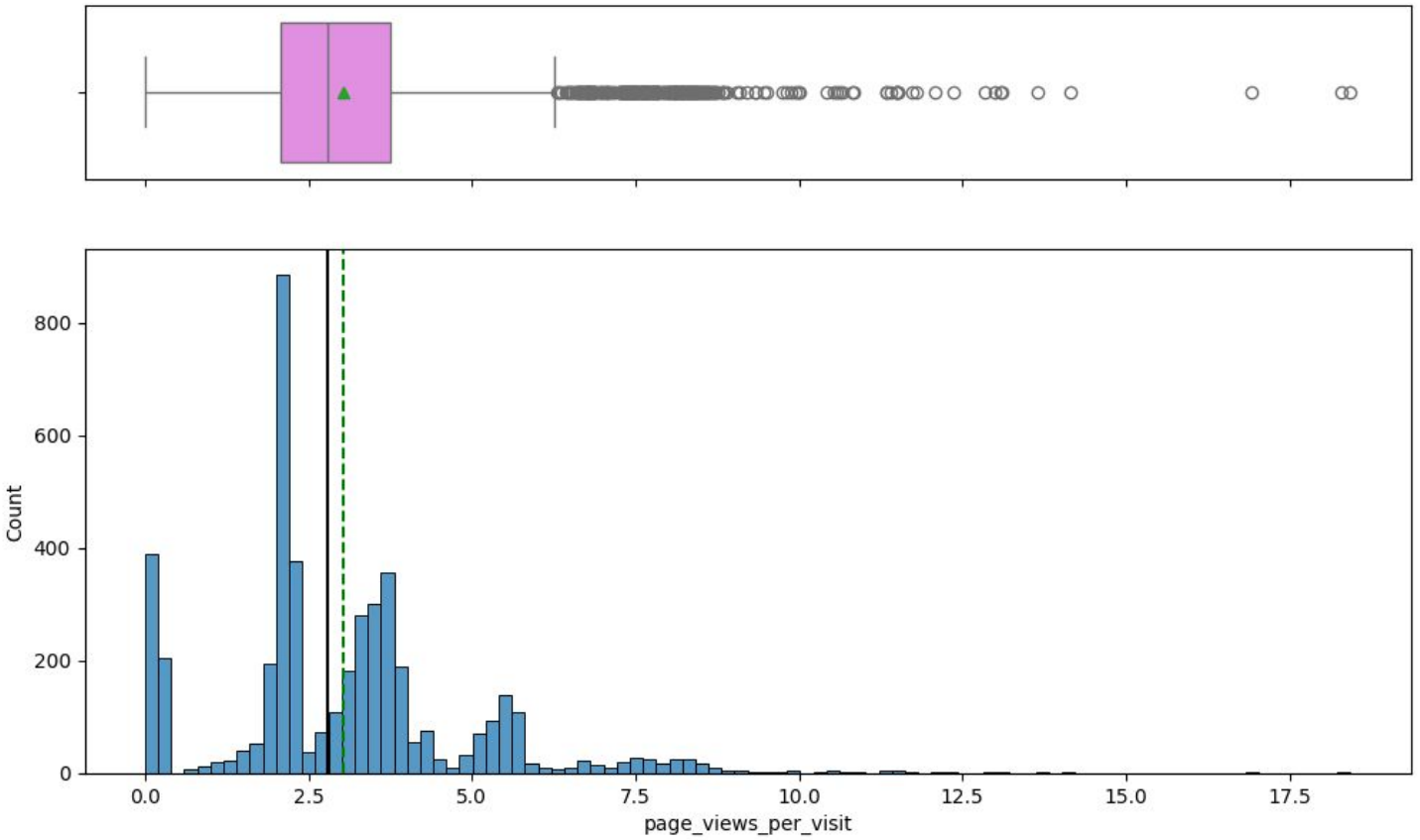
Univariate analysis: Observation on Time spent on website



- Boxplot reveals that there are no outliers.
- The distribution is skewed to the right as shown in the histogram.
- 25 percent of the leads spent between 0 to 120 unit time on the website

EDA Results

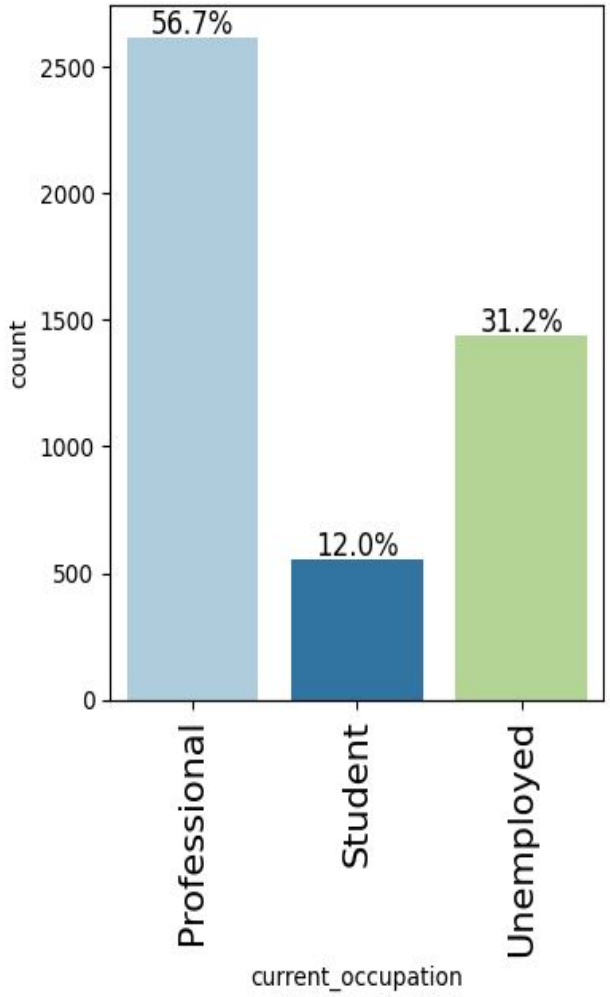
Univariate analysis: Observation on Page views per visit



- Boxplot reveals that there are outliers.
- There are some leads who visited a significant number of pages on the website
- The median number of pages viewed per visit is less than the average number of pages viewed per visit

EDA Results

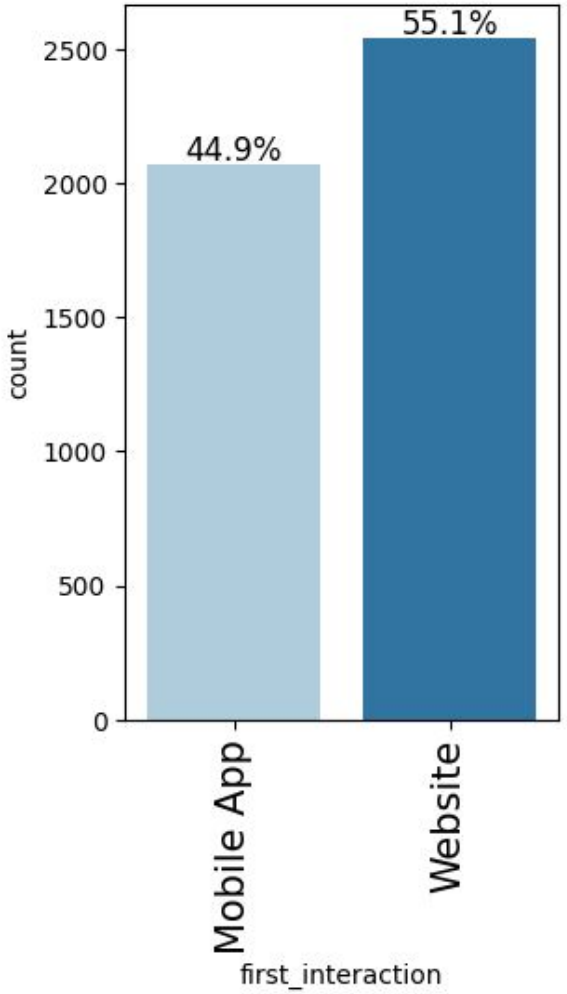
Univariate analysis: Observation on Current occupation



- Here is a barplot of current occupation of leads.
- 57 percent of the leads are professionals, 12 percent are students and 31.2 percent are unemployed.
- These percentages suggests that professionals are constantly trying to upskill in order to reach the utmost height in their career. Likewise, the unemployed people are also curious about online learning, trying to learn skills that can help them land a job. Students are quite busy with their studies and are less likely to show interest in offerings from Edtech companies.

EDA Results

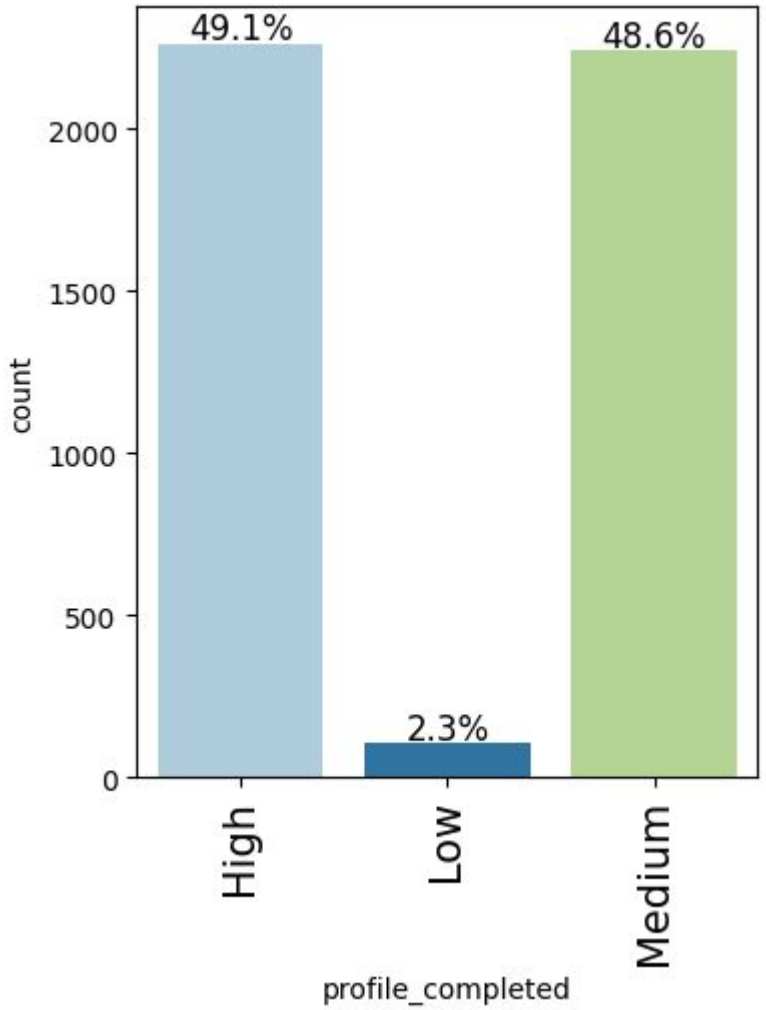
Univariate analysis: Observation on First interaction



- Here is a bar plot showing the device that leads first used to interact with Extraalearn.
- 44.9 percent of the leads first used a Mobile App while 55.1 percent of the leads first used the website to interact with Extraalearn.

EDA Results

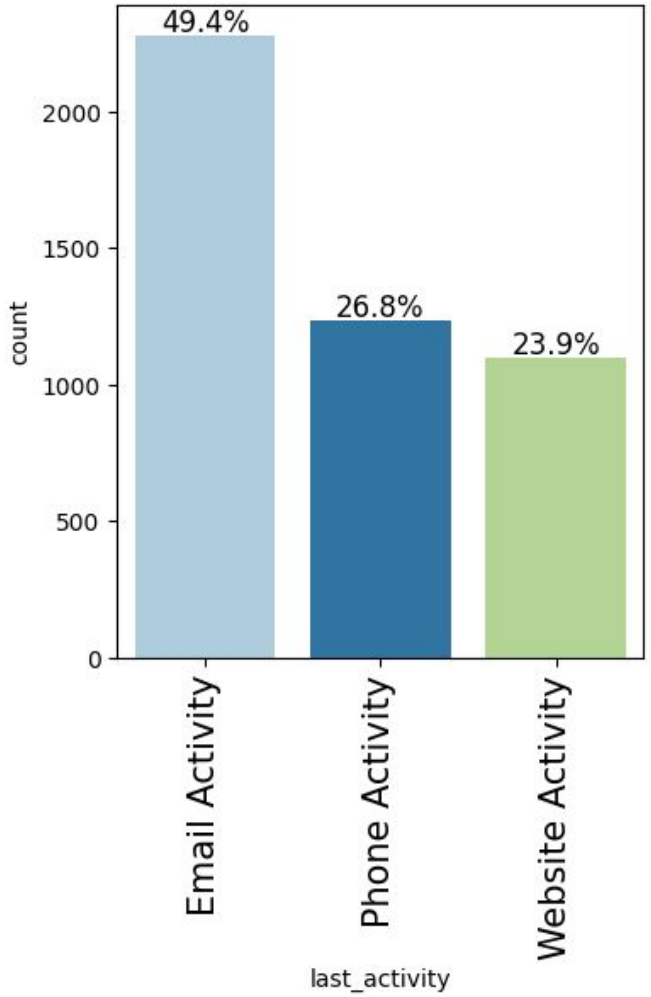
Univariate analysis: Observation on Profile completed



- Here is a bar plot showing how much of the profile has been filled by the leads on the website or mobile app.
- 49.1 percent of the leads filled more than 75percent of their profile, 48.6 percent of the leads filled between 50 to 75 percent of their profile, while 2.3 percent filled less than 50 percent of their profile on Extraalearn's website or mobile app.
  Leads seems to be comfortable providing their profile information to Extraalearn

EDA Results

Univariate analysis: Observation on Last activity

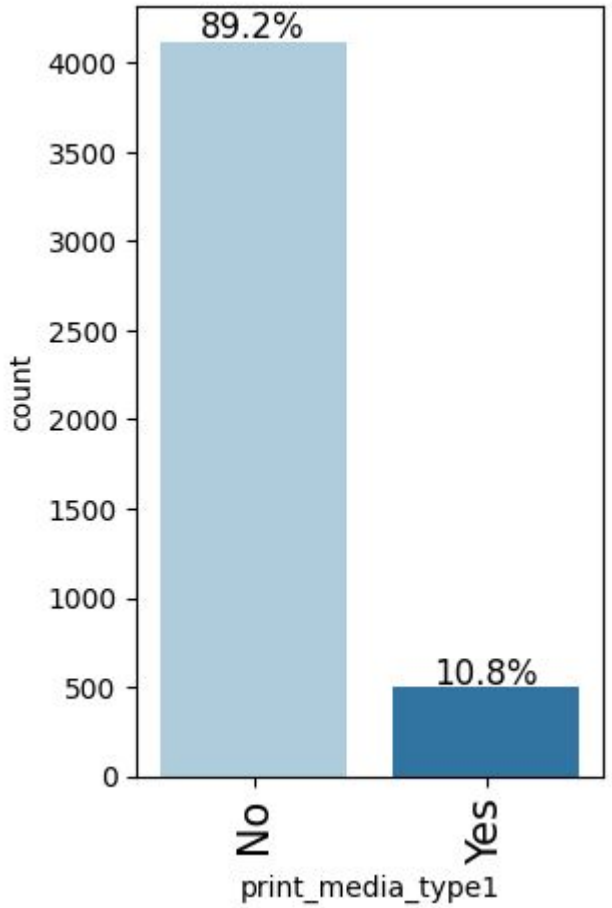


- Here is a bar plot showing the medium of interaction of the leads with Extraalearn.

- 49.4 percent of the leads interacted via Email, 36.8 percent of the leads interacted via Phone, while 23.9 percent interacted via the website.
- Most leads seem to prefer Emails compared to other means of interaction with Extraalearn.

EDA Results

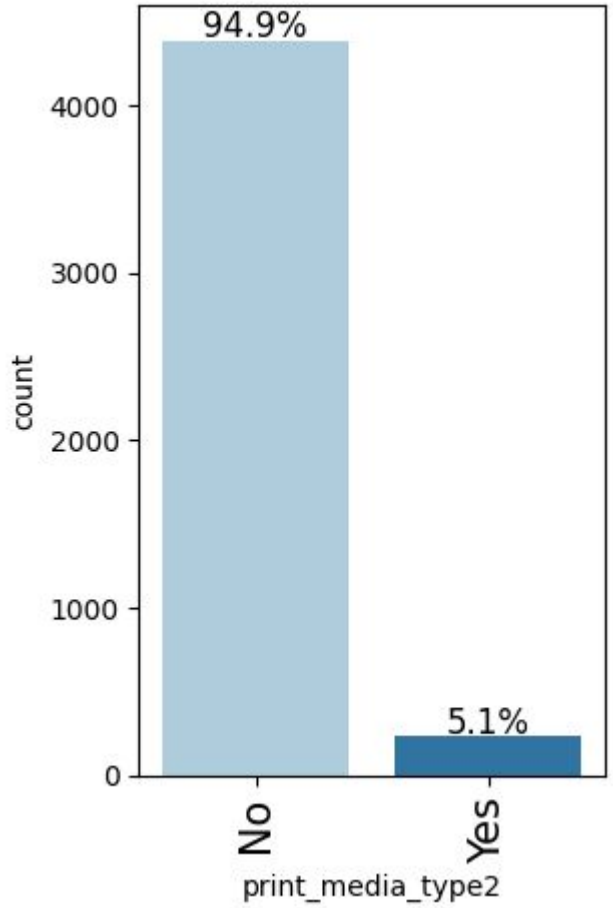
Univariate analysis: Observation on Print media type 1



- Here is a bar plot showing whether the leads had seen the ad of Extraalearn in the Newspaper
- 89.2 percent answered “No“ while 10.8 percent answered “Yes”.
- Most of the leads had not seen Extraalearn’s ad in the Newspaper. This could be because not many people actually read Newspapers these days, so Extraalearn might want to advertised less in Newspapers

EDA Results

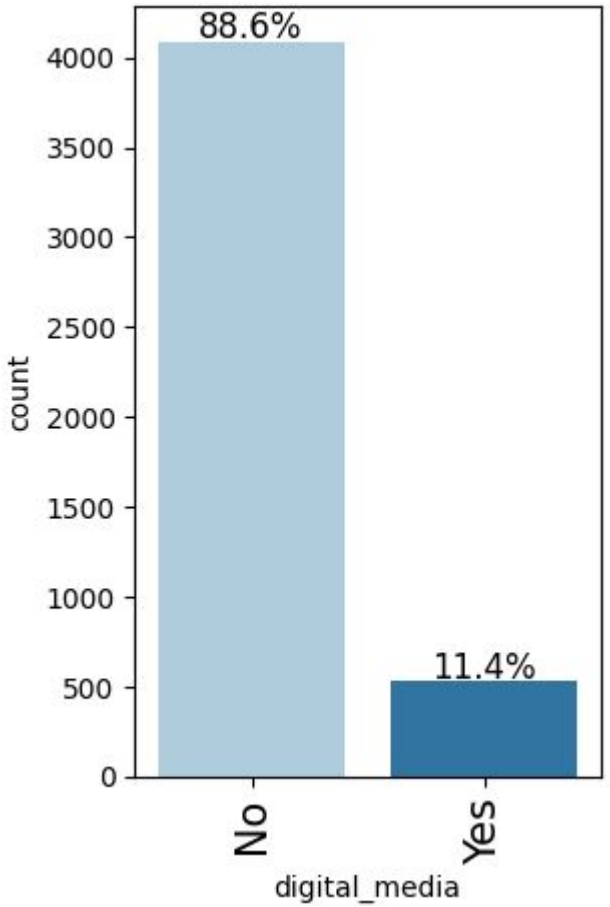
Univariate analysis: Observation on Print media type 2



- Here is a bar plot showing whether the leads had seen the ad of Extraalearn in the Magazine
- 94.9 percent answered “No“ while 5.1 percent answered “Yes”.
- Most of the leads had not seen Extraalearn’s ad in the Magazine. This could be because not many people actually read Magazines these days, so Extraalearn might want to advertised less in Magazines

EDA Results

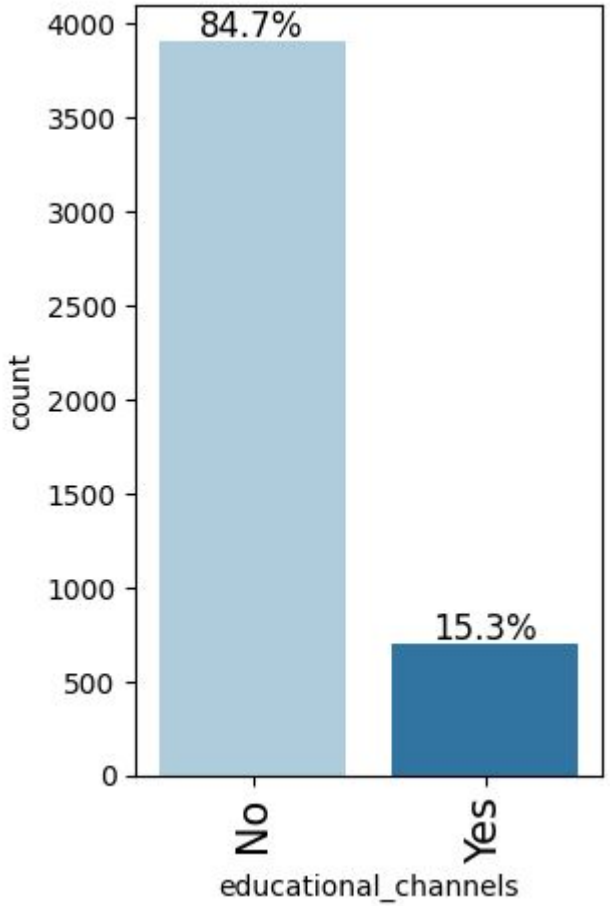
Univariate analysis: Observation on Digital media



- Here is a bar plot showing whether the leads had seen the ad of Extraalearn on digital media
- 88.6 percent answered “No“ while 11.4 percent answered “Yes”.
- Most of the leads had not seen Extraalearn’s ad on digital media.

