

GREAT LEARNING

PROJECT ON MACHINE LEARNING



Project designer: *DURGESH KUMAR JHA*

Session: *2024-25*

From the business perspective we have to find the driving factors which influences the trends for the online cutting-edge technologies on their platform, and analyse the data so we can provide the recommendation to the company.

<i>1. Data interpretation</i>	<i>Page 5</i>
<i>2. Univariate analysis with proper interpretation Key Questions.</i>	<i>Page 6</i>
<i>3. Bivariate & Multivariate analysis with proper interpretation Key Questions.</i>	<i>Page 13</i>
<i>4. Logistic Regression model preparation.</i>	<i>Page 17</i>
<i>5. Model performance Improvement with the different scales.</i>	<i>Page 20</i>
<i>6. Final Model with the different improvement with in the model.</i>	<i>Page 25</i>
<i>7. Visualizing the Decision Tree Classifier.</i>	<i>Page 30</i>
<i>8. Final Model performance with the different matrix we used</i>	<i>Page 32</i>
<i>9. Actionable Insights and Business recommendation.</i>	<i>Page 33</i>

PROBLEM STATEMENT

In the present scenario due to Covid-19, the online education sector has witnessed rapid growth and is attracting a lot of new customers. Due to this rapid growth, many new companies have emerged in this industry. With the availability and ease of use of digital marketing resources, companies can reach out to a wider audience with their offerings. The customers who show interest in these offerings are termed as leads. There are various sources of obtaining leads for Edtech companies, like

OBJECTIVE

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

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

DATA DESCRIPTION

Sl. NO	Column Name	Description
1	ID	ID of the lead
2	Age	Age of the lead

3	<i>current_occupation</i>	<ul style="list-style-type: none"> Current occupation of the lead. Values include 'Professional', 'Unemployed', and 'Student'
4	<i>first_interaction</i>	How did the lead first interact with ExtraaLearn. Values include 'Website', 'Mobile App'
5	<i>profile_completed</i>	the percentage of the profile filled by the lead on the website/mobile app Values include Low - (0-50%), Medium - (50-75%), High (75-100%)
6	<i>website_visits</i>	How many times has a lead visited the website
7	<i>time_spent_on_website</i>	Total time spent on the website in seconds
8	<i>page_views_per_visit</i>	Average number of pages on the website viewed during the visits
9	<i>last_activity</i>	Last interaction between the lead and ExtraaLearn
10	<i>print_media_type1</i>	<ul style="list-style-type: none"> Flag indicating whether the lead had seen the ad of ExtraaLearn in the Newspaper.
11	<i>print_media_type2</i>	Flag indicating whether the lead had seen the ad of ExtraaLearn in the Magazine.
12	<i>digital_media</i>	Flag indicating whether the lead had seen the ad of ExtraaLearn on the digital platforms
13	<i>educational_channels</i>	Flag indicating whether the lead had heard about ExtraaLearn in education channels like online forums, discussion threads, educational websites, etc
14	<i>referral</i>	Flag indicating whether the lead had heard about ExtraaLearn through reference.
15	<i>Status</i>	Flag indicating whether the lead was converted to a paid customer or not.

List of Images

S.NO.	TOPIC
Image 1	The top 5 rows of the data set
Image 2	The total number of rows and columns in data set
Image 3	The statistical summary of data set
Image 4	The total number of missing values in data set
Image 5	Box plot and His plot of age distribution in data set
Image 6	Box plot and His plot of <i>time spend</i> in data set
Image 7	Box plot and His plot of pages views per visit in data set
Image 8	Labeled Bar plot of occupation of the lead in data set
Image 9	Labeled Bar plot of number of children in data set
Image 10	Labeled Bar plot of status of profile completion in data set
Image 11	Labeled Bar plot of last activity on web in data set
Image 12	Labeled Bar plot of observation on print media type 1 in data set
Image 13	Labeled Bar plot of observation on print media type 2 in data set
Image 14	Labeled Bar plot of observation on room type reserved in data set
Image 15	Labeled Bar plot of observation on educational channels in data set

Image 16	Labeled Bar plot of observation on referrals in data set
Image 17	Labeled Bar plot of observation on status of lead converted to paid customer
Image 18	Heat map showing the co-relation b/w many numerical variables
Image 19	Image shows relationship b/w age distribution, and the time spend on the web
Image 20	Image shows relationship b/w time spend on the web and their conversion to paid customer
Image 21	Image shows relationship b/w visited web and their conversion to paid customer
Image 22	Image shows relationship b/w visited web and their conversion to paid customer
Image 23	Image shows the relationship between the lead visited on the website and their conversion.
Image 24	Image shows the percentage of profile filled by the lead on mobile or on the website.
Image 25	Image shows the percentage of profile filled by the lead on mobile or on the website.
Image 26	Image shows the conversion rate of leads who see the ad on Newspaper.
Image 27	Image shows the conversion rate of leads who see the ad on Magazine.
Image 28	Image shows conversion rate of leads heard about the Extraalearn on education channels.
Image 29	Image shows the status of the leads who learn about the Extraalearn through the referral's.
Image 30	This image shows the outliers present in our data set.
Image 31	Image shows we have taken 3459 rows for training set and 1153 rows for test set.
Image 32	Image shows the performance of our model we have built with data available.
Image 33	Image shows that recall matrix for Logistic Regression.
Image 34	Confusion matrix on training set by Logistic regression.
Image 35	Image shows Recall value for the test matrix for the Logistic Regression.
Image 36	Confusion matrix on test set by Logistic regression.
Image 37	Image shows recall value for the Naïve bayes classifier with the training data.
Image 38	Confusion matrix on training set by Naive Bayes Classifier.
Image 39	Image shows recall score for the test data for Naïve Bayes Classifier.
Image 40	Confusion matrix on test set by Naive Bayes Classifier.
Image 41	Image shows our recall value, Precision, and F1 score for the training set for KNN Classifier.
Image 42	Confusion matrix on training set by KNN Classifier.
Image 43	Image shows recall value, Precision, and F1 score for the test set for KNN Classifier.
Image 44	Confusion matrix on test set by KNN Classifier.
Image 45	Image shows recall value, Precision, and F1 score for training set for Decision Tree Classifier.
Image 46	Confusion matrix on training set by Decision Tree Classifier.
Image 47	Confusion matrix on test set by Decision Tree Classifier.
Image 48	Image shows we have deal with the Logistic Regression from Multicollinearity.
Image 49	Image shows we have deal with the high p-value variables.
Image 50	This is our final model as we have only significant features available in this model.
Image 51	This is our final model as we have only significant features available in this model.
Image 52	Above Image shows the ROC curve which is tiled around the curve.
Image 53	Image shows recall value, Precision, and F1 score training set for tuned Logistic Regression.
Image 54	Confusion matrix on training set for tuned Logistic Regression.
Image 55	Image shows recall value, Precision, and F1 score for test set for tuned Logistic Regression.
Image 56	Confusion matrix on test set for tuned Logistic Regression.
Image 57	Above Image shows the different values with the different K values.
Image 58	Image shows recall value, Precision, and F1 score for the training set for tuned KNN Model.
Image 59	Confusion matrix on training set for tuned KNN Model.
Image 60	Image shows recall value, Precision, and F1 score for the test set for tuned KNN Model.
Image 61	Confusion matrix on test set for tuned KNN Model.
Image 62	Image shows recall value, Precision, F1 score for the training set for tuned Decision Tree.
Image 63	Confusion matrix on training set for tuned Decision Tree Classifier.
Image 64	Image shows recall value, Precision and F1 score for test set tuned Decision Tree Classifier.
Image 65	Confusion matrix on test set for tuned Decision Tree Classifier.
Image 66	Image showing the visualizing the Decision Tree.
Image 67	Image showing final model with all the matrix

Data Overview

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

Image 1

Above Image shows the top 5 rows of data set.