EDA Results

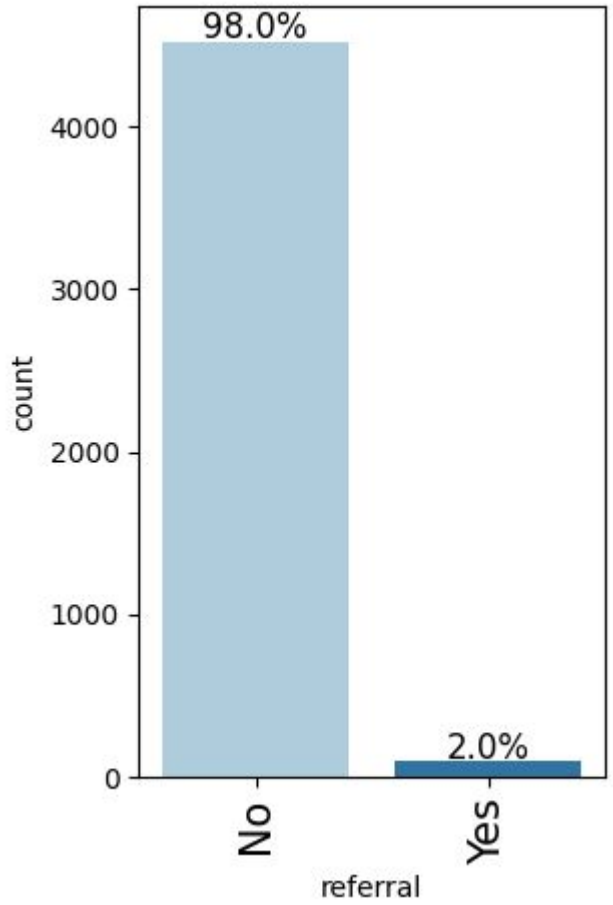
Univariate analysis: Observation on Educational channels



- Here is a bar plot showing whether the leads had heard about ExtraaLearn in the education channels like online forums, discussion threads, educational websites, etc.
- 84.7 percent answered “No“ while 15.3 percent answered “Yes”.
- Most of the leads had not heard about Extraalearn in educational channels.

EDA Results

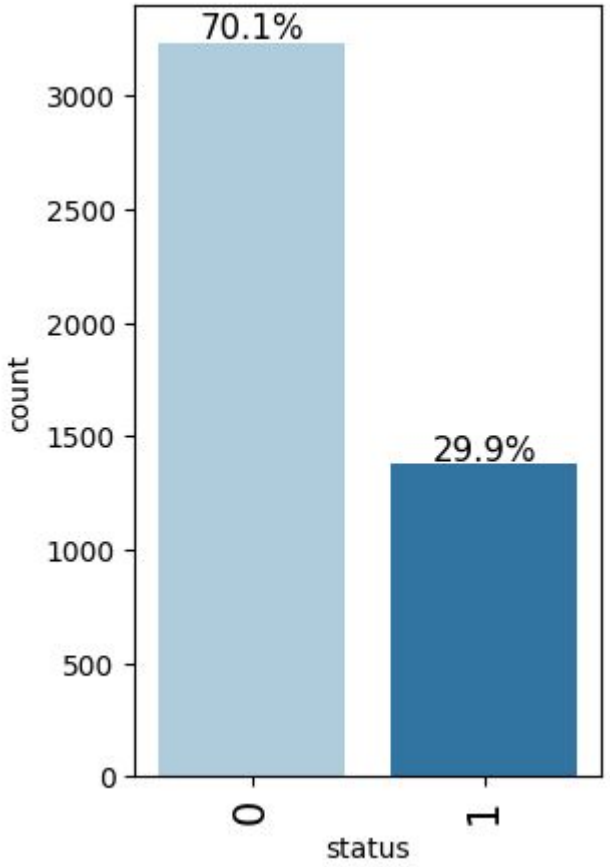
Univariate analysis: Observation on Referral



- Here is a bar plot showing whether the lead had heard about ExtraaLearn through reference.
- 98.0 percent answered “No“ while 2.0 percent answered “Yes”.
- Most of the leads had not heard about Extraalearn through reference

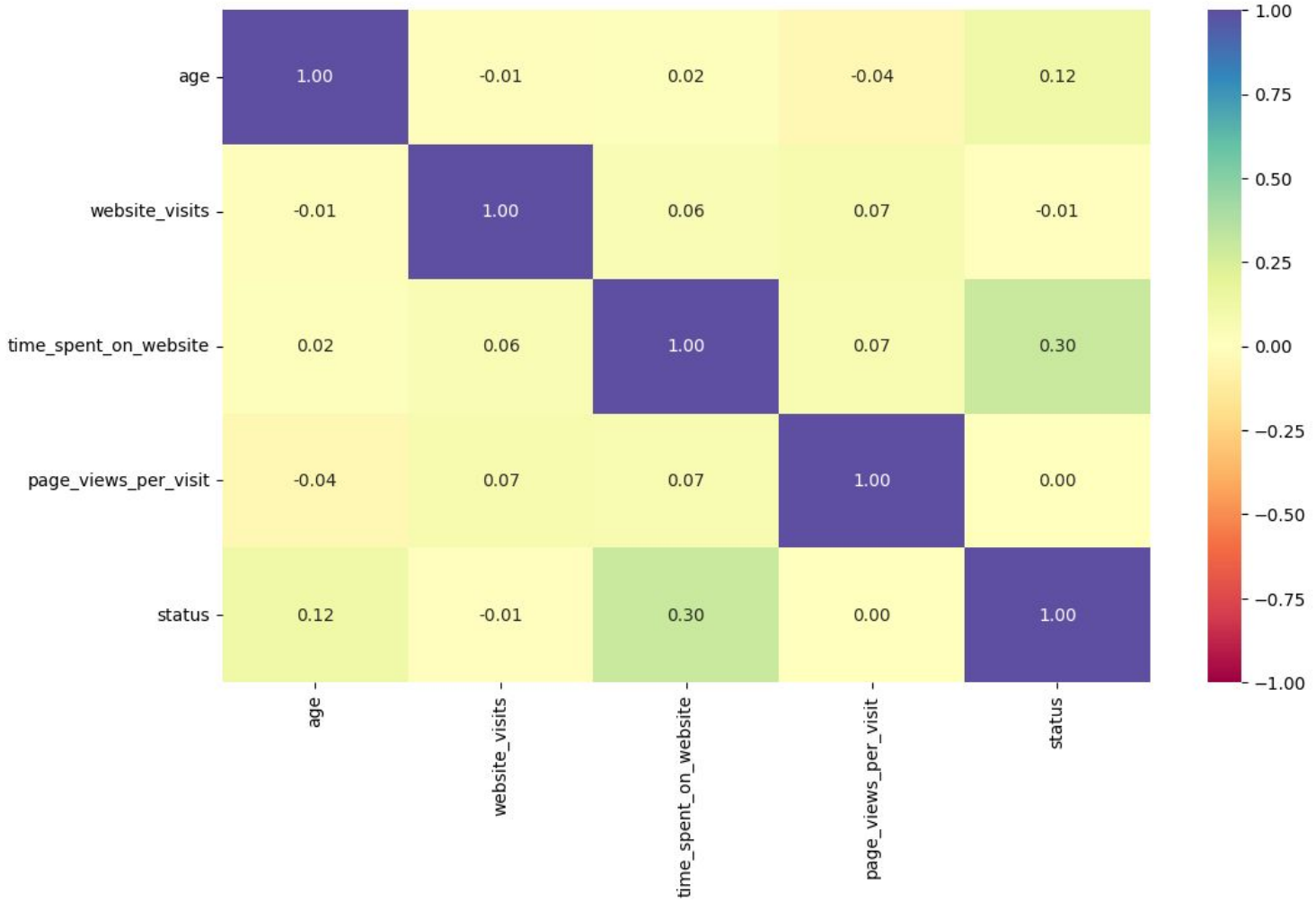
EDA Results

Univariate analysis: Observation on Status



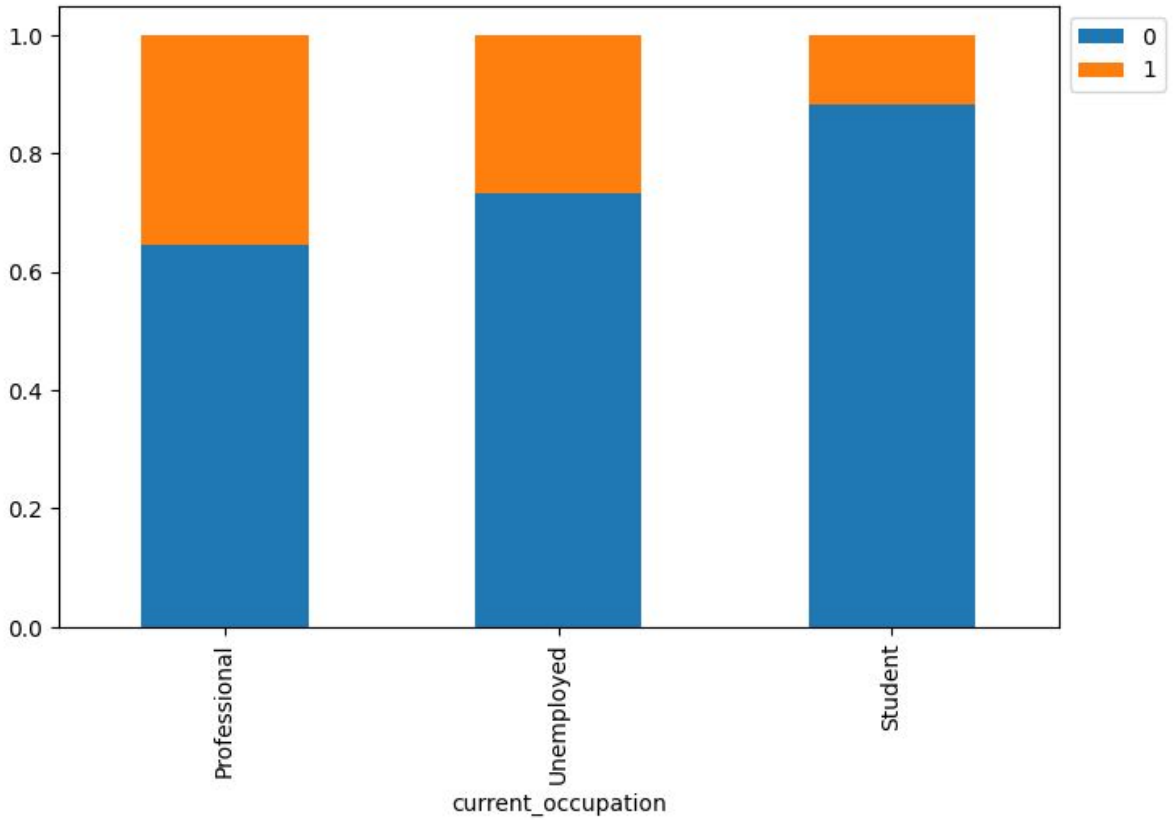
- Here is a bar plot showing whether the lead was converted to a paid customer or not.
- Coding “not converted” as 0 and “converted” as 1, we find that 70.1 percent were converted while 29.9 percent were not converted.

EDA Results
Bivariate analysis



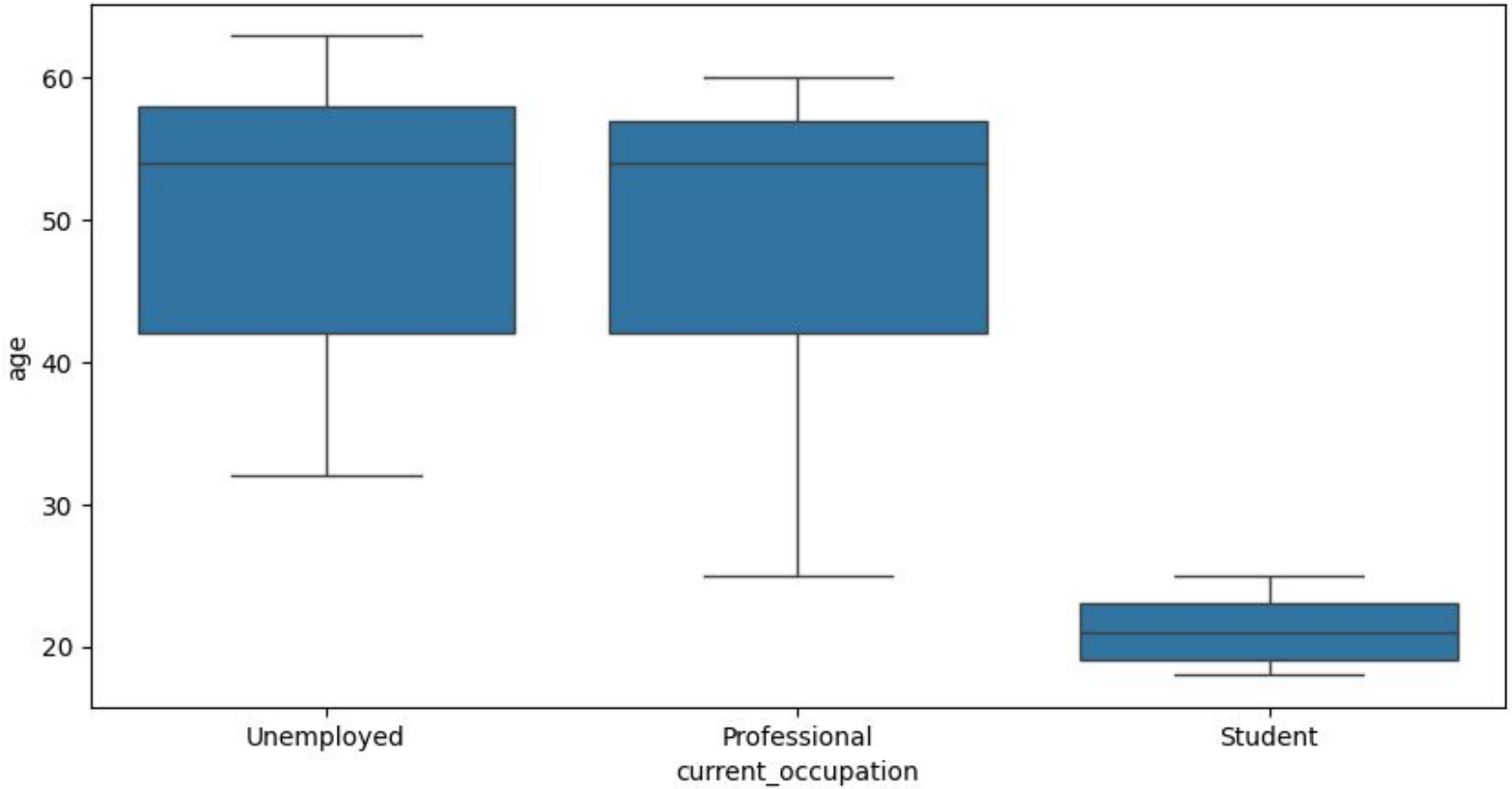
- The variables time_spent_on_website and status have the highest correlation
- There is no strong positive correlation between any of the variables
- The variable website_visits is negatively correlated with status
- Page_views_per_visit has no correlation with status.
- The correlations between the variables are all weak
- We will need further analysis to determine which variable significantly influence the conversion of leads to customers.

EDA Results
Bivariate analysis: Current occupation vs Status



- Leads who are professionals are most likely to convert to paying customers
- Leads who are students are the least likely to convert to a paying customers

EDA Results
Bivariate analysis: Current occupation, Age



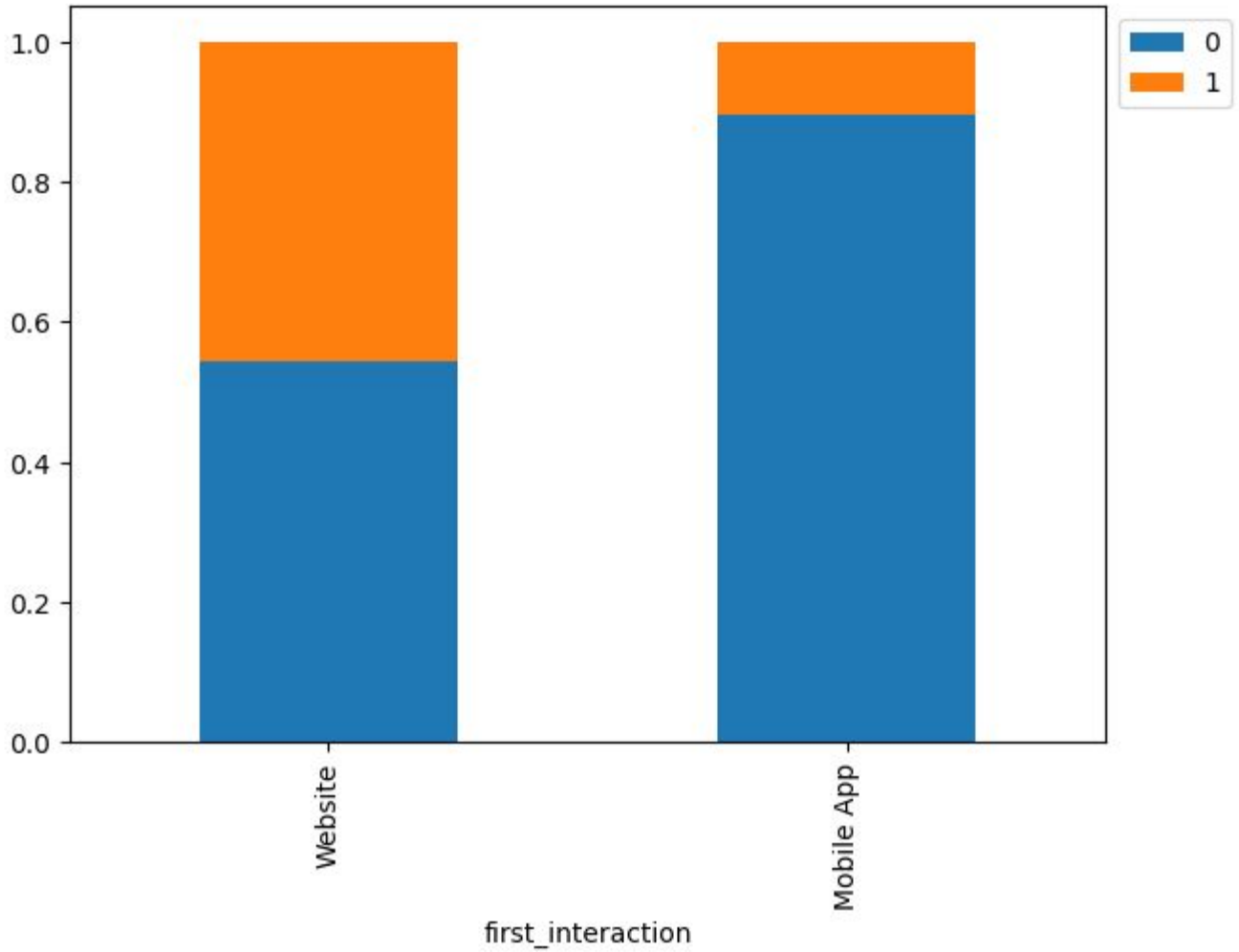
- Leads who above 25 years old are either unemployed or professionals
- Leads who are more than 60 years old are unemployed
- Leads who are less than 25 years old are students
- The youngest professional is around 25 years old

EDA Results
Bivariate analysis: Current occupation, Age

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

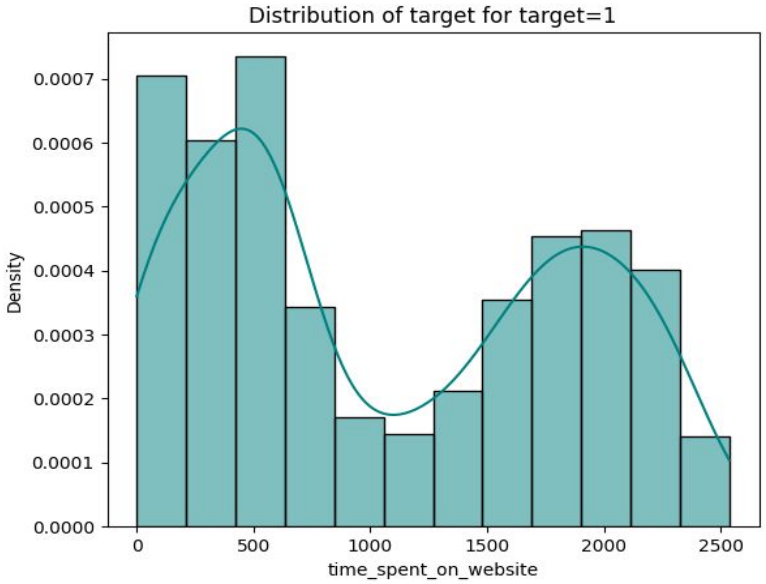
- The oldest lead is 63 years old and is unemployed.
- The youngest lead is 18 years old and is a student
- The average age of professional leads is around 49 while that of unemployed leads is around 50 years

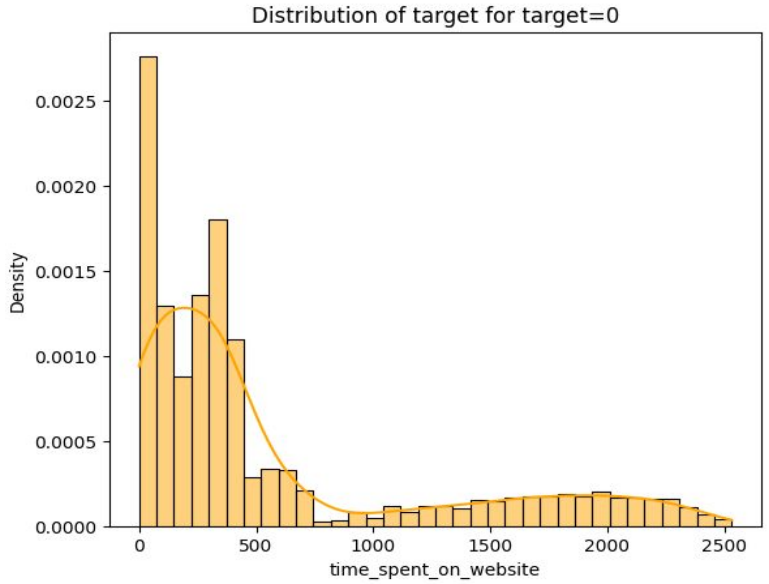
EDA Results
Bivariate analysis: First interaction vs Status

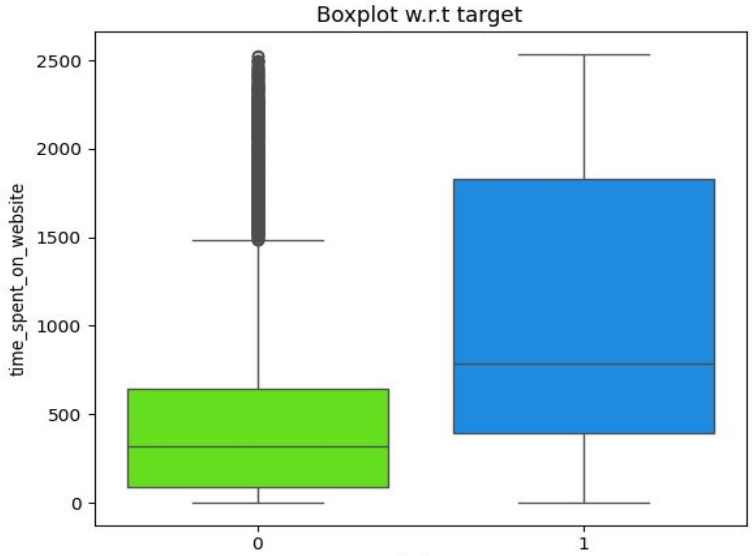


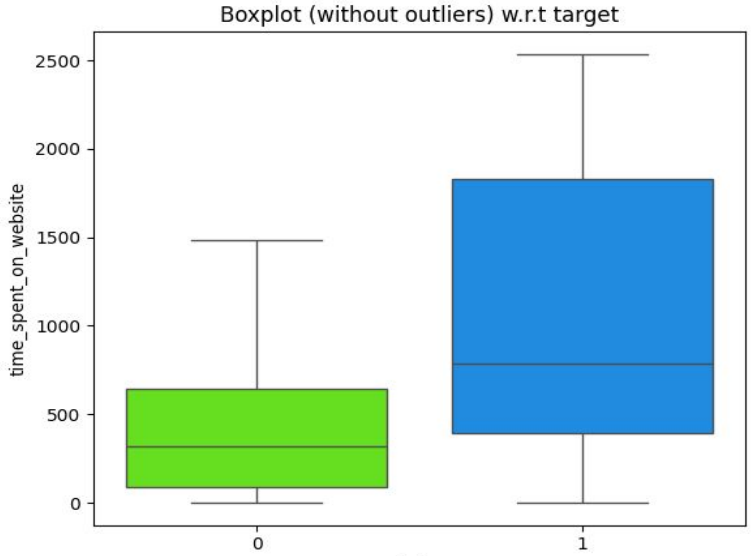
- Leads whose first interaction with Extraalearn was via website are more likely to be converted to customers

EDA Results
Bivariate analysis: Time spent on website vs Status







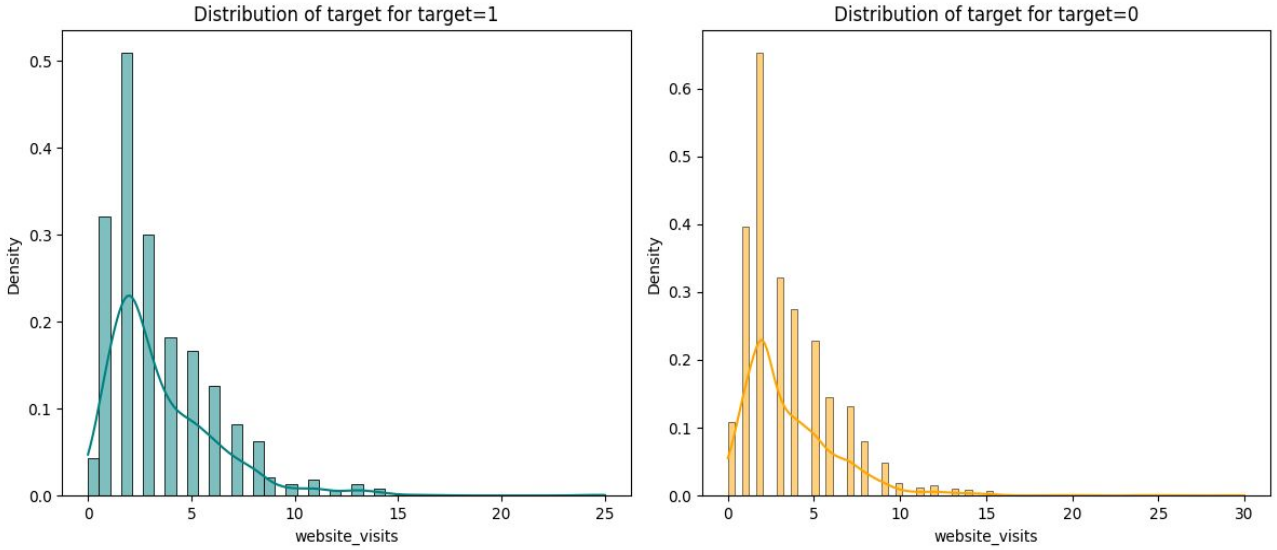


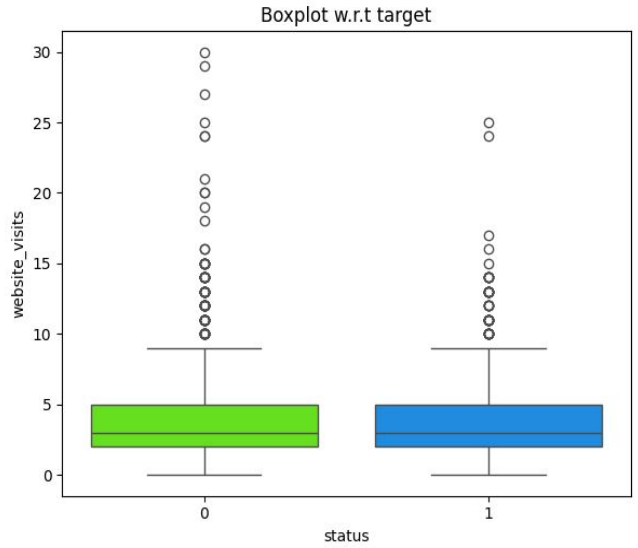
EDA Results
Bivariate analysis: Time spent on website vs Status

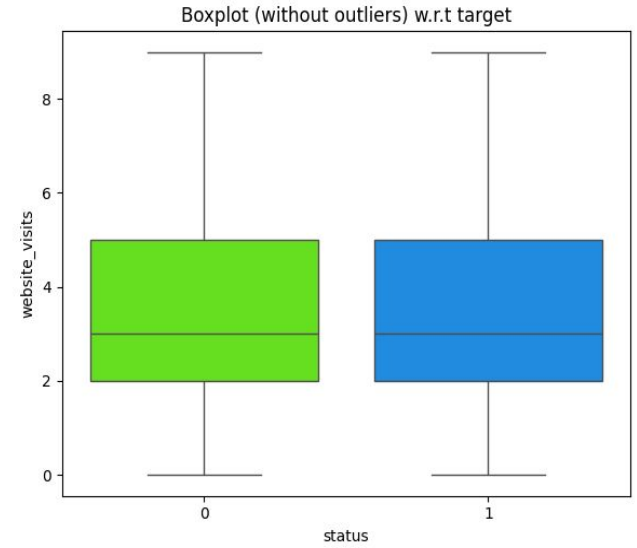
- For leads who were eventually converted to paying customers, the amount of time they spent on the website seem to vary.

- The second plot shows the distribution of time_spent_on_website for leads were not converted to paying customers
- This distribution is rightly skewed, more leads tend to spend fewer time on the website
- There are two Boxes of boxplots: boxplot w.r.t target and boxplot (without outliers) w.r.t target
- For the first Box of boxplot, there is a clear difference between the median. While leads who were not converted are likely to spend less time on the website, there are some leads who spent significantly higher amount of time than others. For leads who were converted , the median time spent is much higher, and the IQR is wider, indicating more variability in time spent on the website among convetrted leads.
- For the second Box of boxplots, the outliers have been removed. We can see that leads who were eventually converted spent more time on average on the website, while those who were unconverted generally spent much less time.
- The median time spent on the website by converted leads is 789 while the median time for unconverted lead is 317

EDA Results
Bivariate analysis: Website visit vs Status





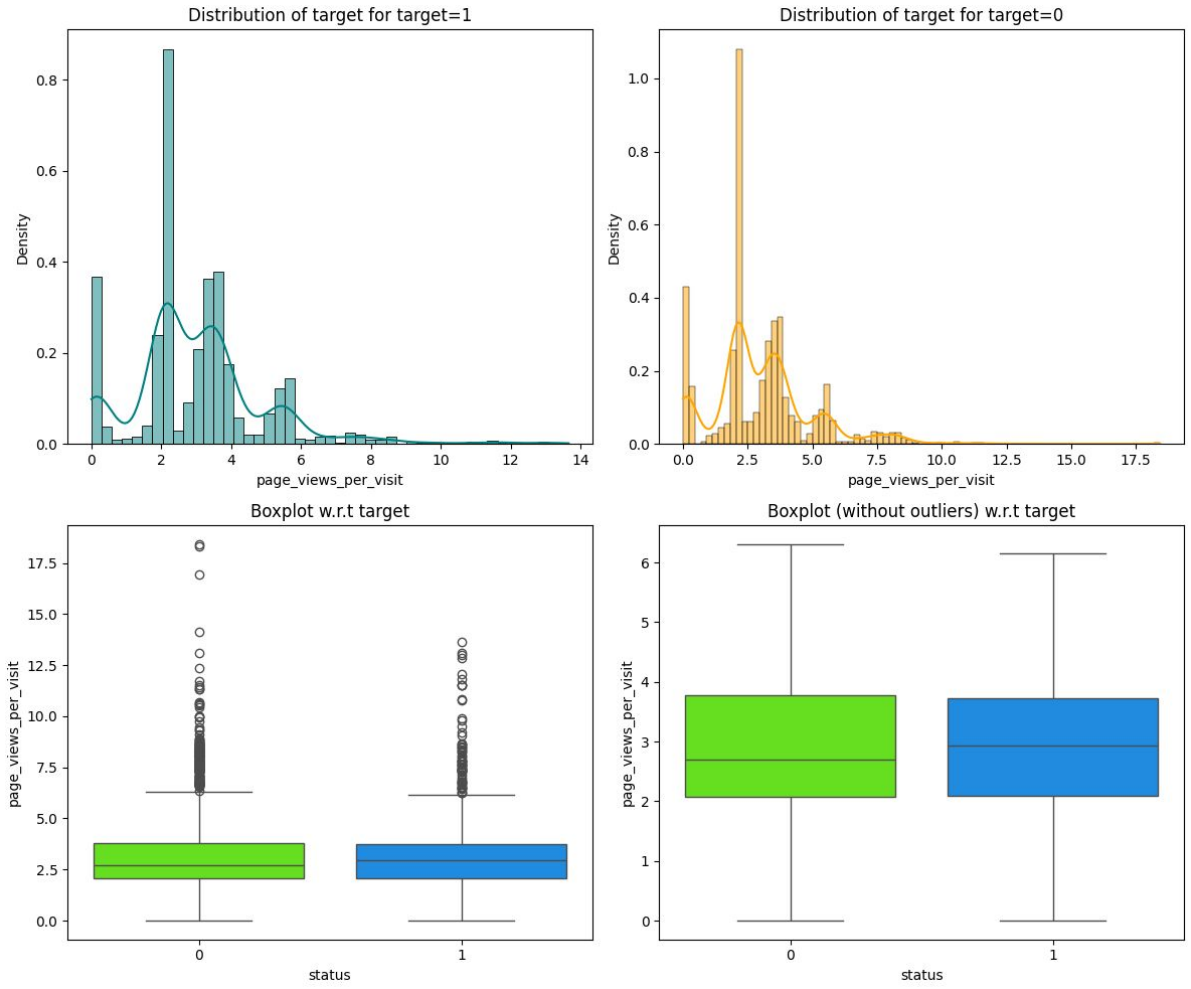


EDA Results
Bivariate analysis: Website visit vs Status

- For leads who were eventually converted to paying customers, the number of time they visit the website seem to vary.

- The second plot shows the distribution of website_visit for leads that were not converted to paying customers
- This distribution is rightly skewed, some leads tend to visit the website frequently
- There are two Boxes of boxplots: boxplot w.r.t target and boxplot (without outliers) w.r.t target
- For the first Box of boxplot, there isn't much difference between the median. There are outliers, there are some unconverted leads who visited the website way too frequently. This is also true for converted leads.
- For the second Box of boxplots, the outliers have been removed.

EDA Results
Bivariate analysis: Page views per visit vs Status

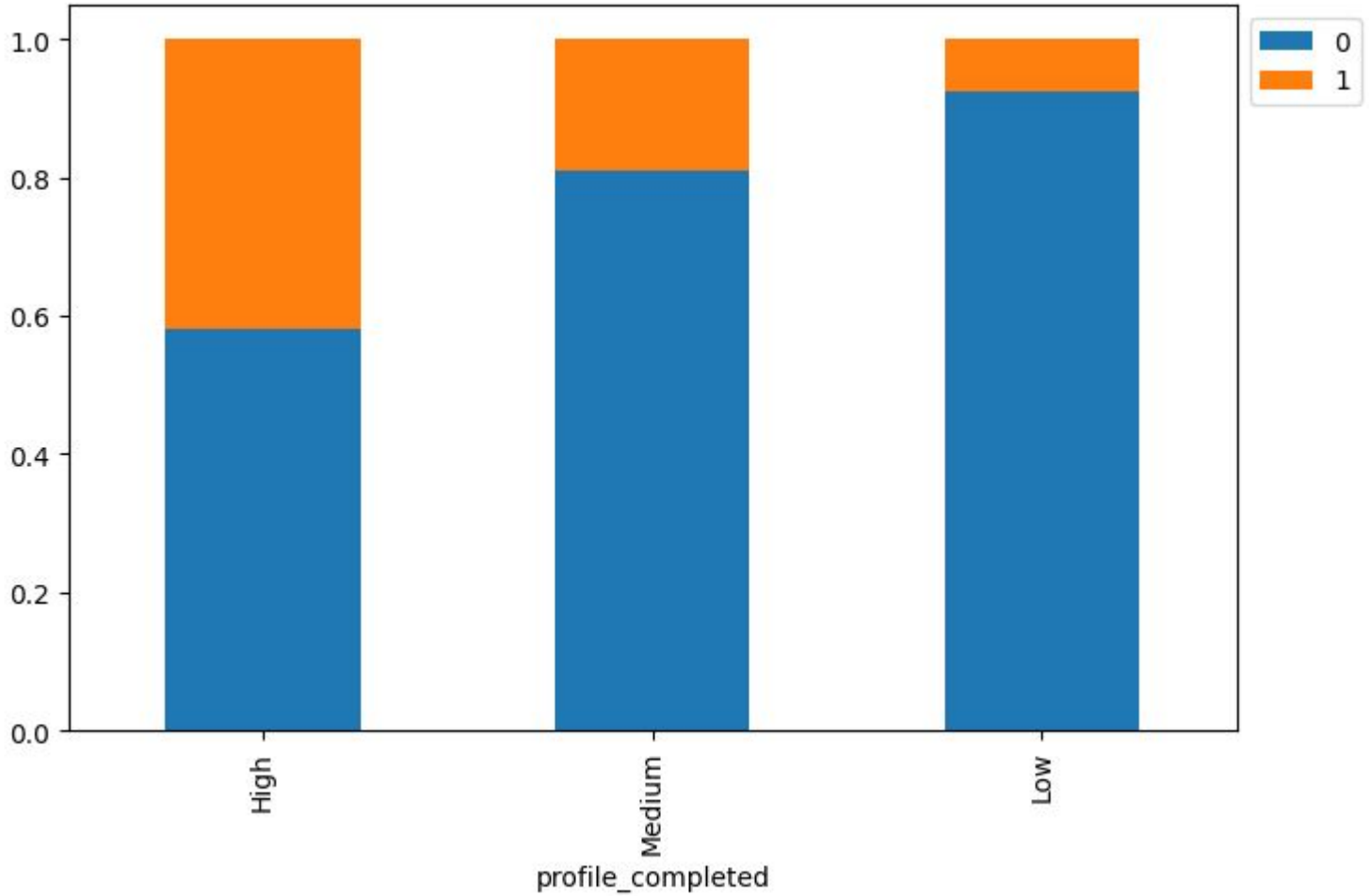


EDA Results
Bivariate analysis: Page views per visit vs Status

- For leads who were eventually converted to paying customers, the number of time they visit the website seem to vary. Most converted leads visited the website twice.

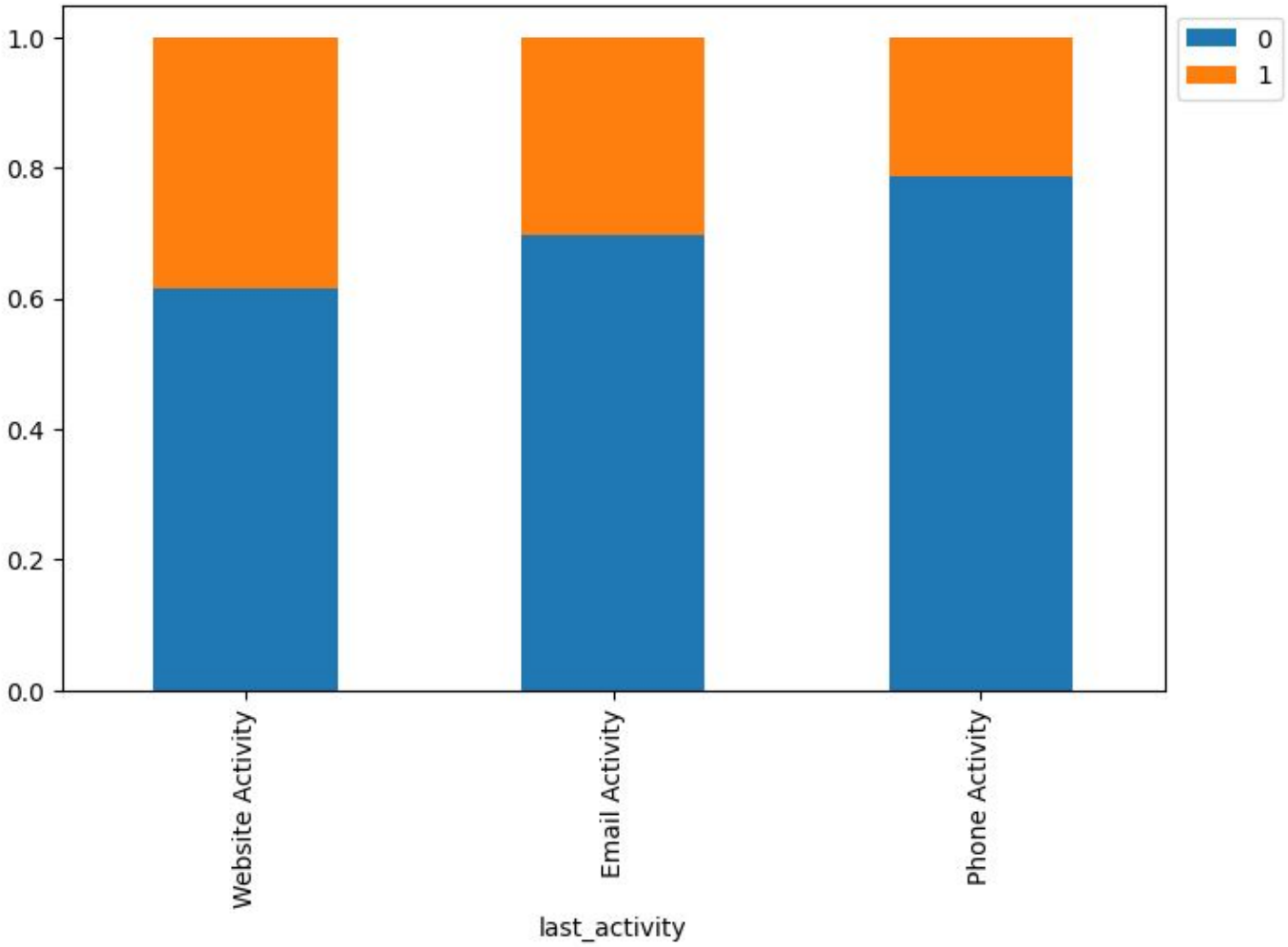
- The second plot shows the distribution of website_visit for leads that were not converted to paying customers
- This distribution is rightly skewed, some leads tend to visit the website frequently yet remained unconverted
- There are two Boxes of boxplots: boxplot w.r.t target and boxplot (without outliers) w.r.t target
- For the first Box of boxplot, there isn't much difference between the median. There are outliers, there are some unconverted leads who visited the website way too frequently. This is also true for converted leads.
- For the second Box of boxplots, the outliers have been removed.

EDA Results
Bivariate analysis: Profile completed vs Status



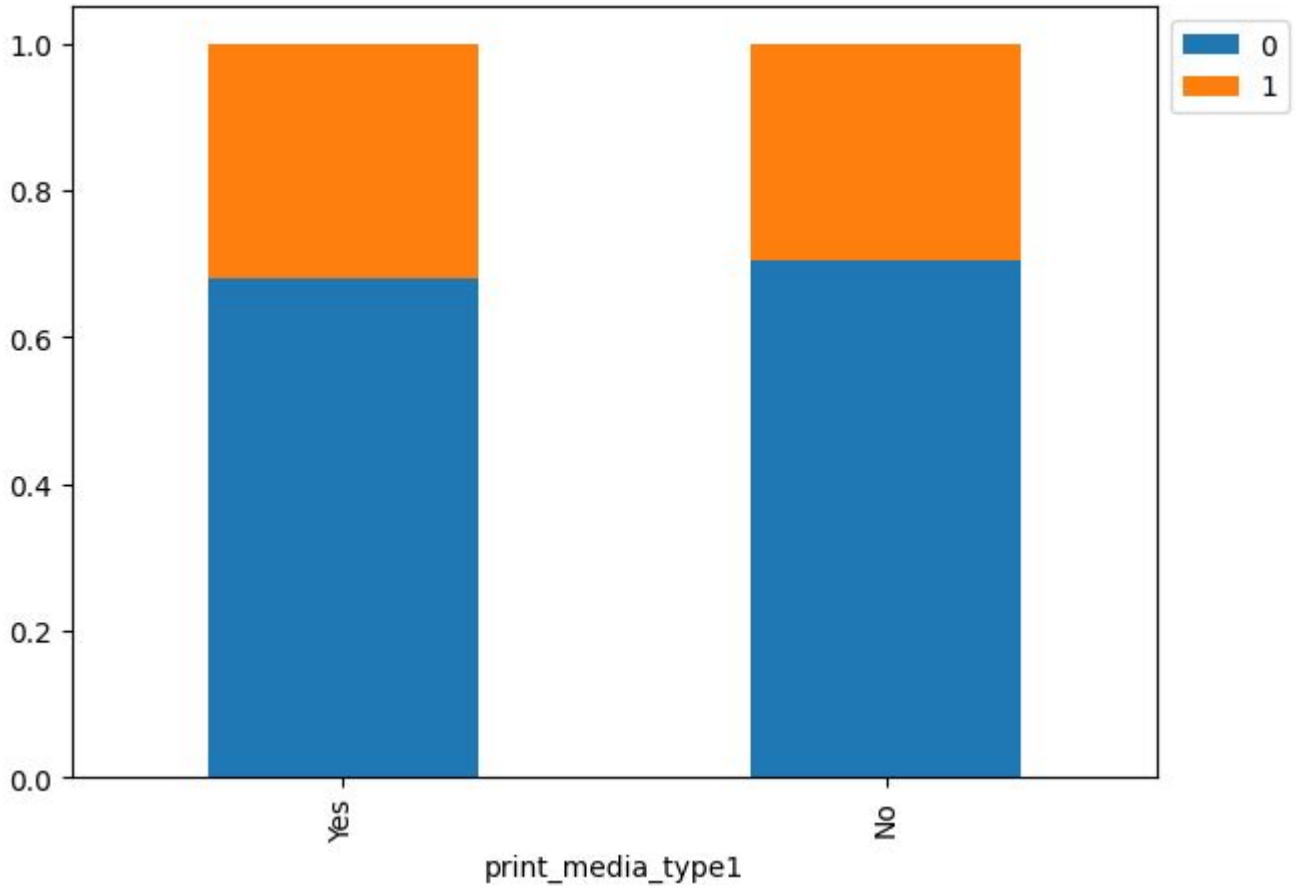
- Leads who filled more than 75 percent of their profile on Extraalearn platform were the most converted
- Leads whose filled less than 50 percent of their profile on Extraalearn platform were the least converted