	ID	age	current_occupation	first_interaction	profile_completed	website_visits	time_spent_on_website	page_views_per_visit	last_activity	print_media_type1	print_media_type2
4607	EXT4608	35	Unemployed	Mobile App	Medium	15	360	2.17000	Phone Activity	No	
4608	EXT4609	55	Professional	Mobile App	Medium	8	2327	5.39300	Email Activity	No	
4609	EXT4610	58	Professional	Website	High	2	212	2.69200	Email Activity	No	
4610	EXT4611	57	Professional	Mobile App	Medium	1	154	3.87900	Website Activity	Yes	
4611	EXT4612	55	Professional	Website	Medium	4	2290	2.07500	Phone Activity	No	

Image 2

Above Image shows the bottom 5 rows of data set.

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

Image 3

Above Image shows the statistical summary of data set.

Data columns (total 15 columns):

#	Column	Non-Null Count	Dtype
0	ID	4612 non-null	object
1	age	4612 non-null	int64
2	current_occupation	4612 non-null	object
3	first_interaction	4612 non-null	object
4	profile_completed	4612 non-null	object
5	website_visits	4612 non-null	int64
6	time_spent_on_website	4612 non-null	int64
7	page_views_per_visit	4612 non-null	float64
8	last_activity	4612 non-null	object
9	print_media_type1	4612 non-null	object
10	print_media_type2	4612 non-null	object
11	digital_media	4612 non-null	object
12	educational_channels	4612 non-null	object
13	referral	4612 non-null	object
14	status	4612 non-null	int64

Image 4

Above Image shows that we don't have any missing values in the data set.

Above Image shows we have total 15 columns and 4612 rows in our data set.

There are 4 integer columns and 10 object type columns and 1 float value type columns

Our target variable is of integer type

Number of duplicate rows = 0

We have no Duplicate values in our data set

Exploratory Data Analysis (EDA)

Unilateral data

Observation by age

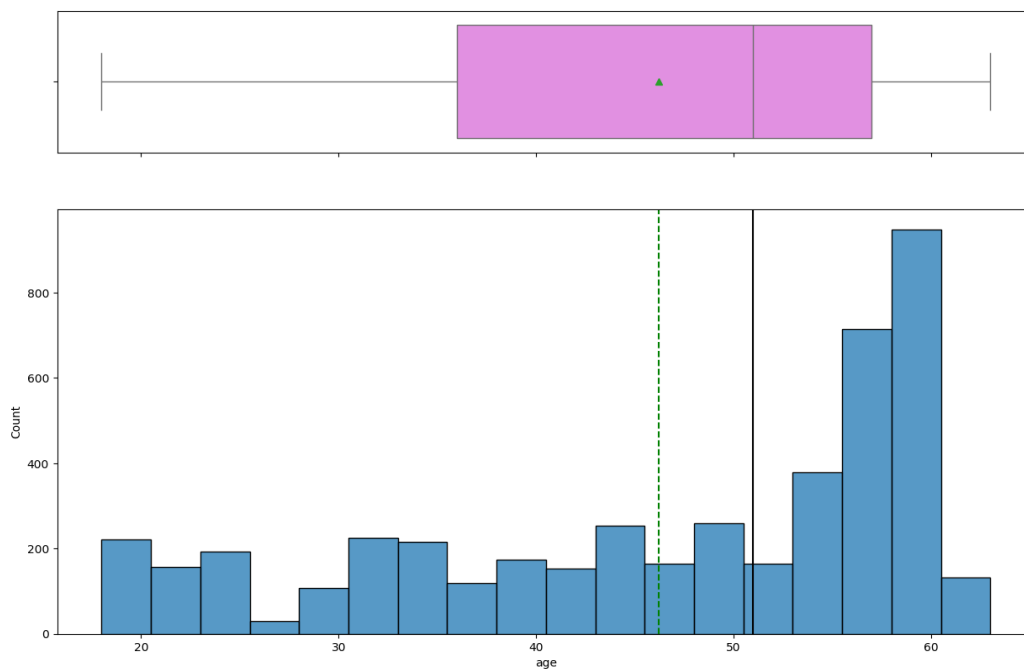


Image 5

Above image shows that the age distribution is left skewed distribution.