EDA Results
Bivariate analysis: Last activity vs Status



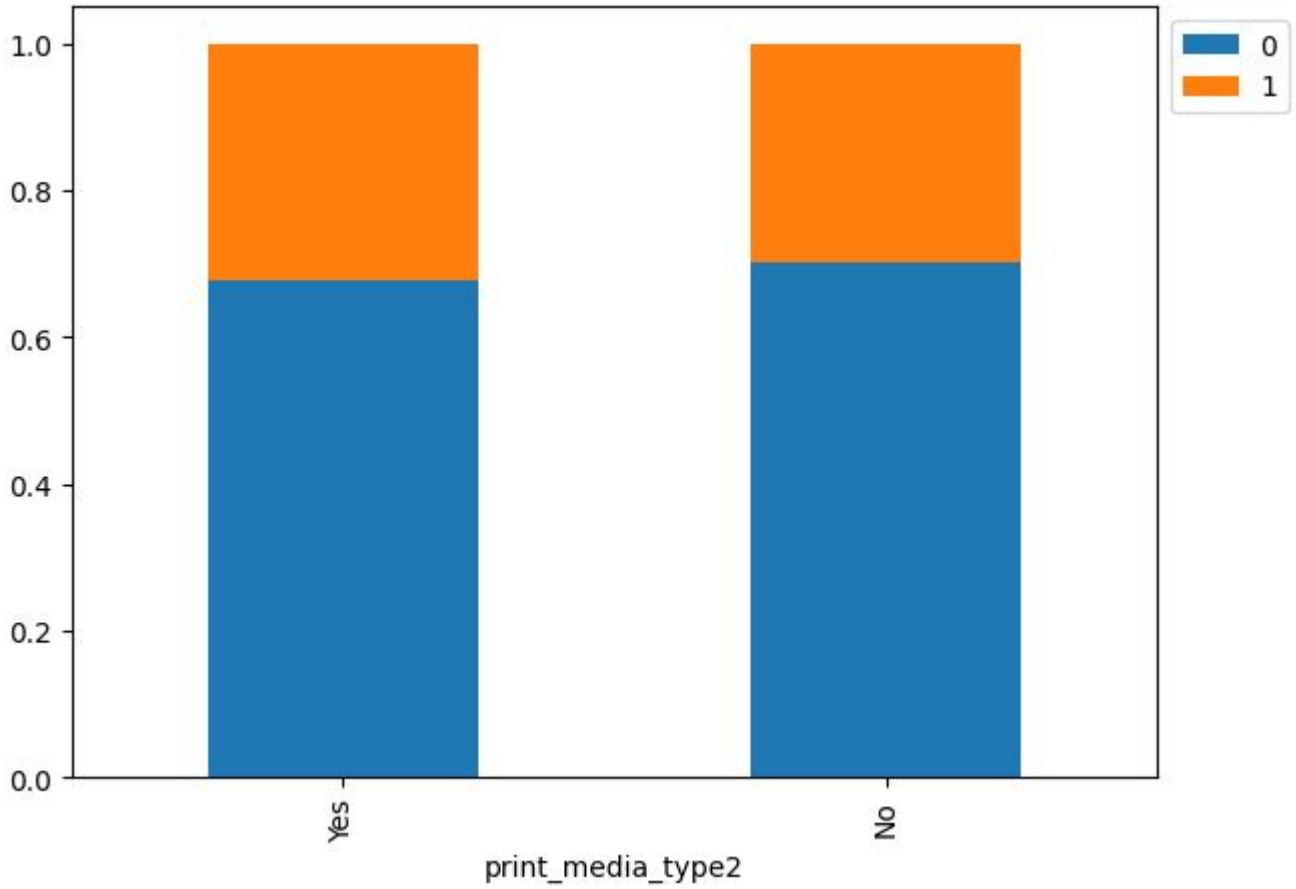
- Leads whose last medium of interaction with Extraalearn was through the Website were most likely to be converted to paid customers, followed by those who used Email.
- Those whose last interaction was via Phone were least likely to be converted to paid customers.

EDA Results
Bivariate analysis: Print media type 1 vs Status



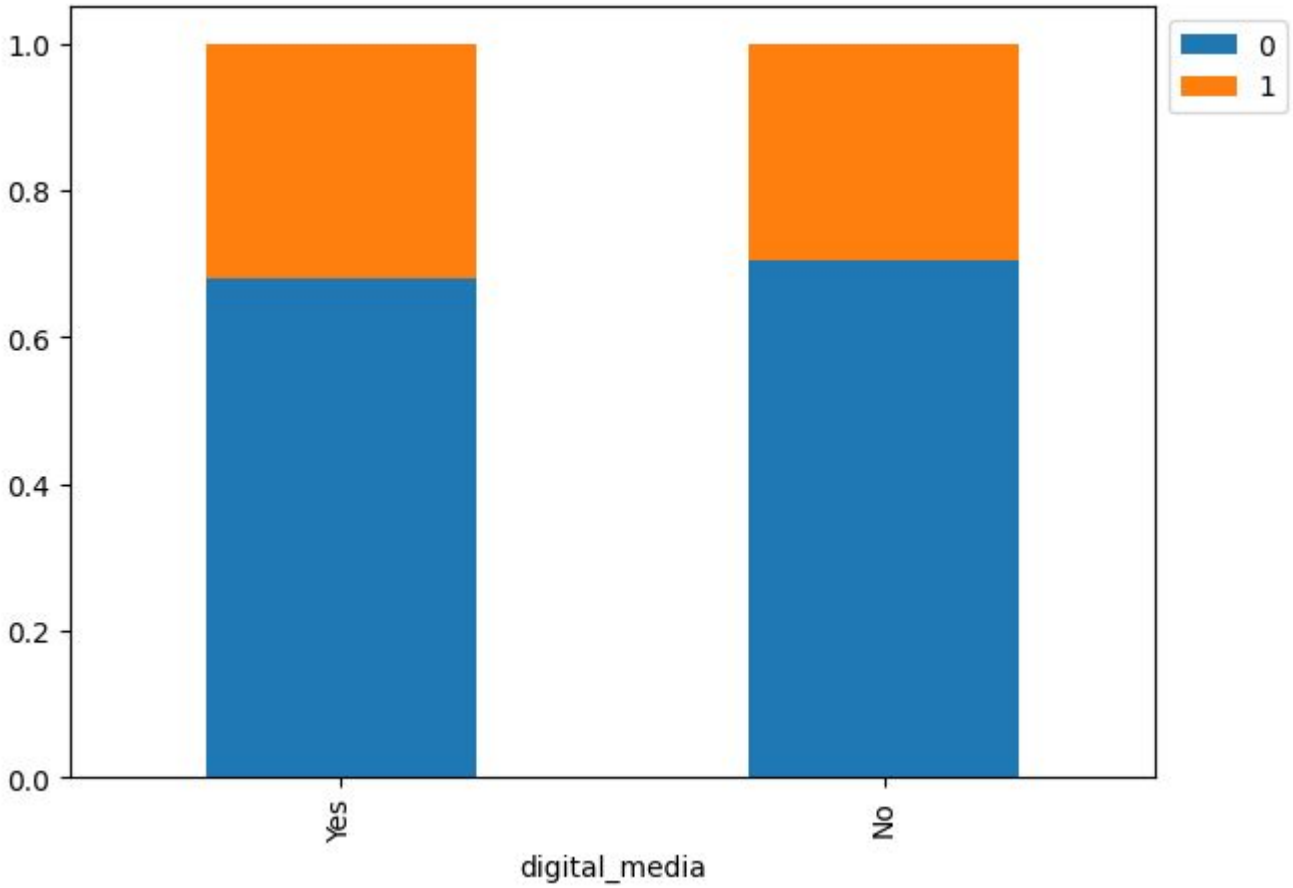
- Among leads who had seen the ad in the Newspaper, there are more unconverted leads than converted ones
  - Among leads who had not seen the ad in the Newspaper, there are more unconverted leads than converted ones
  - The proportion of converted leads who saw the newspaper ad appears to be the same as the proportion of converted leads who did not see it.

EDA Results
Bivariate analysis: Print media type 2 vs Status



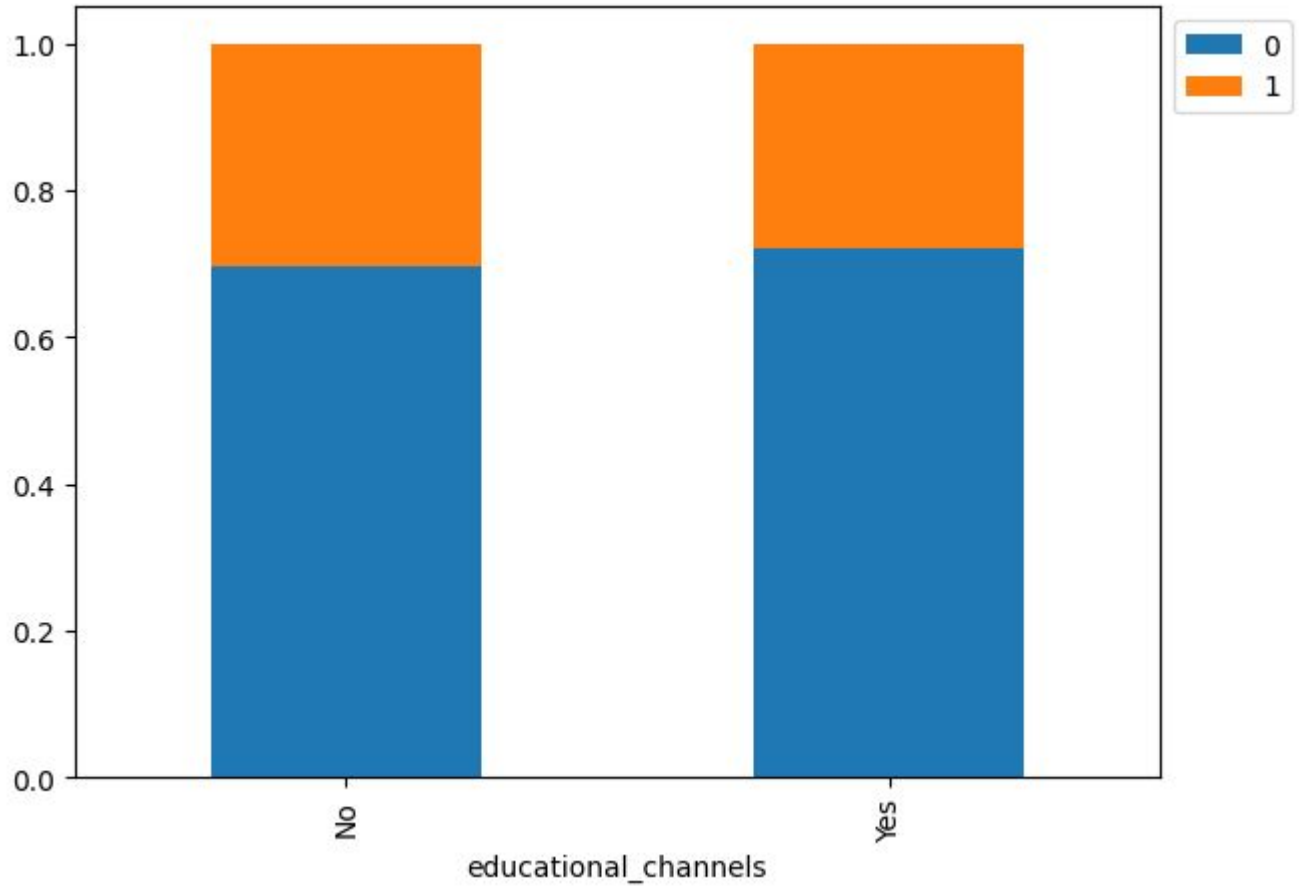
- Among leads who had seen the ad in the Magazine, there are more unconverted leads than converted ones
  - Among leads who had not seen the ad in the Magazine, there are more unconverted leads than converted ones
  - The proportion of converted leads who saw the Magazine ad appears to be the same as the proportion of converted leads who did not see it.

EDA Results
Bivariate analysis: Digital media vs Status



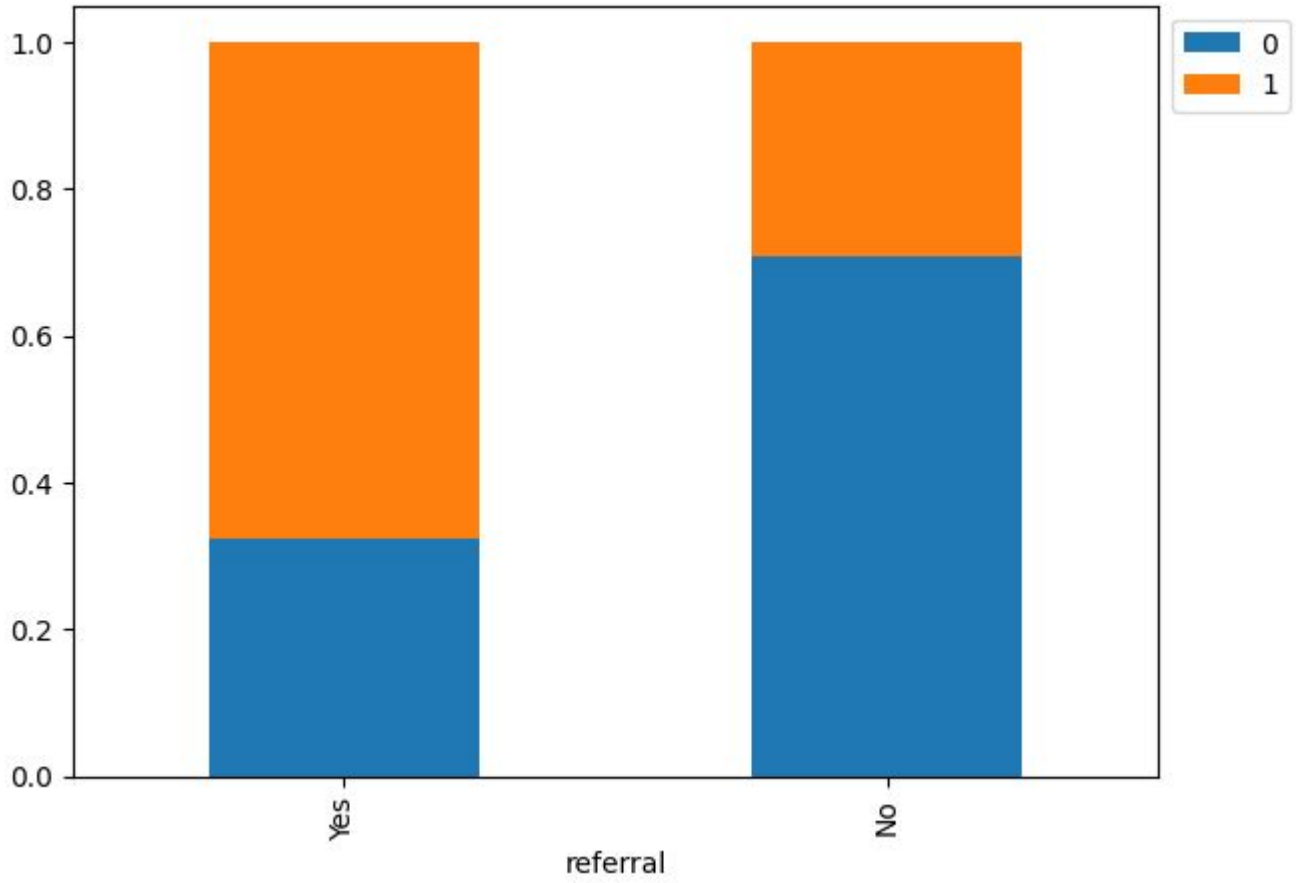
- Among leads who had seen the ad in digital media, there are more unconverted leads than converted ones
  - Among leads who had not seen the ad in digital media, there are more unconverted leads than converted ones
  - The proportion of converted leads who saw the ad in digital media appears to be the same as the proportion of converted leads who did not see it.

EDA Results
Bivariate analysis: Educational channels vs Status



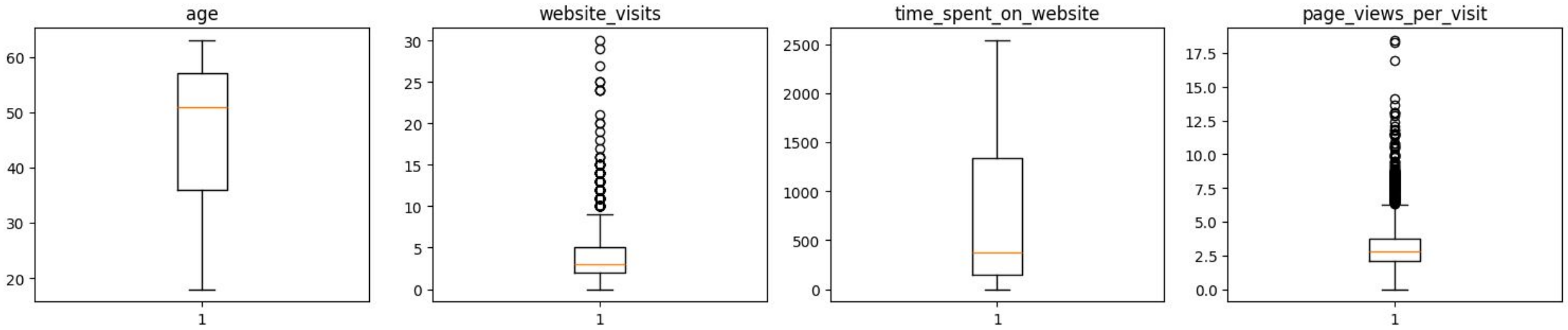
- Among leads who had heard about Extraalearn in education channels like online forums, discussion threads, educational websites, etc., there are more unconverted leads than converted ones
- Among leads who had not heard about Extraalearn in education channels like online forums, discussion threads, educational websites, etc., there are more unconverted leads than converted ones
- The proportion of converted leads who learned about Extraalearn through educational channels seems to be the same as the proportion of converted leads who discovered Extraalearn through other sources.

EDA Results
Bivariate analysis: Referral vs Status



- Leads that were referred to Extraalearn have high chance of being converted to paid customers.
- Leads that were not referred have a higher chance of not being converted to paid customers.
- The chance of converting a referred lead is higher than converting a lead that was not referred.

EDA Results: Outlier check



- There are four boxplots in the image above.
- There are outliers in the second and fourth boxplots. There are some converted leads who visited the website too often, and there some who also viewed way too many pages, per visit, than other leads.

Data Preparation for Modelling

- The training set has 3228 rows and 16 columns
- The test set has 1284 rows and 16 columns
- Percentage of classes in training set: 0.70415 (unconverted leads), 0.29858 (converted leads)
- Percentage of classes in test set: 0.69509 (unconverted leads), 0.30491 (converted leads)

Model Evaluation Criterion

Logit Regression Results

Dep. Variable:	status	No. Observations:	3228
Model:	Logit	Df Residuals:	3211
Method:	MLE	Df Model:	16
Date:	Sat, 23 Nov 2024	Pseudo R-squ.:	0.3589
Time:	11:30:54	Log-Likelihood:	-1256.8
converged:	True	LL-Null:	-1960.4
Covariance Type:	nonrobust	LLR p-value:	4.576e-290

- A pseudo R-square value of 0.3589 suggests the model explains about 35.89% of the variance in the outcome variable, which is considered moderately strong.

Model Evaluation Criterion

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.3415	0.309	-7.580	0.000	-2.947	-1.736
age	-0.0023	0.005	-0.443	0.657	-0.012	0.008
website_visits	-0.0031	0.019	-0.169	0.865	-0.039	0.033
time spent on website	0.0012	7.12e-05	17.558	0.000	0.001	0.001
page_views_per_visit	-0.0342	0.026	-1.301	0.193	-0.086	0.017
current occupation Student	-2.1984	0.254	-8.642	0.000	-2.697	-1.700
current_occupation_Unemployed	-0.5794	0.110	-5.244	0.000	-0.796	-0.363
first_interaction_Website	2.7226	0.127	21.498	0.000	2.474	2.971
profile completed Low	-2.4844	0.467	-5.323	0.000	-3.399	-1.570
profile_completed_Medium	-1.7699	0.109	-16.178	0.000	-1.984	-1.555
last activity Phone Activity	-0.6989	0.126	-5.562	0.000	-0.945	-0.453
last_activity_Website Activity	0.4660	0.120	3.880	0.000	0.231	0.701
print_media_type1_Yes	0.2258	0.154	1.469	0.142	-0.075	0.527
print media type2 Yes	0.4621	0.220	2.097	0.036	0.030	0.894
digital_media_Yes	0.1142	0.155	0.737	0.461	-0.190	0.418
educational channels Yes	0.1286	0.144	0.892	0.372	-0.154	0.411
referral_Yes	1.5923	0.379	4.198	0.000	0.849	2.336

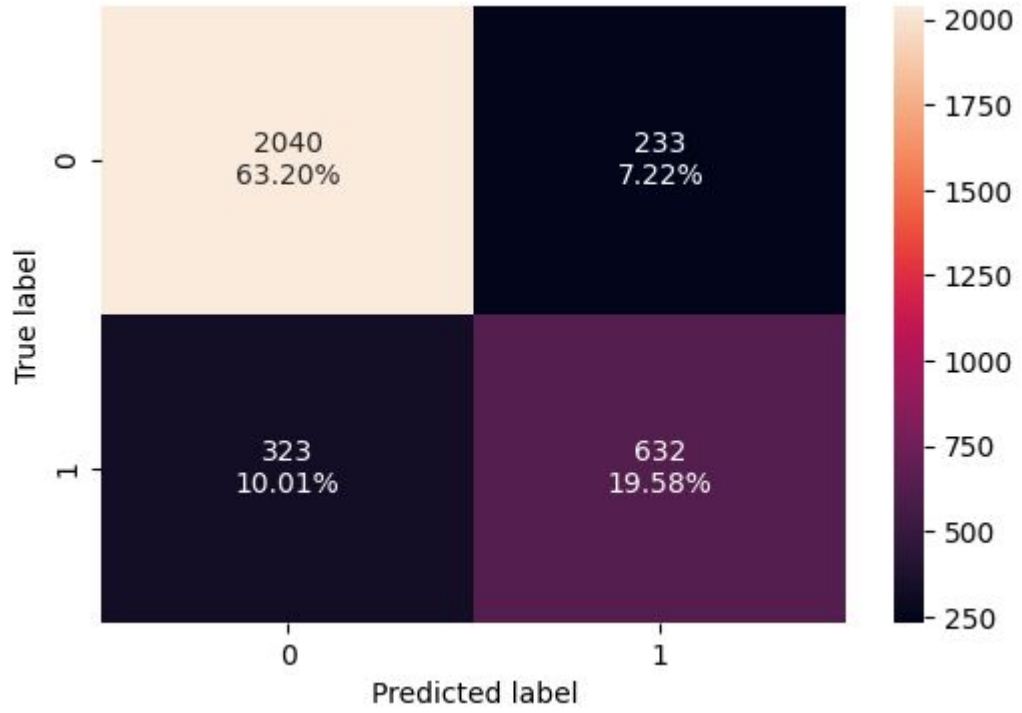
Model Evaluation Criterion: Observations