The mean value of age is around 51 years.

Observations on website visits

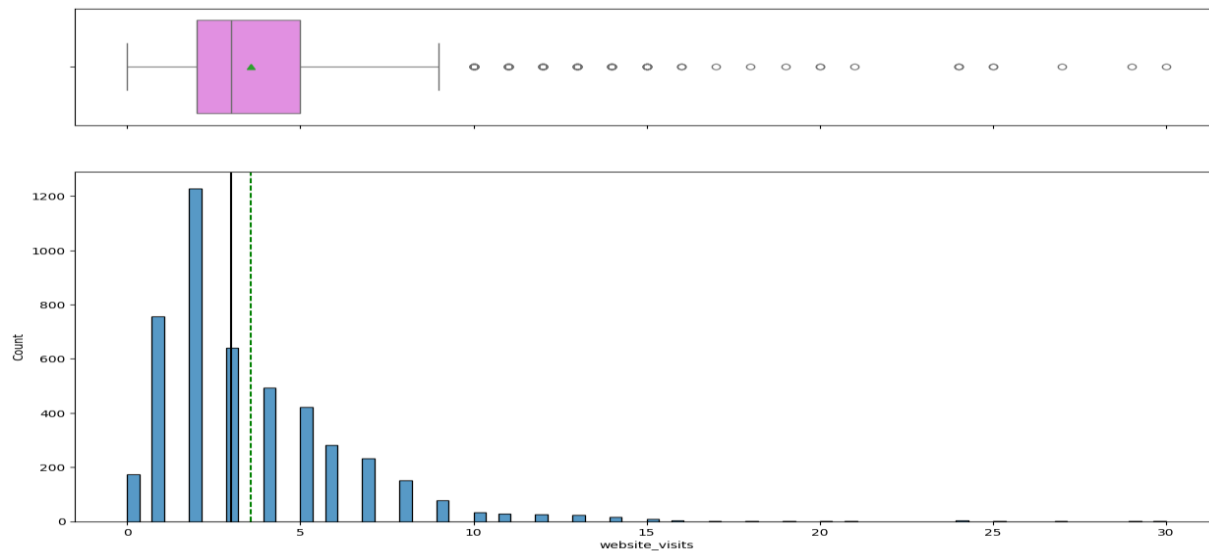


Image 6

Above Image shows that the data for website visit is right skewed.

The mean value of website visit is around 3.

Observations on number of time spent on website

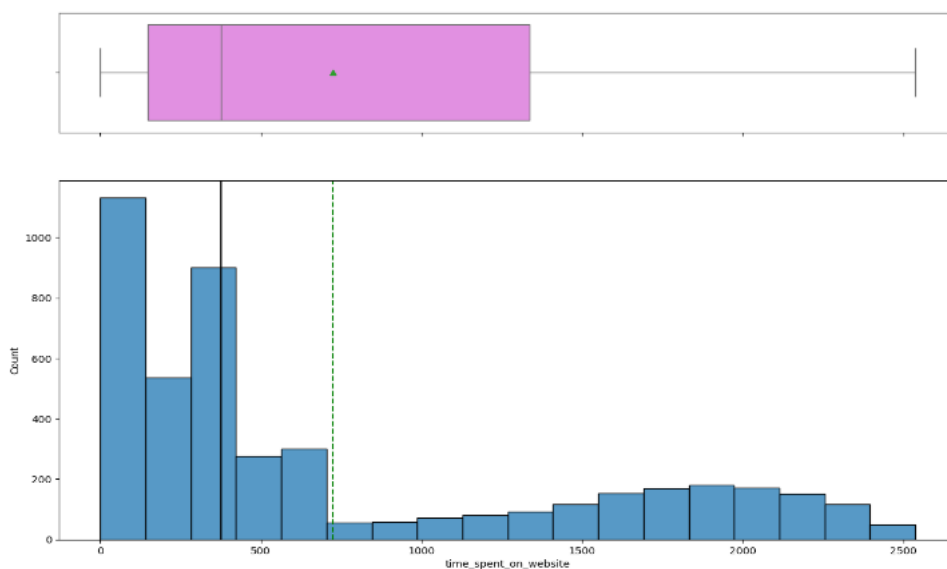


Image 7

Above Image shows that the data for time spend is right skewed.

Image shows the time spend on the web site.

Observations on number of page views per visit

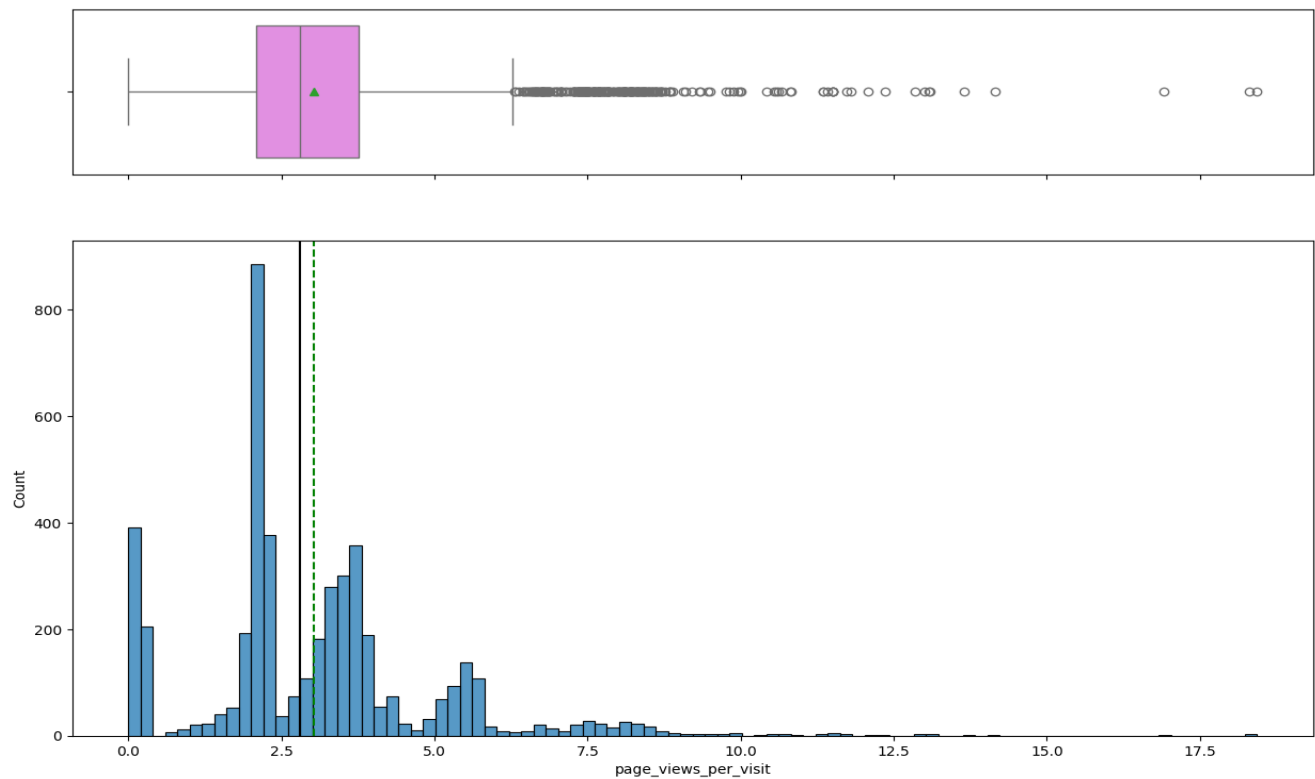


Image 8

Above Image shows that the data for pages views per visit is right skewed

Image shows on the number of pages views per visit by persons on the website.