- Negative values of the coefficient show that the probability of a lead being converted decreases with the increase of the corresponding attribute value. For example the chance of leads being converted decreases as their age increase.
- Positive values of the coefficient show that the probability of a lead being converted increases with the increase of the corresponding attribute value.
- p-value of a variable indicates if the variable is significant or not. If we consider the significance level to be 0.05 (5%), then any variable with a p-value less than 0.05 would be considered significant. For instance, time_spent_on_website would be considered significant

Model Performance

Training performance:

	Accuracy	Recall	Precision	F1
0	0.82776	0.66178	0.73064	0.69451



- The F1_score of the model is ~0.695 and we will try to maximize it further
- The variables used to build the model might contain multicollinearity, which will affect the p-values
- We will have to remove multicollinearity from the data to get reliable coefficients and p-values

Multicollinearity

	feature	VIF
0	const	38.00376
1	age	1.97547
2	website_visits	1.01606
3	time_spent_on_website	1.02318
4	page_views_per_visit	1.01711
5	current_occupation_Student	2.02624
6	current_occupation_Unemployed	1.07083
7	first_interaction_Website	1.00842
8	profile_completed_Low	1.03701
9	profile_completed_Medium	1.02992
10	last_activity_Phone Activity	1.13621
11	last_activity_Website Activity	1.14076
12	print_media_type1_Yes	1.00208
13	print_media_type2_Yes	1.00265
14	digital_media_Yes	1.00396
15	educational_channels_Yes	1.00517
16	referral_Yes	1.00847

- None of the variables show moderate or high multicollinearity since VIF is less than 5 for each variable.
- We still want to know which variables do not affect the target variable, so we will consider dropping high p-valu.e

Dropping Variables with high p-value

Logit Regression Results

Dep. Variable:	status	No. Observations:	3228
Model:	Logit	Df Residuals:	3217
Method:	MLE	Df Model:	10
Date:	Mon, 02 Dec 2024	Pseudo R-squ.:	0.3576
Time:	05:17:31	Log-Likelihood:	-1259.3
converged:	True	LL-Null:	-1960.4
Covariance Type:	nonrobust	LLR p-value:	3.443e-295

- After dropping variables with high p-values, we get the regression result above

Dropping Variables with high p-value

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.4977	0.141	-17.732	0.000	-2.774	-2.222
time spent on website	0.0012	7.06e-05	17.546	0.000	0.001	0.001
current occupation Student	-2.1419	0.204	-10.489	0.000	-2.542	-1.742
current occupation Unemployed	-0.5878	0.110	-5.336	0.000	-0.804	-0.372
first interaction Website	2.7178	0.126	21.518	0.000	2.470	2.965
profile completed Low	-2.4736	0.466	-5.306	0.000	-3.387	-1.560
profile completed Medium	-1.7612	0.109	-16.166	0.000	-1.975	-1.548
last activity Phone Activity	-0.6965	0.125	-5.560	0.000	-0.942	-0.451
last activity Website Activity	0.4629	0.120	3.867	0.000	0.228	0.698
print media type2 Yes	0.4616	0.220	2.101	0.036	0.031	0.892
referral Yes	1.6006	0.382	4.195	0.000	0.853	2.349

- After dropping variables with high p-values, we get the regression result above
- Current_occupation_students, current_occupation_unemployed, profile_completed_low, profile_completed_medium and last_activity_phone_Activity all have negative coefficients, implying that a decrease in any of these variables will lead to an increase in status

“New” Model Performance

Training performance

	Accuracy	Recall	Precision	F1
0	0.82714	0.65969	0.73001	0.69307

- **The F1_score of the model is 0.69307 compared to the F1 score of 0.69451 of the previous model. So there is a slight decrease in F1 score after dropping variables with high p-values.**
- **All the variables left have p-value<0.05.**
- **We can say that lg1 is a good model for making inference.**

Converting coefficient to odds

	A	B	C	D	E
Odds	**0.08227**	**1.00124**	**0.11744**	**0.55554**	**15.14700**
Change_odd%	**-91.77267**	**0.12387**	**-88.25648**	**-44.44627**	**1414.70043**

KEY:

- **A** = const
- **B** = time_spent_on_website
- **C** = current_occupation_Student
- **D** = current_occupation_Unemployed
- **E** = first_interaction_Website

Converting coefficient to odds

	F	G	H	I	J	K
Odds	**0.08428**	**0.17184**	**0.49833**	**1.58873**	**1.58662**	**4.95606**
Change_odd%	**-91.57176**	**-82.81597**	**-50.16732**	**58.87331**	**58.66203**	**395.60635**

KEY:

- **F** = profile_completed_Low
- **G** = profile_completed_Medium
- **H** = last_activity_Phone Activity
- **I** = last_activity_Website Activity
- **J** = print_media_type2_Yes
- **K** = referral_Yes

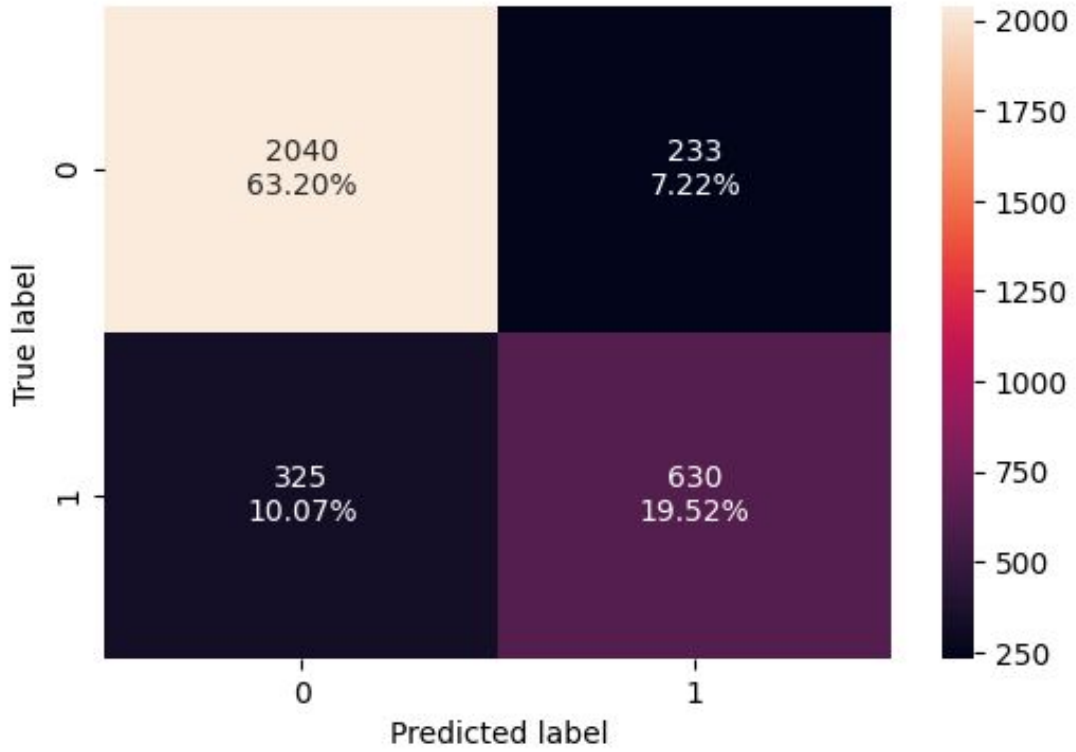
Coefficient Interpretation

- time_spent_on_website: Holding all other features constant a 1 unit change in the time spent of website will increase the odds of a lead being converted to a paying customer by 1.00124 times or a 0.12387% increase in the odds of a lead being converted.
- current_occupation_student: The odds of a lead whose current occupation is “student” are 0.11744 times less than a lead that is not a student, or a 88.25648% fewer odds of a lead being converted to a paying customer.
- print_media_type2_Yes: Holding all other features constant, a 1 unit change in the number of leads who had seen the ad of ExtraaLearn in the Magazine will increase the odds of a lead being converted to a paying customer by 1.58662 times or a 58.66203% increase in the odds of a lead being converted.
- All the other features can be interpreted in similar ways.

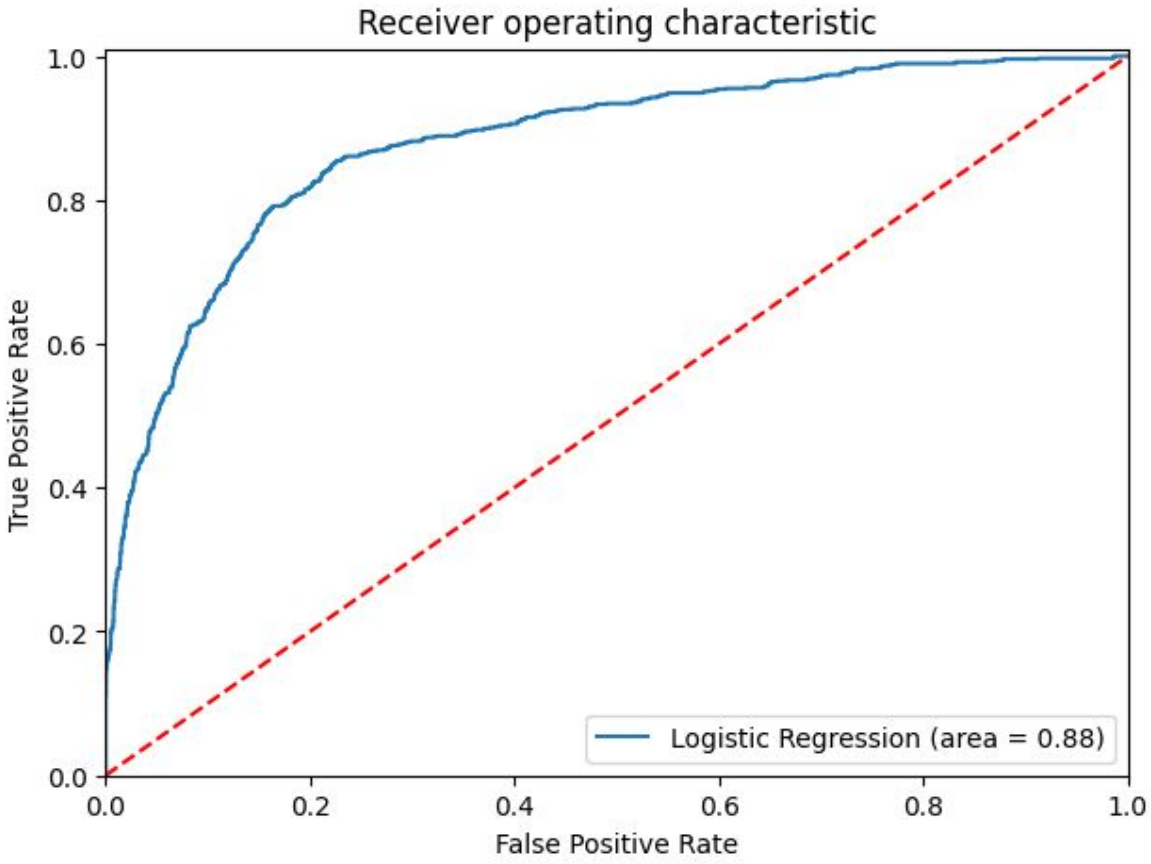
Coefficient Interpretation

Training performance:

	Accuracy	Recall	Precision	F1
0	**0.82714**	**0.65969**	**0.73001**	**0.69307**



ROC_AUC

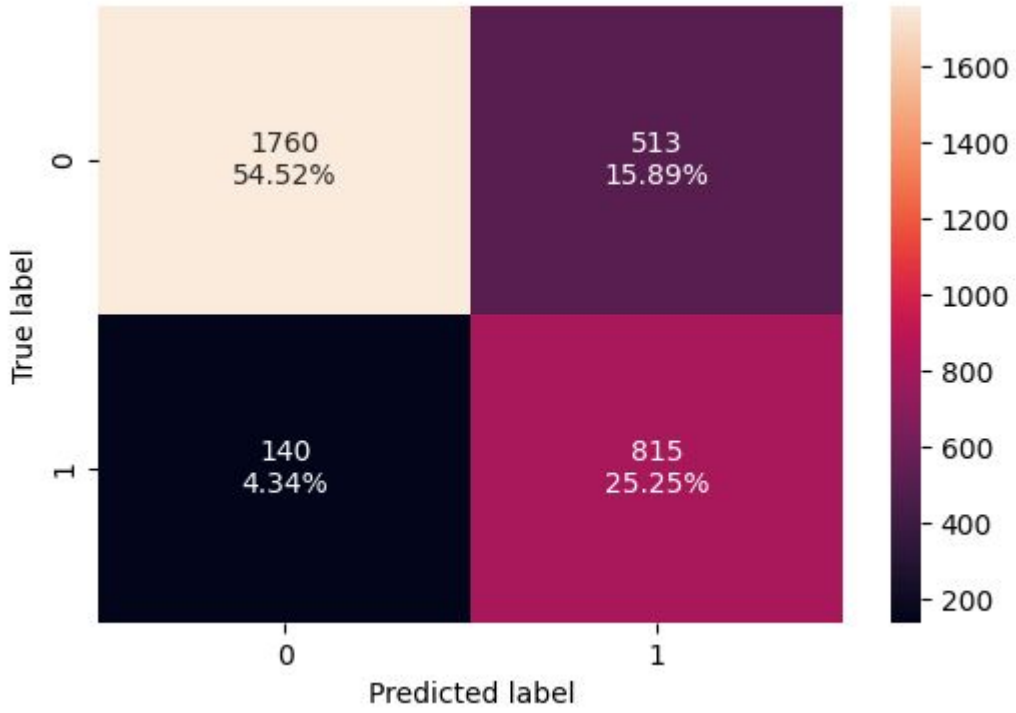


- Logistic Regression model is giving a good performance on training set
- ROC-AUC score of 0.88 on training is quite good

ROC_AUC

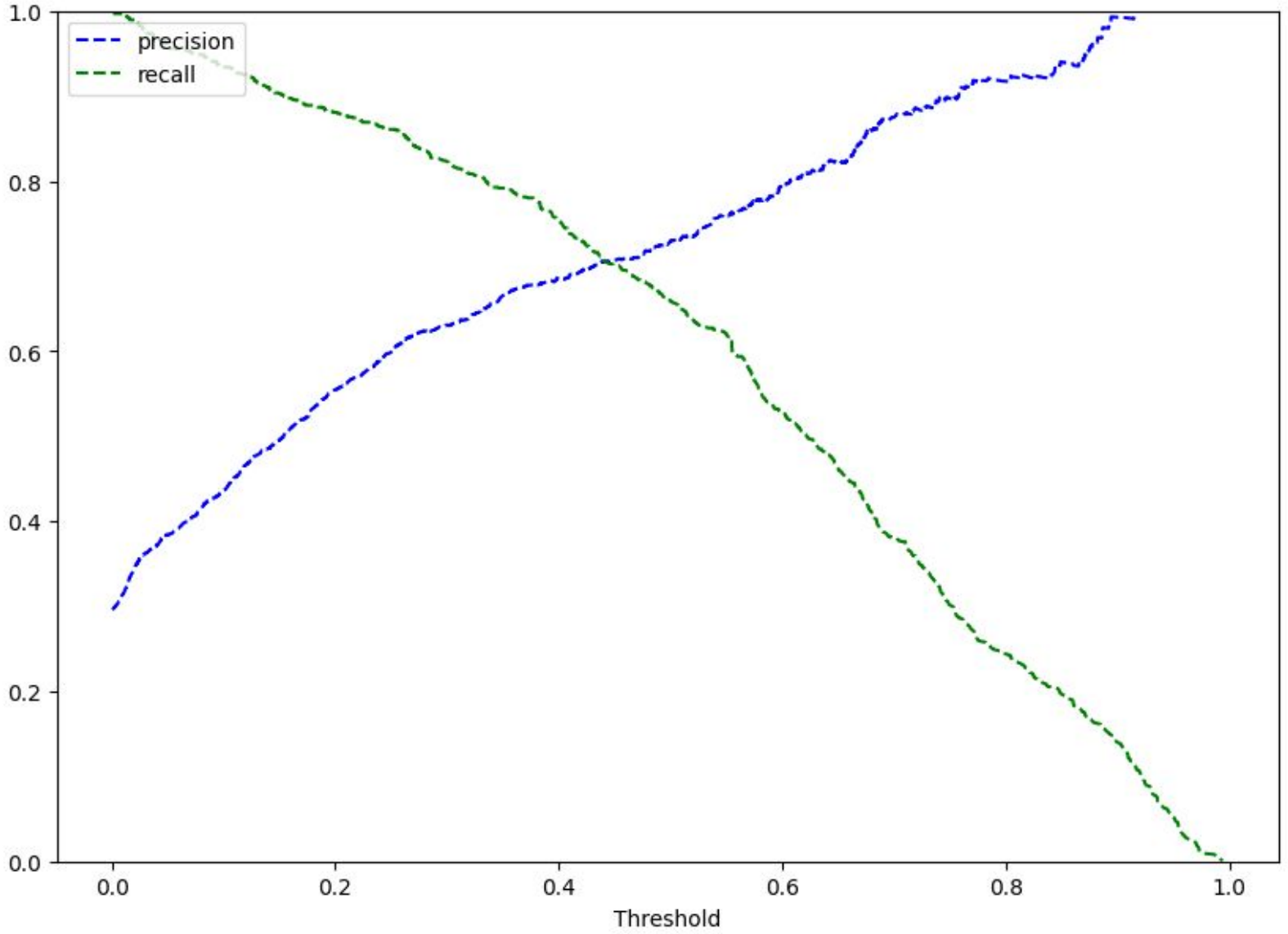
Training performance:

	Accuracy	Recall	Precision	F1
0	0.79771	0.85340	0.61370	0.71397



- Accuracy and Precision of model has reduced but the other metrics have increased.
- In particular, Recall has been improved, thus minimizing False Positives. This is what we want.
- So, the model is still giving a good performance.

Finding a better threshold via the Precision-Recall curve

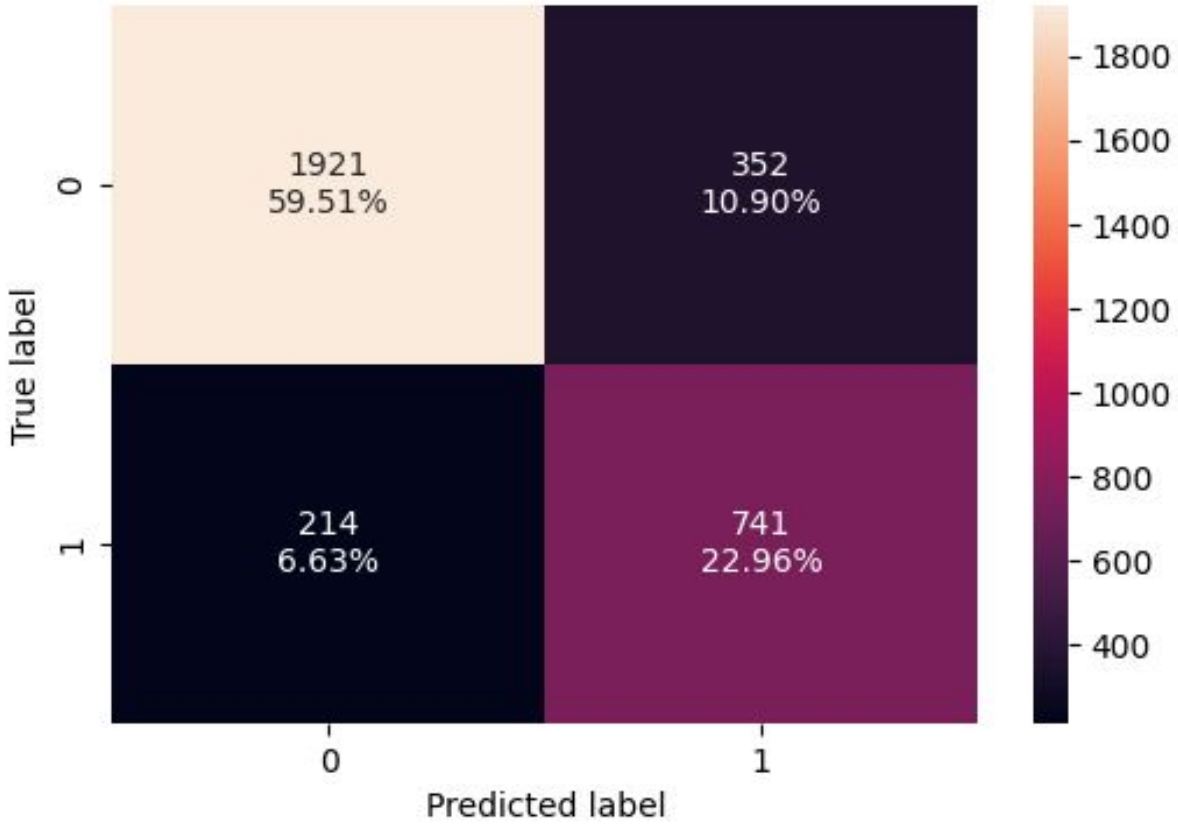


At the threshold of 0.48, we get balanced recall and precision.

Checking model performance on training set

Training performance:

	Accuracy	Recall	Precision	F1
0	0.82466	0.77592	0.67795	0.72363

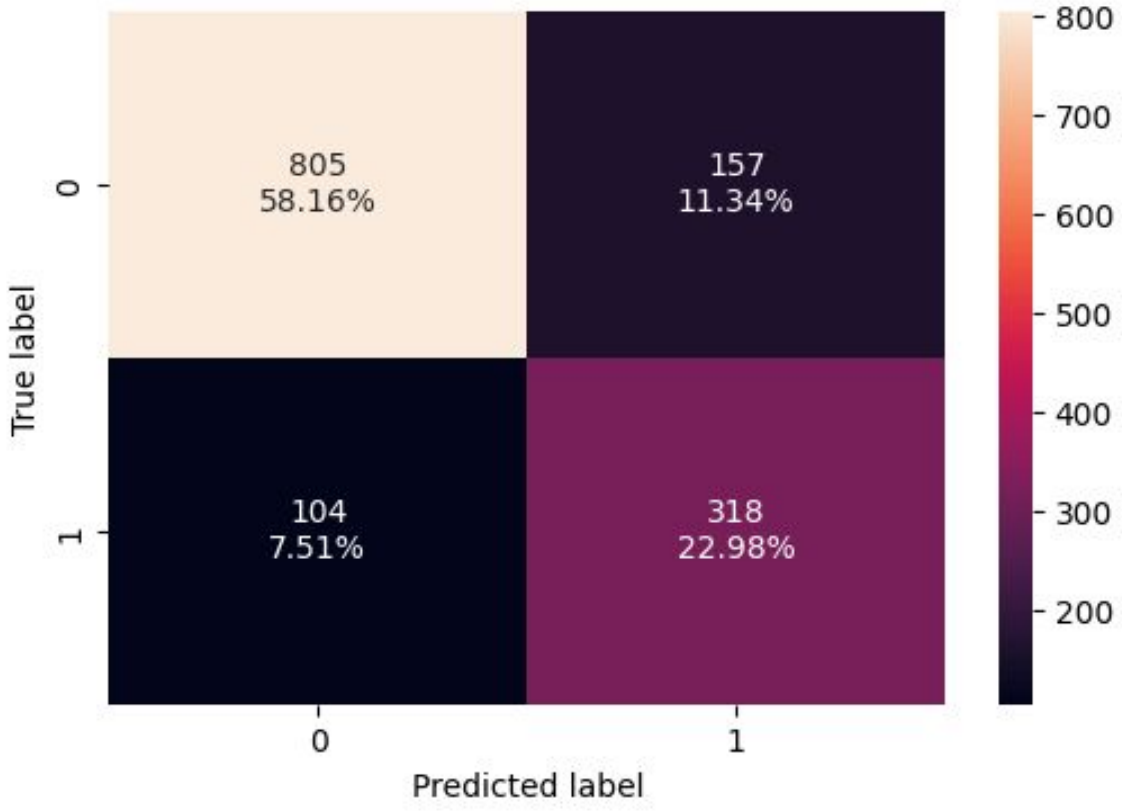


- Recall has decreased compared to the previous model, while Precision has improved. So, we might have to avoid a threshold 0.38 since Recall is our primary concern.
- From the Precision-Recall curve, it seems we might need to reduce the threshold in order to increase Recall.
- Let us check the performance on test set

Checking model performance on test set

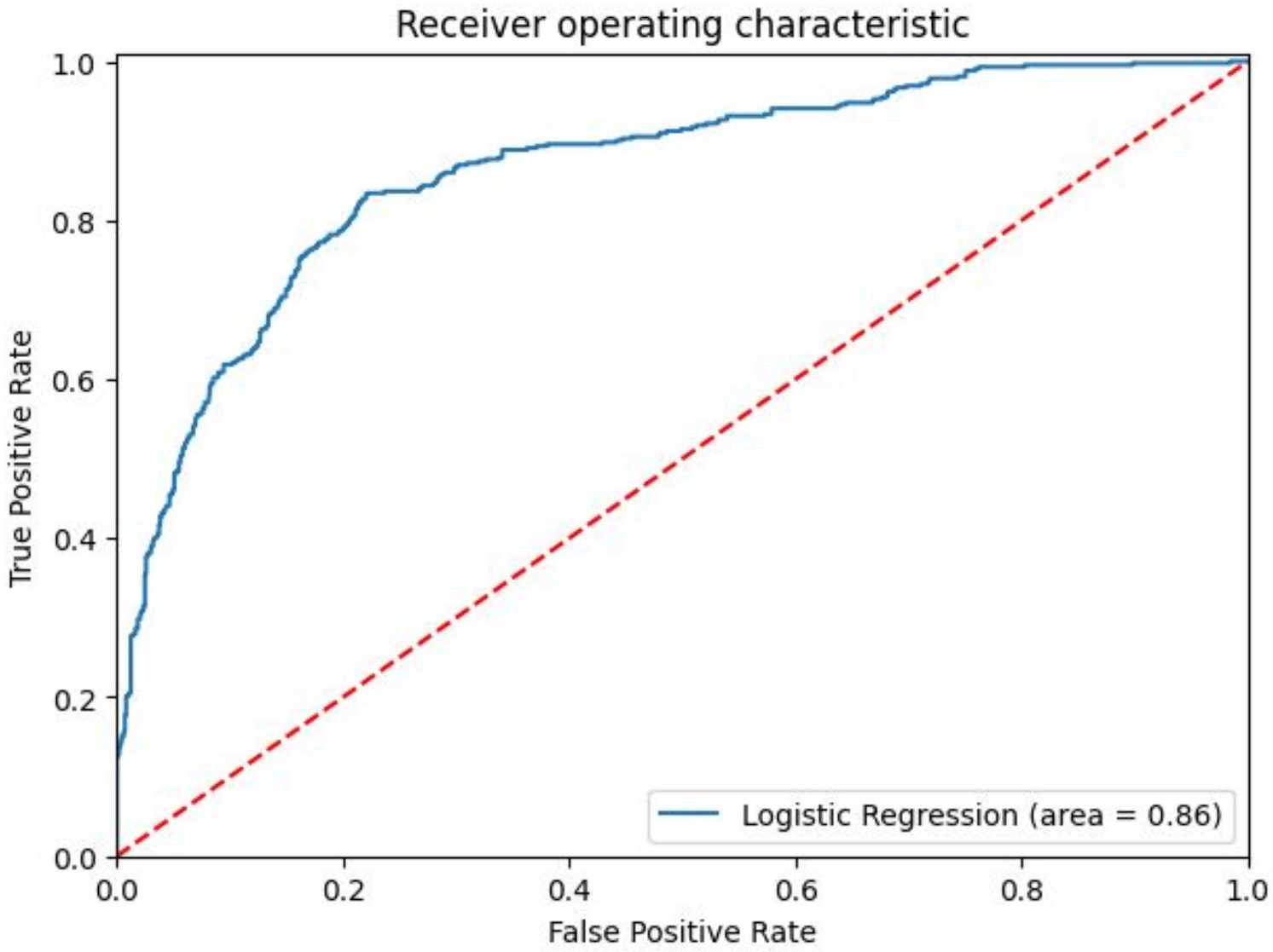
Test performance:

	Accuracy	Recall	Precision	F1
0	0.81142	0.75355	0.66947	0.70903



- Recall has decreased compared to the previous model, while Precision has improved. So, we might have to avoid a threshold 0.48 since Recall is our primary concern.
- From the Precision-Recall curve, it seems we might need to reduce the threshold in order to increase Recall.
- Let us check the performance on test set

ROC

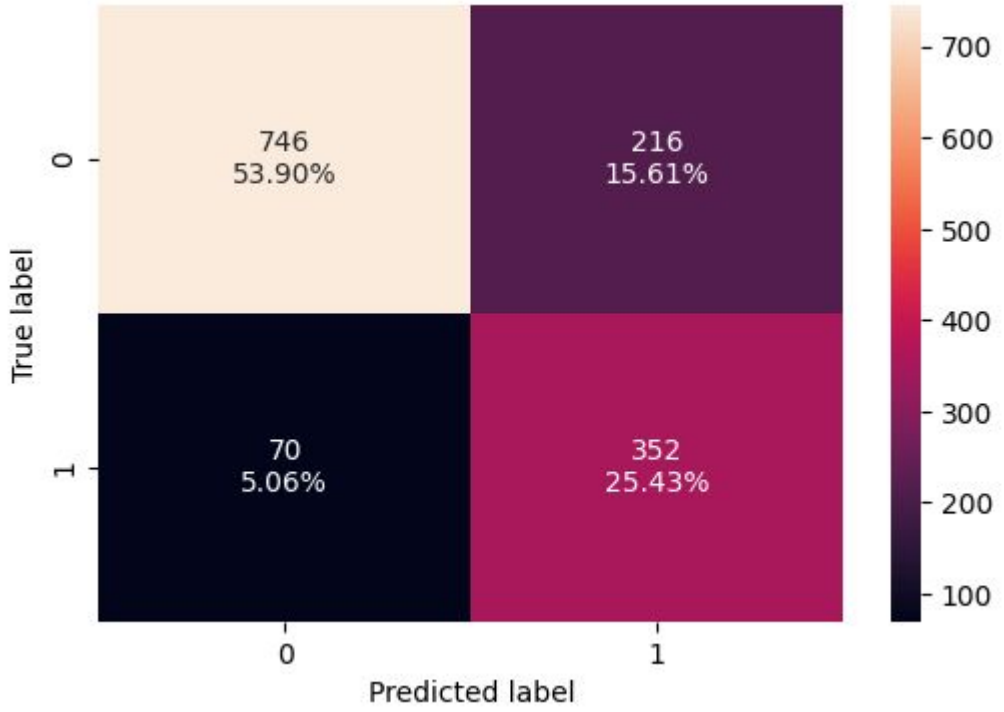


- Logistic Regression model is giving a good performance on training set
- ROC-AUC score of 0.86 on testing is quite good

Using model with threshold=0.26

Test performance:

	Accuracy	Recall	Precision	F1
0	0.79335	0.83412	0.61972	0.71111

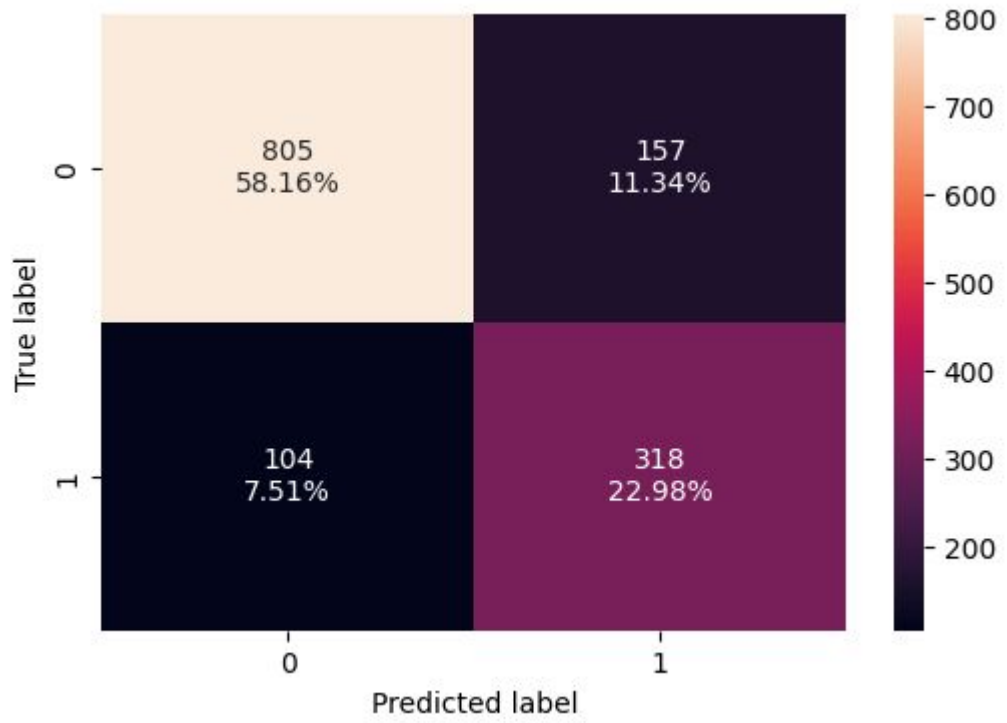


- With threshold=0.26, we can see that Recall has significantly increased, as desired

Using model with threshold=0.38

Test performance:

	Accuracy	Recall	Precision	F1
0	0.81142	0.75355	0.66947	0.70903



- Clearly, Recall reduces as the threshold increases while Precision reduces.

Model performance summary

Training performance comparison:

Logistic	Regression-default Threshold	Logistic Regression-0.26 Threshold	Logistic Regression-0.38 Threshold
Accuracy	0.82714	0.79771	0.82466
Recall	0.65969	0.85340	0.77592
Precision	0.73001	0.61370	0.67795
F1	0.69307	0.71397	0.72363

Testing performance comparison:

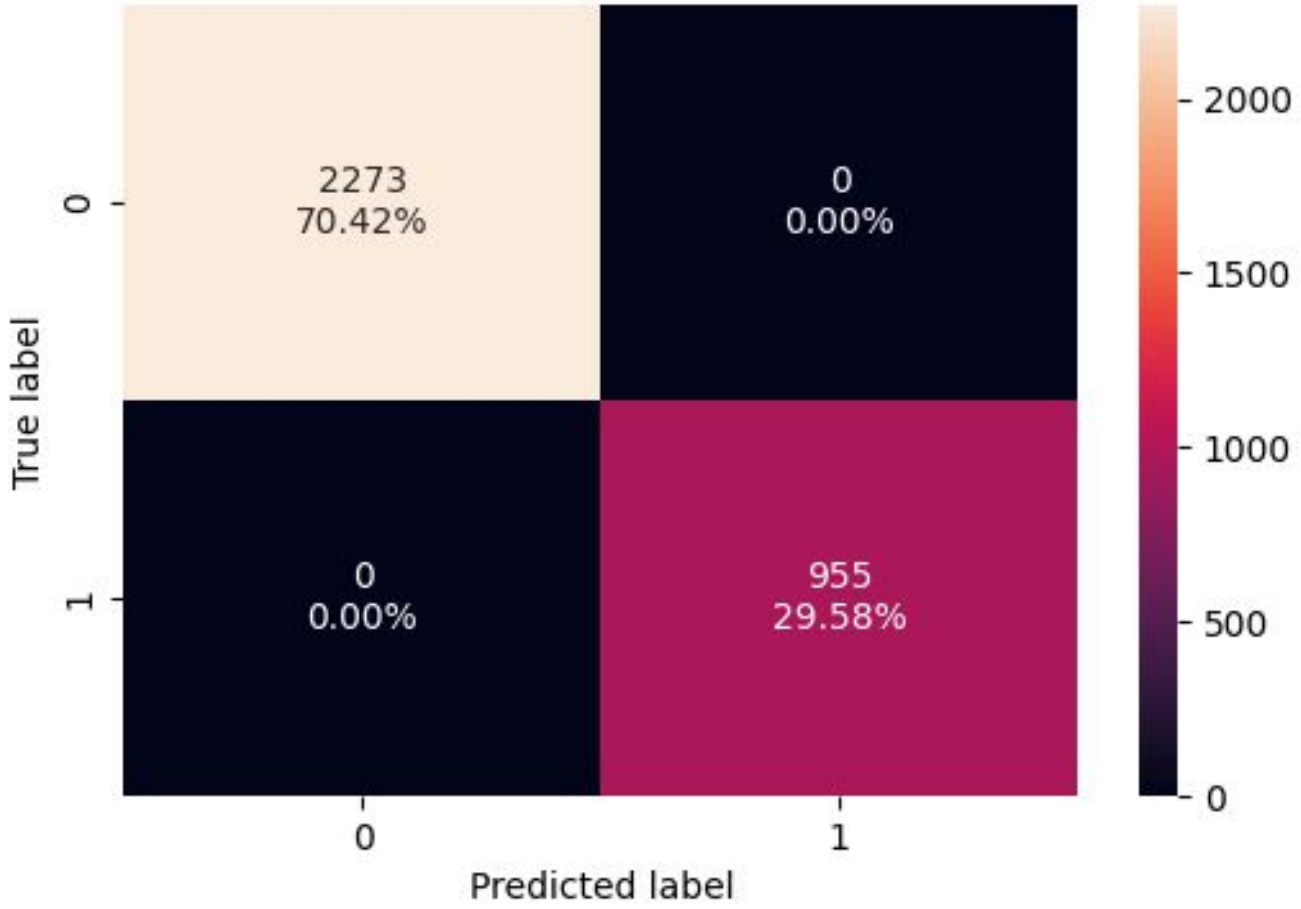
Logistic	Regression-default Threshold	Logistic Regression-0.26 Threshold	Logistic Regression-0.38 Threshold
Accuracy	0.81142	0.79335	0.81142
Recall	0.75355	0.83412	0.75355
Precision	0.66947	0.61972	0.66947
F1	0.70903	0.71111	0.70903

Observations from Logistic Regression model

- We have been able to build a predictive model that can be used by Extraalearn to predict which leads are likely to be converted with an F1 score of ~0.7 on the training set and formulate marketing policies accordingly.
- Using the model with threshold=0.26 the model will give a high recall but low precision score - Extraalearn will be able to predict which leads will be converted and will be able to minimize the possibility of losing a potential customer but might lose on resources.
- Using the model with the default threshold will give a moderate precision but lowers recall score - Extraalearn will be able to save resources by correctly predicting which leads are likely to convert but might lose a potential customer.

Decision Tree: Sklearn

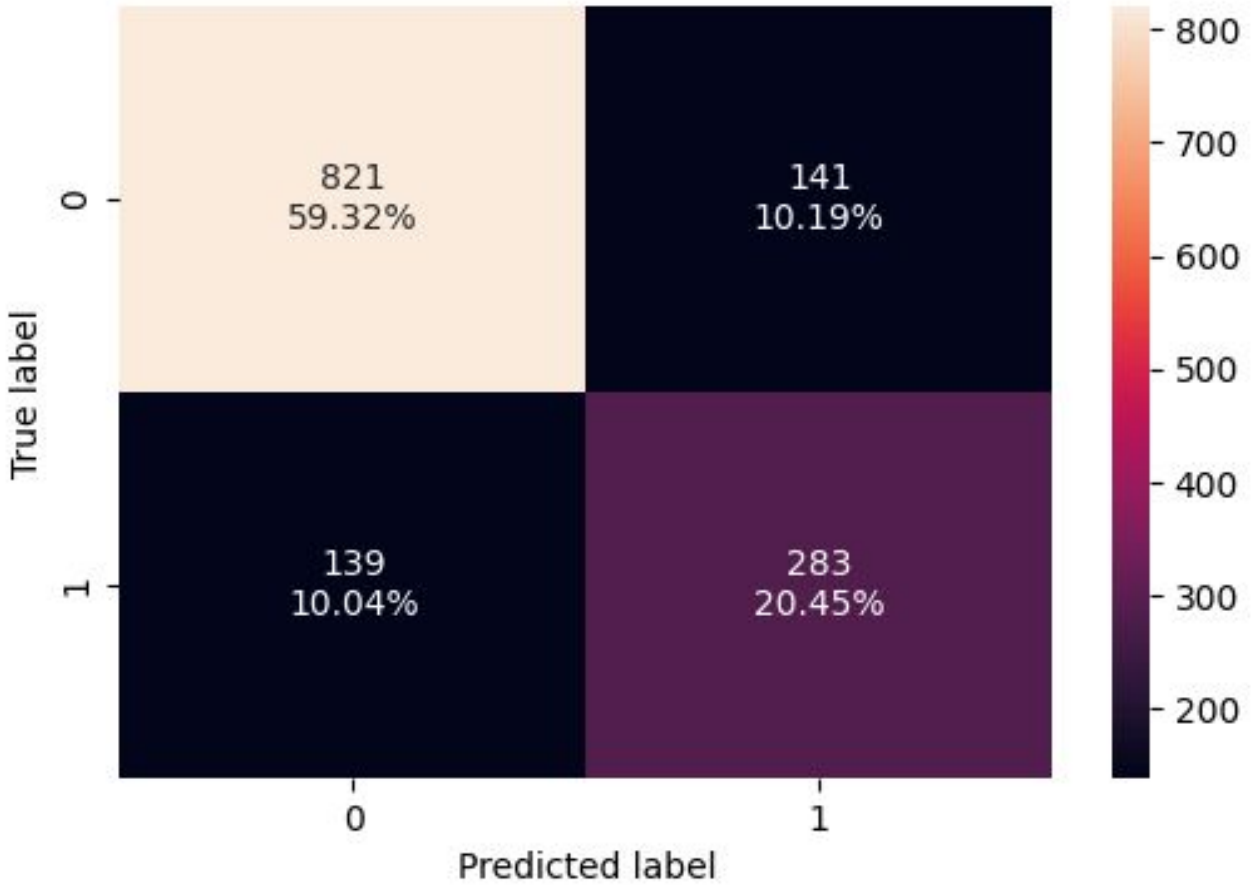
- We fit decision tree on train data with Decision tree classifier having random state = 1.



	Accuracy	Recall	Precision	F1
0	**1.00000**	**1.00000**	**1.00000**	**1.00000**

Decision Tree: Sklearn

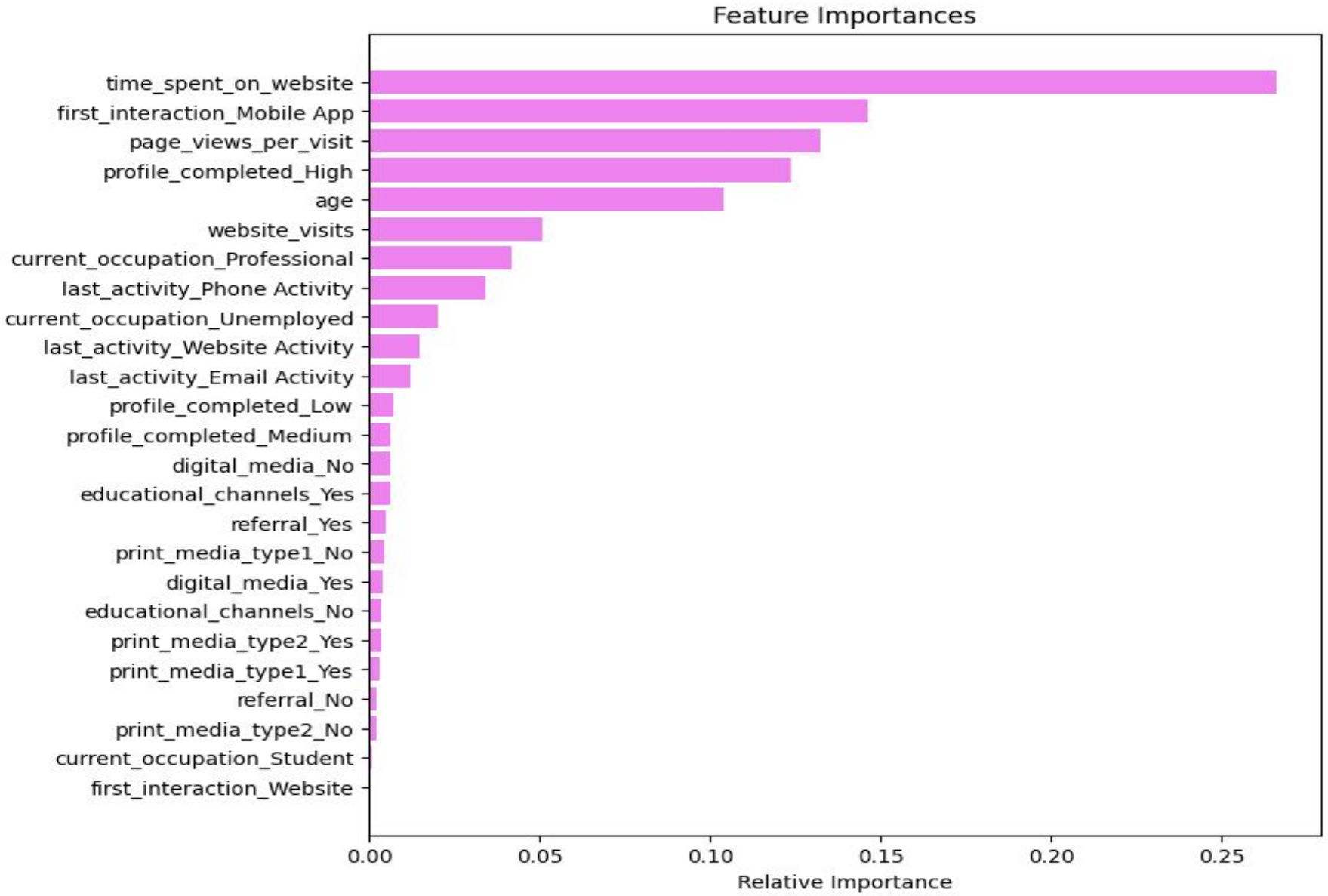
- Confusion matrix for Test data



	Accuracy	Recall	Precision	F1
0	**0.79769**	**0.67062**	**0.66745**	**0.66903**

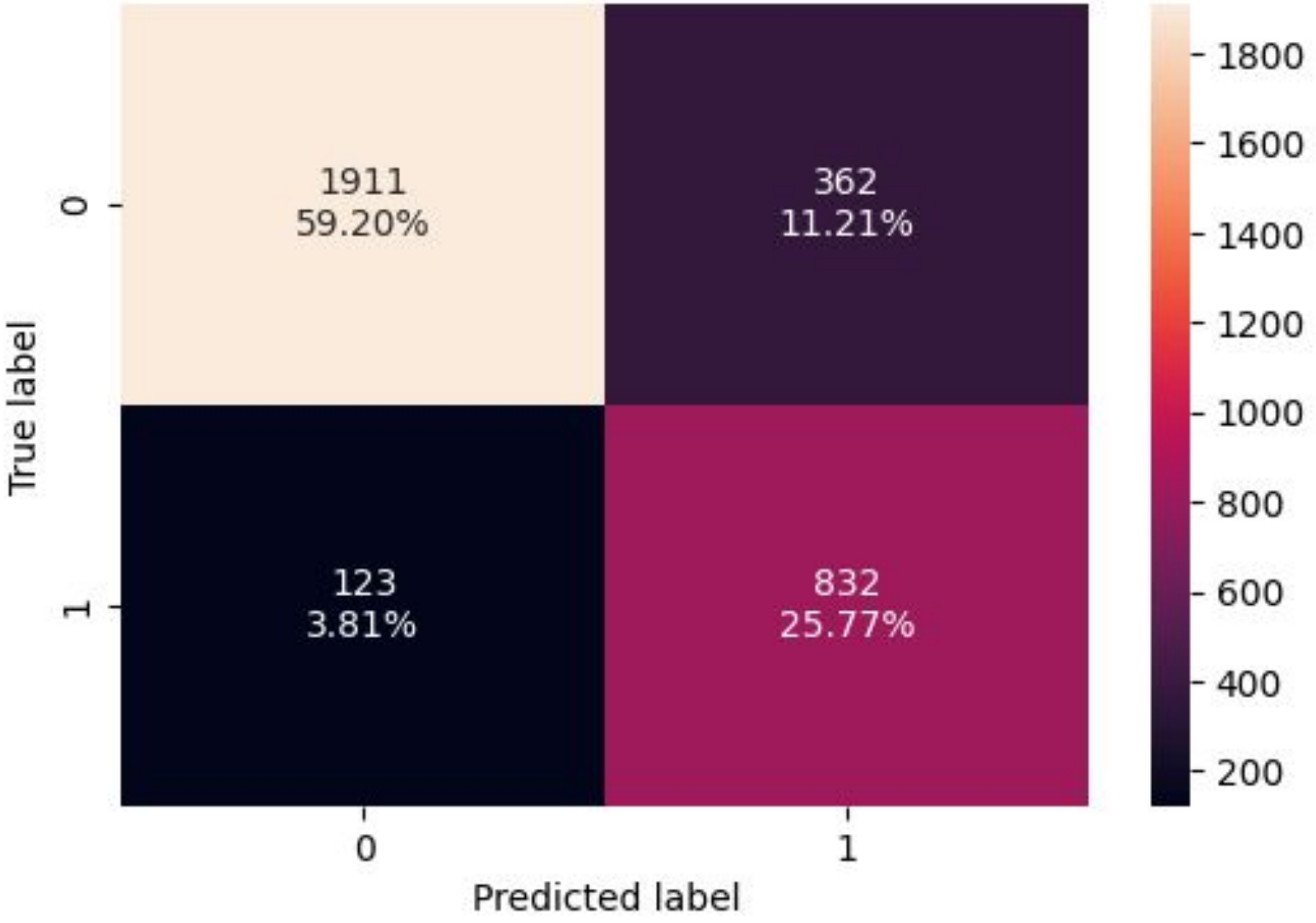
- Model is giving good and generalized results on training and test set.

Feature Importances



- Time spent on website is the most important feature in determining whether a lead is converted.

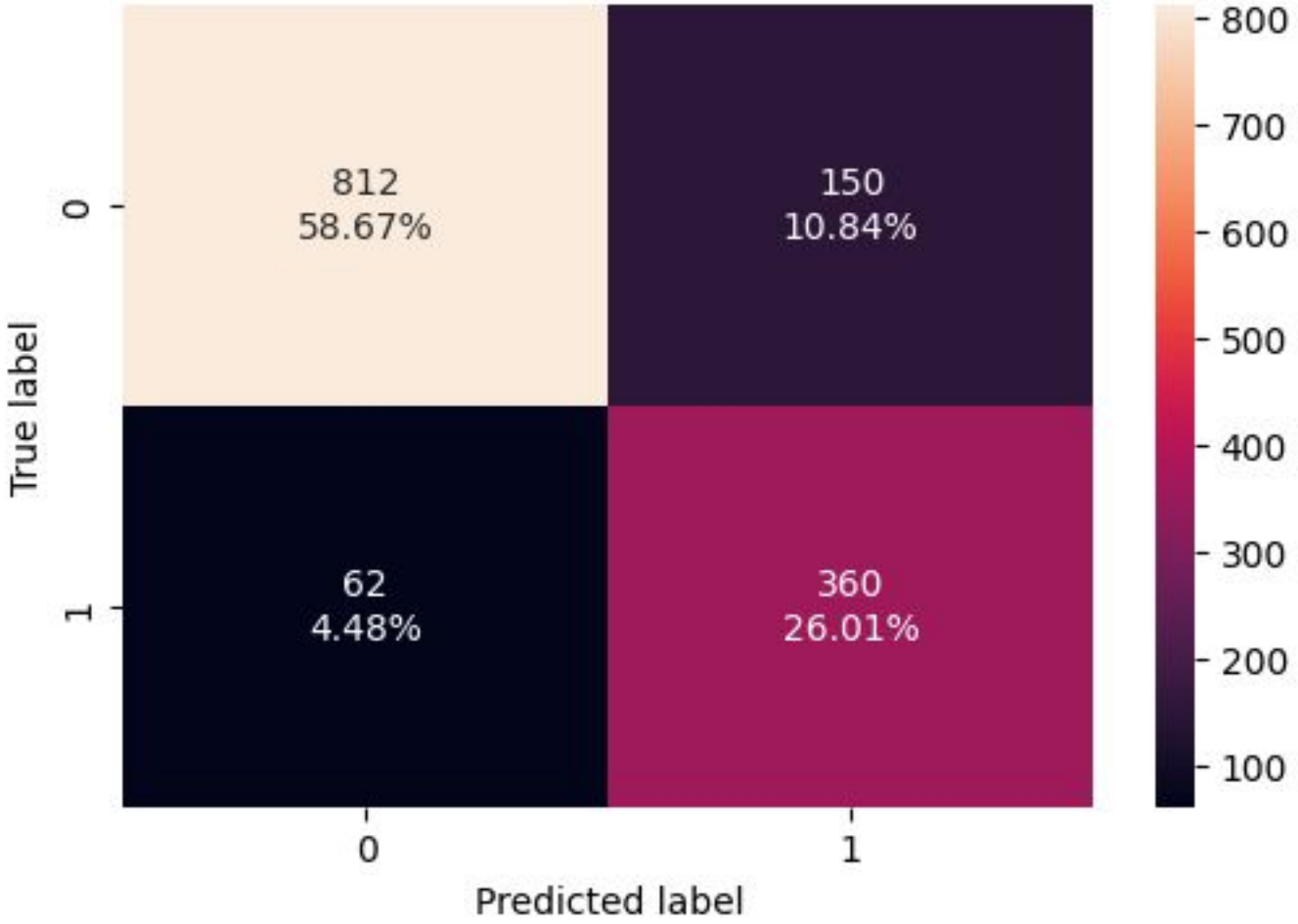
Pre-pruning the Tree: performance on training set



	Accuracy	Recall	Precision	F1
0	0.84975	0.87120	0.69682	0.77431

- Recall has improved on the training set when compared to the initial model.

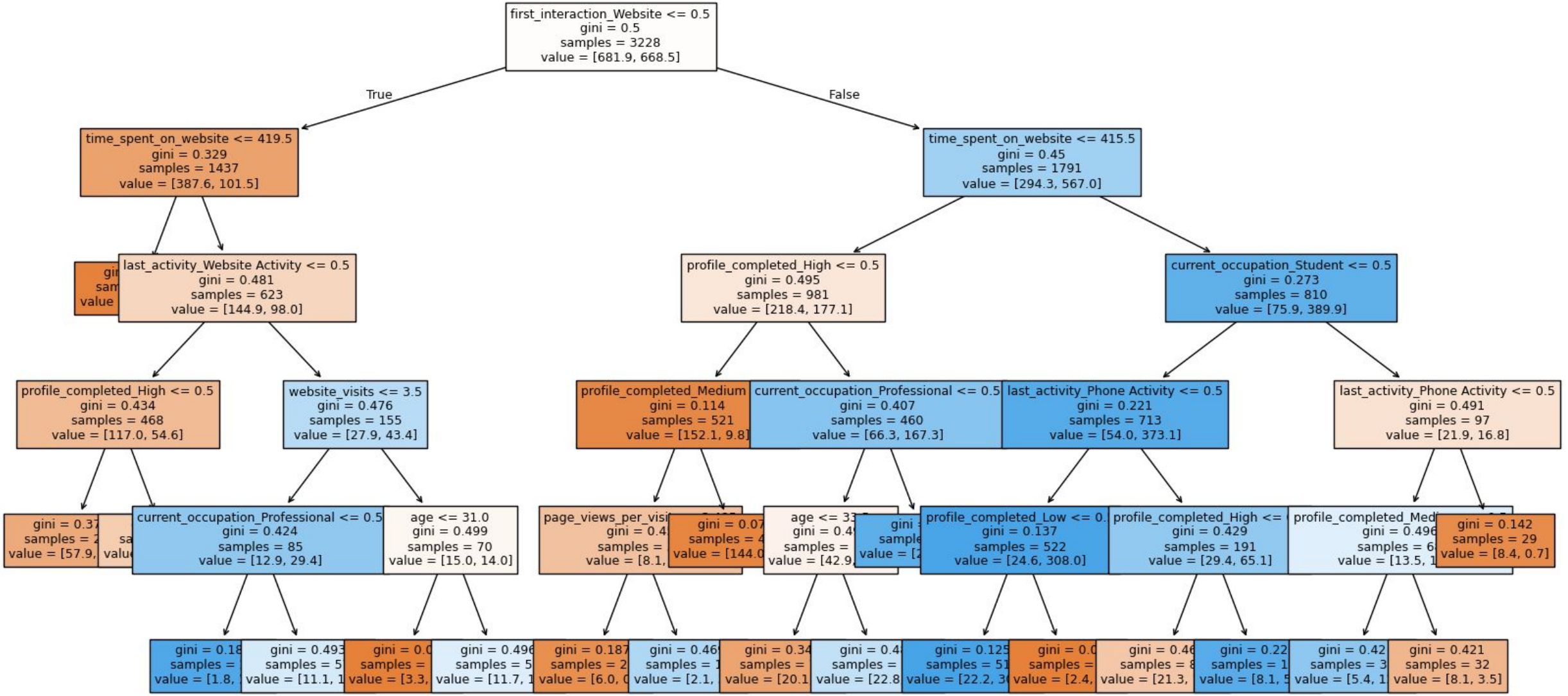
Pre-pruning the Tree: performance on test set



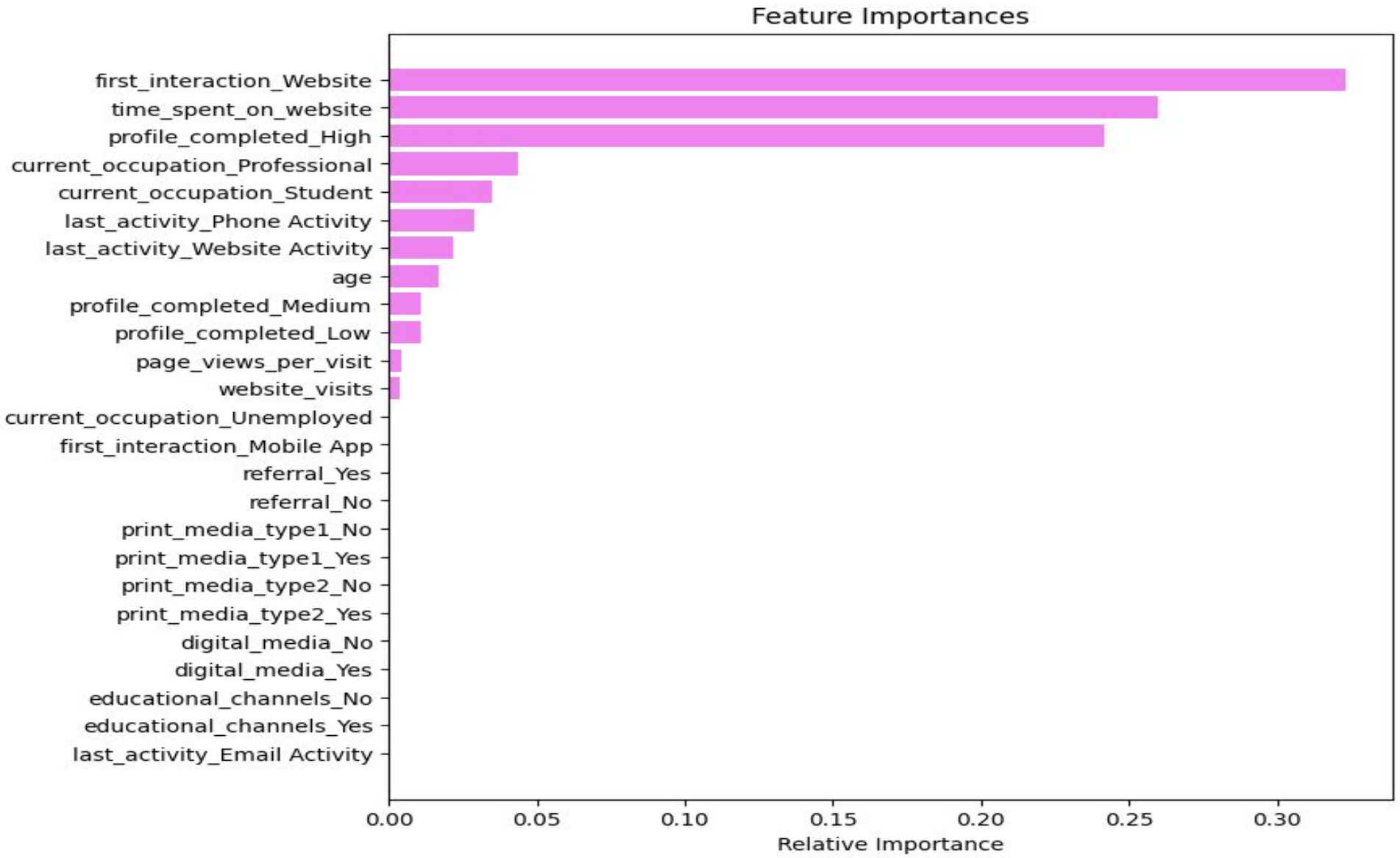
	Accuracy	Recall	Precision	F1
0	0.84682	0.85308	0.70588	0.77253

- Recall has improved on the test set when compared to the initial model.

Visualizing the Tree

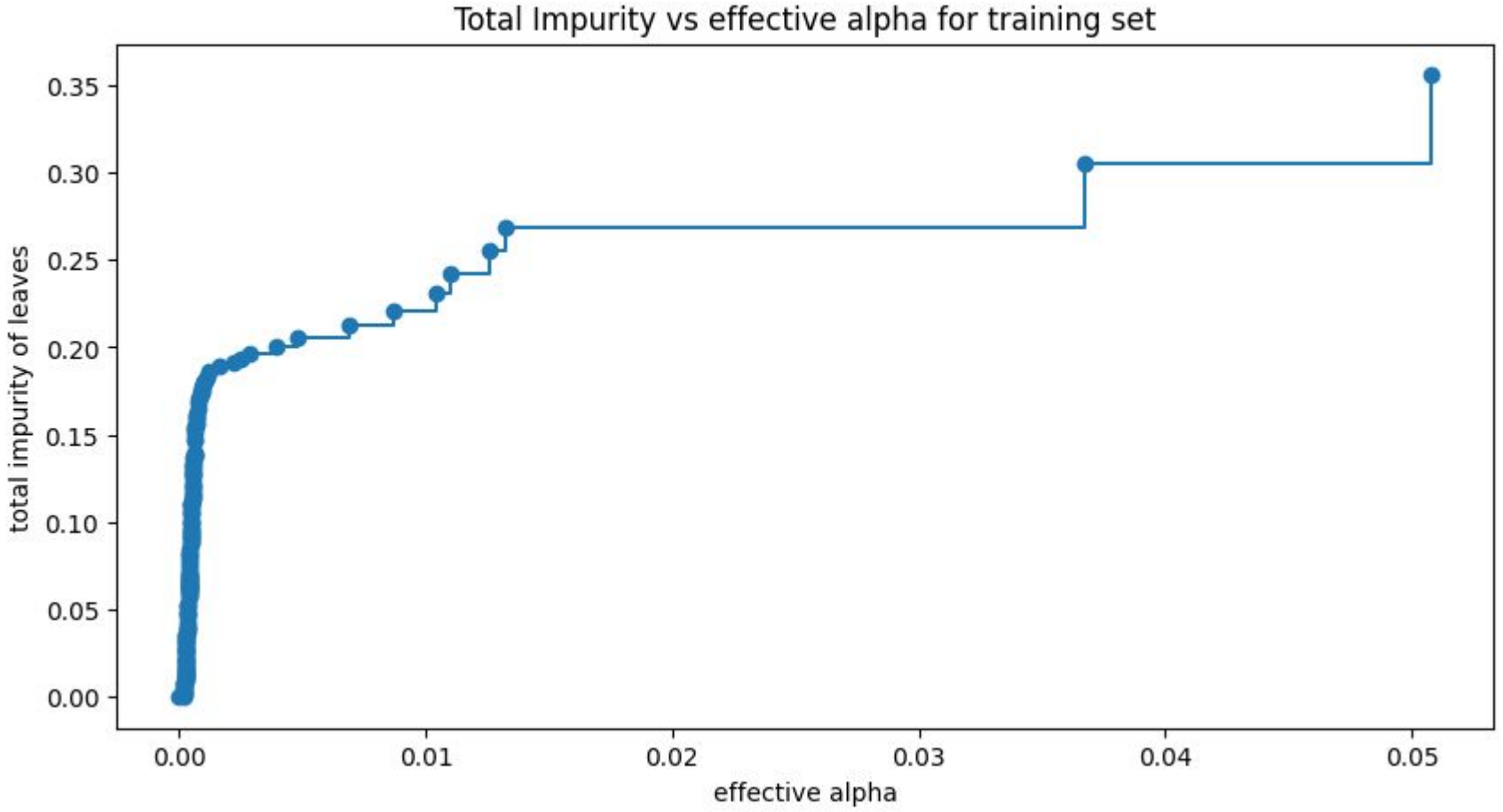


Feature Importance in the Tree Building



- First interaction website is now the most important feature in determining whether a lead is converted.

Cost Complexity Pruning



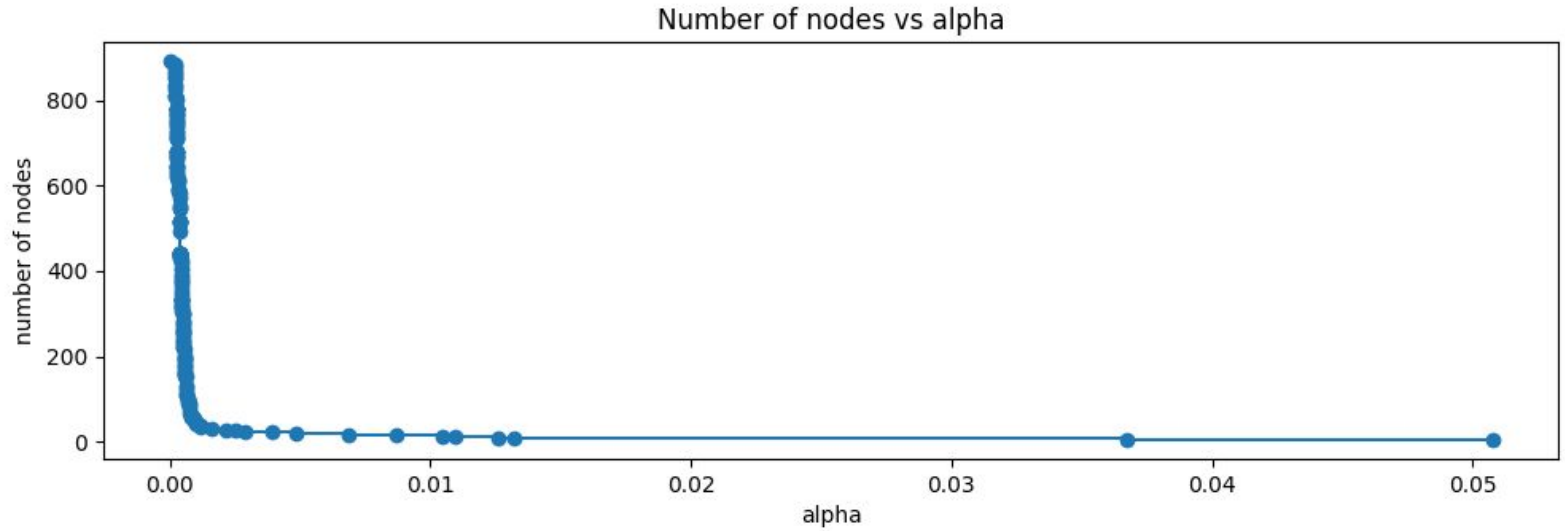
- We can make our model simpler by finding an optimal value for alpha.

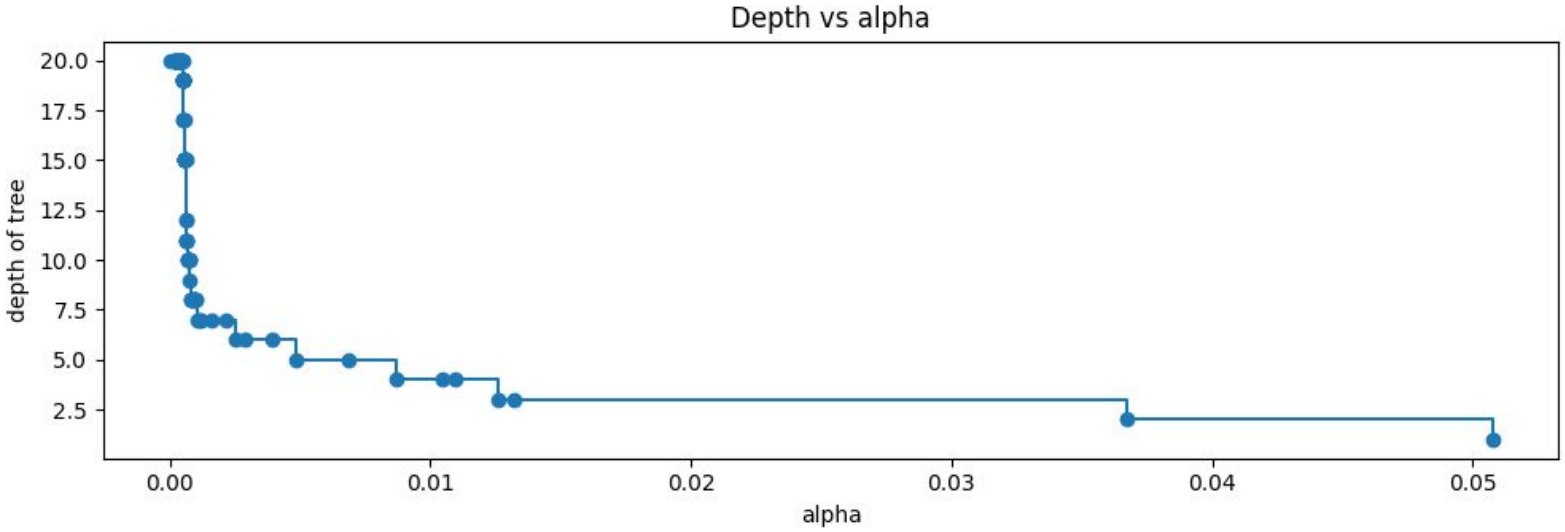
Cost Complexity Pruning

Fitting Decision Tree on training data

- Number of nodes in the last tree is: 1 with ccp_alpha: 0.06098340060606085814

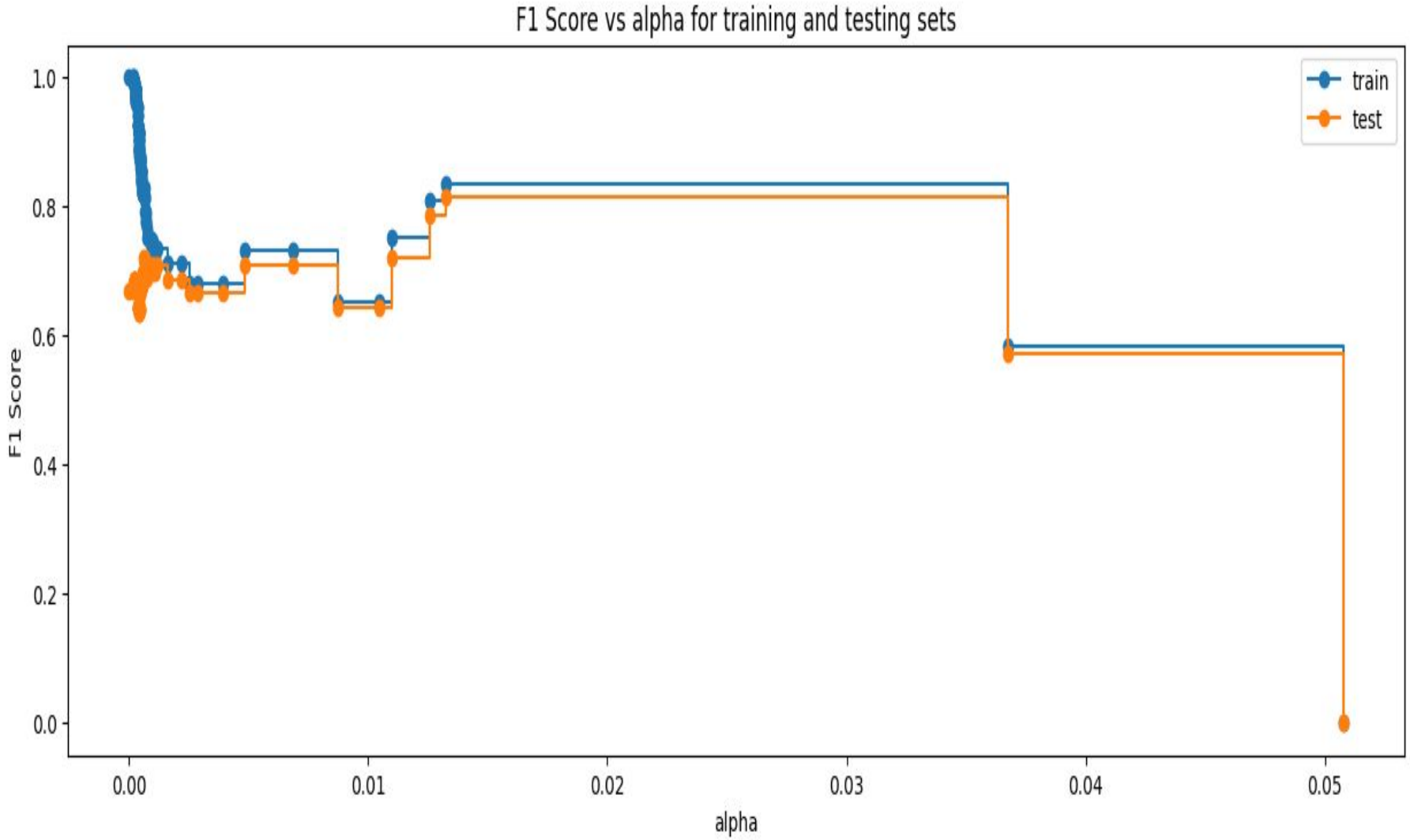
Cost Complexity Pruning





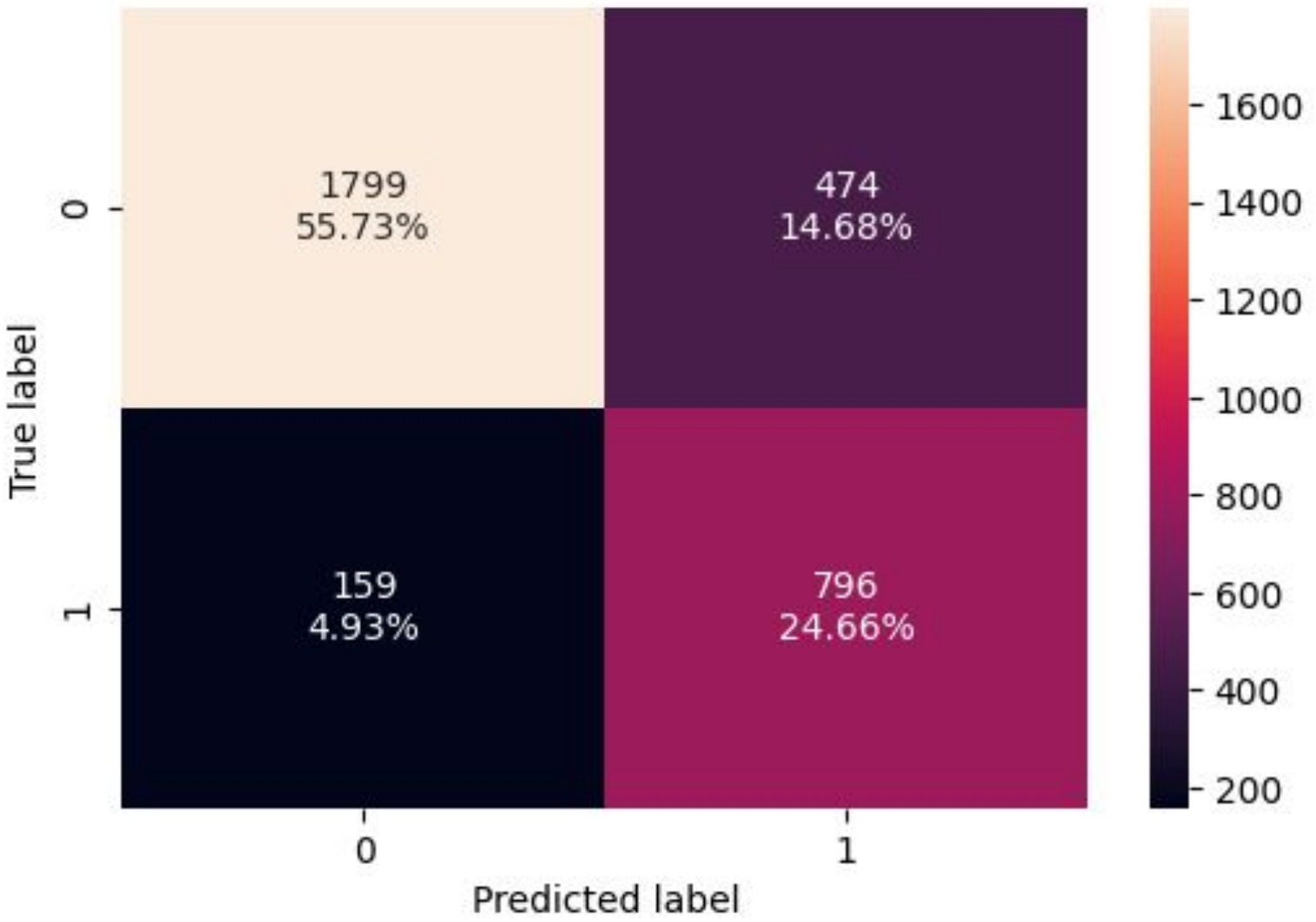
- The number of nodes decreases with increasing alpha
- The depth of nodes decreases with increasing alpha

Recall Score vs alpha for training and test sets



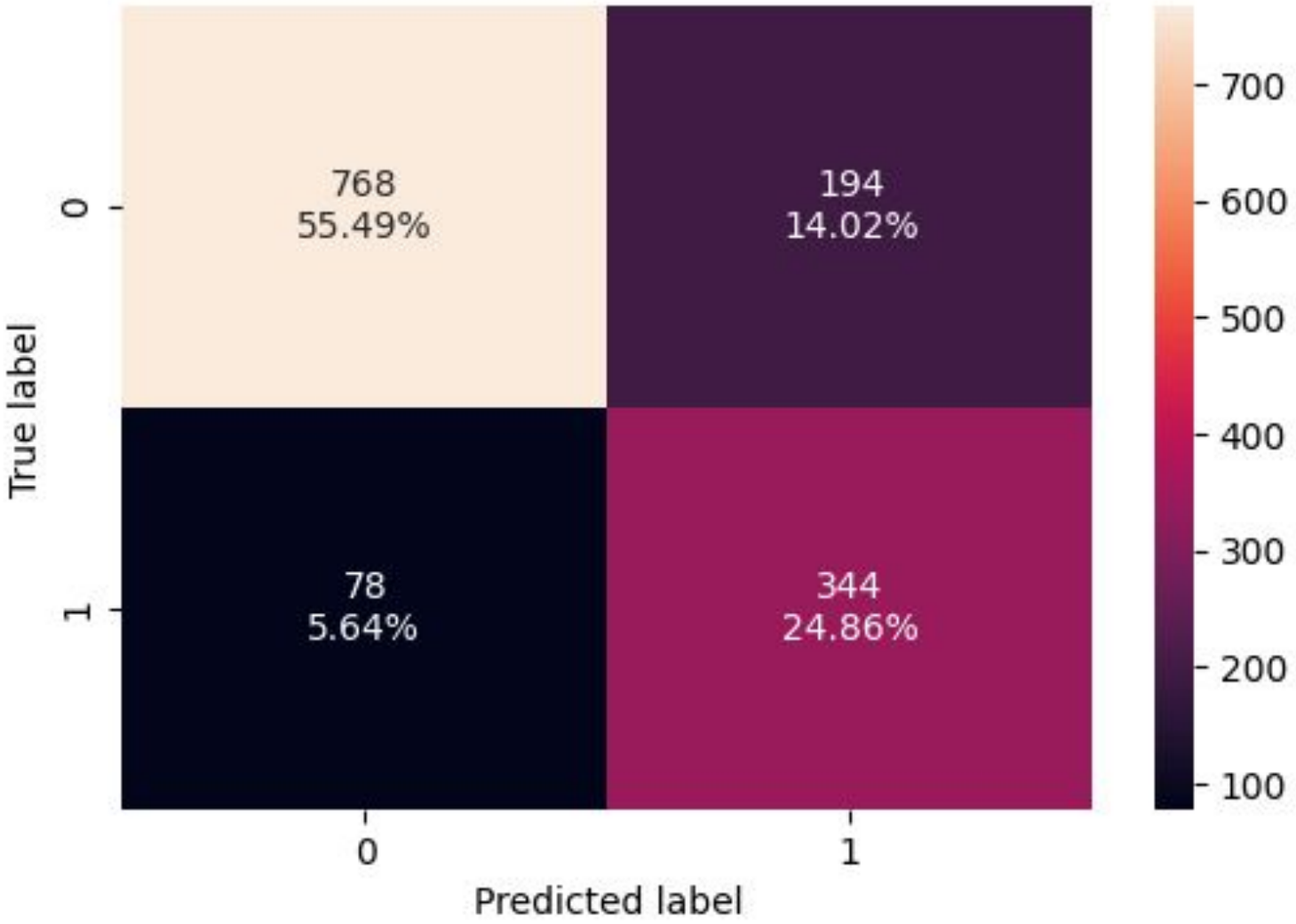
- The optimal alpha is chosen based on the maximum testing Recall
- DecisionTreeClassifier(ccp_alpha=0.013232021543488195, random_state=1)
- Optimal alpha is about 0.0132

Performance on training set

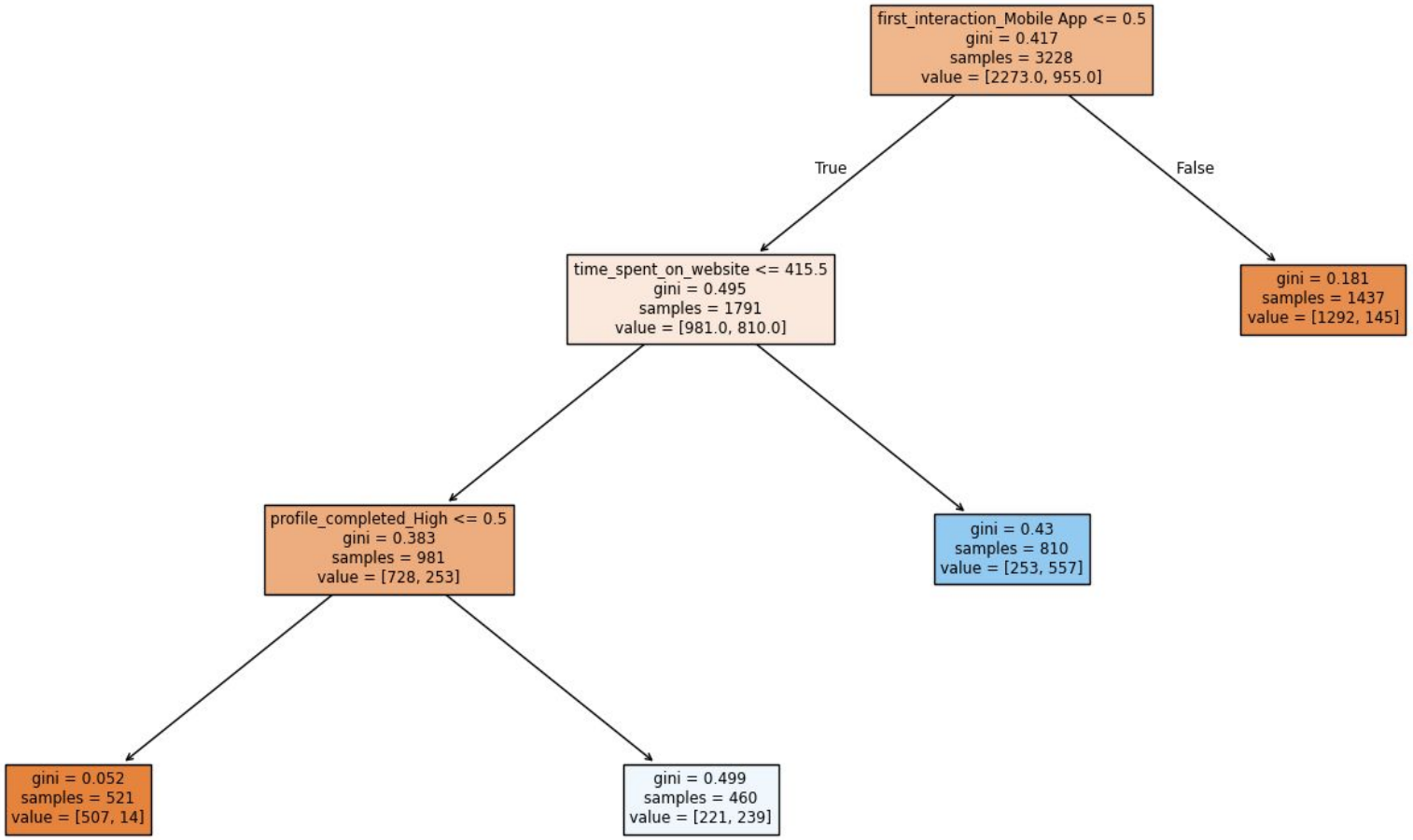


	Accuracy	Recall	Precision	F1
0	0.80390	0.83351	0.62677	0.71551

Performance on test set

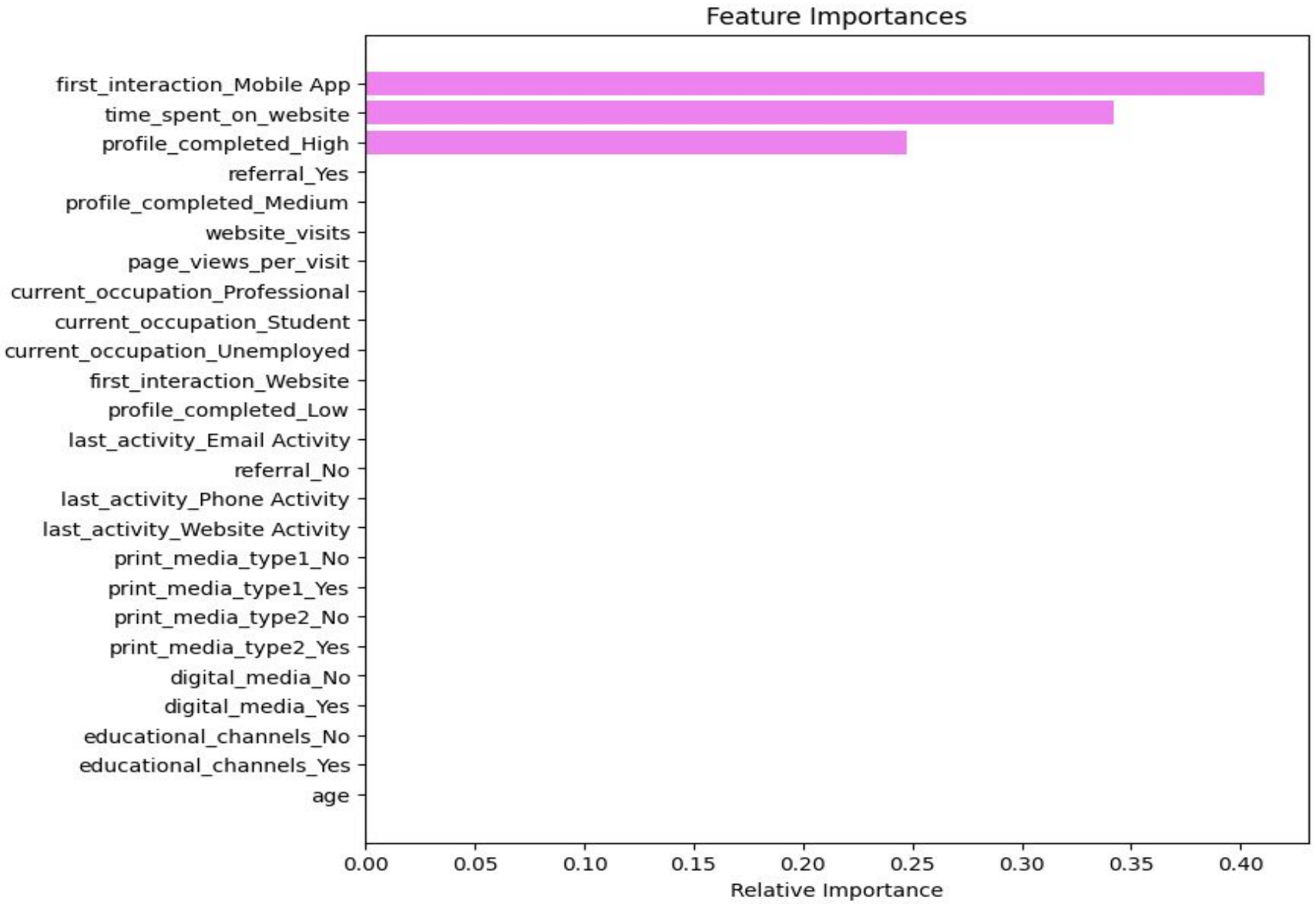


Visualising the Tree



- The has been simplified a lot

Feature Importance



- The most important feature is now first_interaction_Mobile_App
- Only 3 features are important in the overall model

Training and Testing Performance Comparison

TRAINING PERFORMANCE COMPARISON

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	1.00000	0.84975	0.80390
Recall	1.00000	0.87120	0.83351
Precision	1.00000	0.69682	0.62677
F1	1.00000	0.77431	0.71551

TESTING PERFORMANCE COMPARISON

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.79769	0.84682	0.80347
Recall	0.67062	0.85308	0.81517
Precision	0.66745	0.70588	0.63941
F1	0.66903	0.77253	0.71667

Training and Testing Performance Comparison

- Decision Tree with sklearn achieves the highest Recall value on Training set. A perfect score of 1 could indicate overfitting to the data, thereby making it difficult for the model to generalise well to unseen data.
- Both Decision Tree pre-pruning and Decision Tree post-pruning have the high recall value greater than 80% on testing set.
- Decision Tree pre-pruning has the highest Recall value of approximately 0.85 on testing set.
- It turns out that we really don’t need to prune the tree

Actionable Insights and Key Take away

- First_interaction: More people first interacted with EXTRAALEARN via the website than MobileApp. First_interaction_website is the most important feature of the model. So, EXTRAALEARN should create more opportunities for people to reach them via their website, they can also make their website more attractive to leads.
- Time_spent_on_website: the more time the leads spend on EXTRAALEARN website, the more they are likely to be converted. EXTRAALEARN can add more interesting features to their website to keep leads engaged.
- Profile_completed_high and Profile_completed_medium are two important features in our model that predicts whether a lead is converted. Measures that encourage Leads to complete their profile should be put in place. For instance, the profile section should require only necessary information from the Leads, this will help the leads complete their information faster.