Observations on number of adults

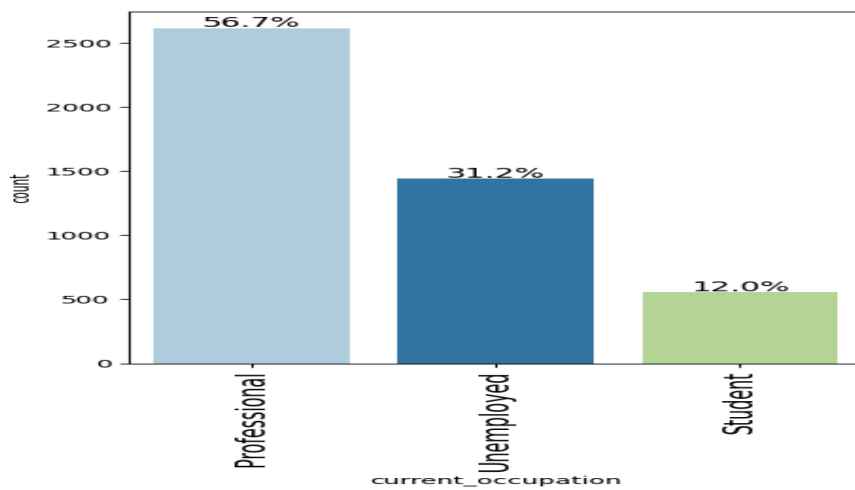


Image 9

Above Image shows the occupation of the lead as mostly professionals are visiting the site.

Observations on number of children

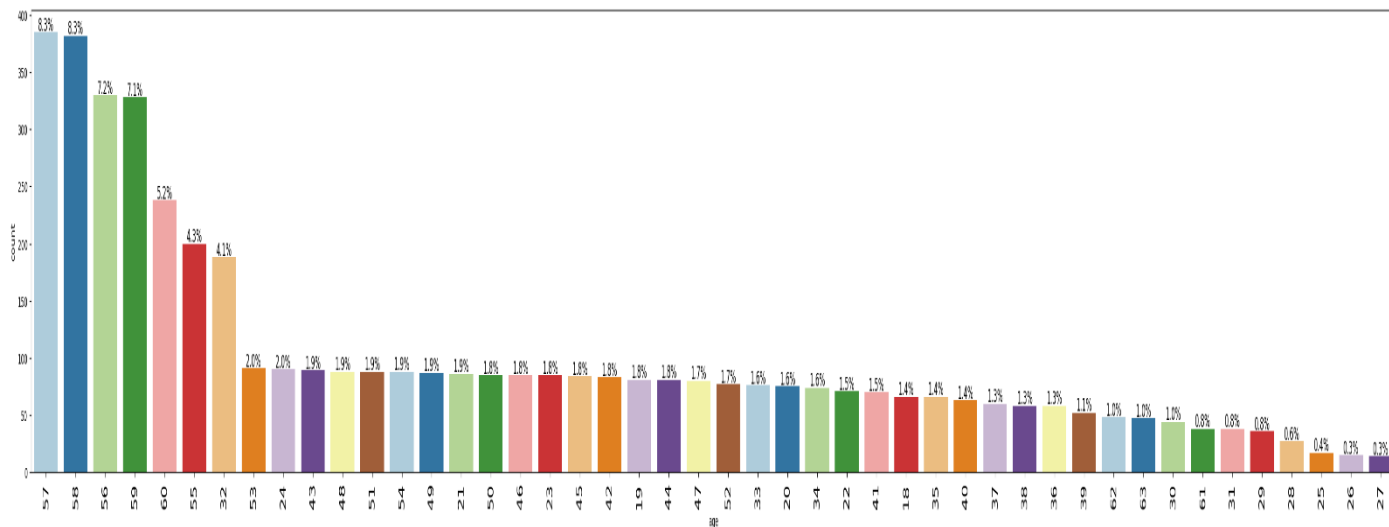


Image 10

Above image shows mostly aged people as around 8.3% or are of age 57 and 58 years, followed by 7.2% for age 56 years, least with 0.3% for 26 and 27 years.

Observations on profile completed

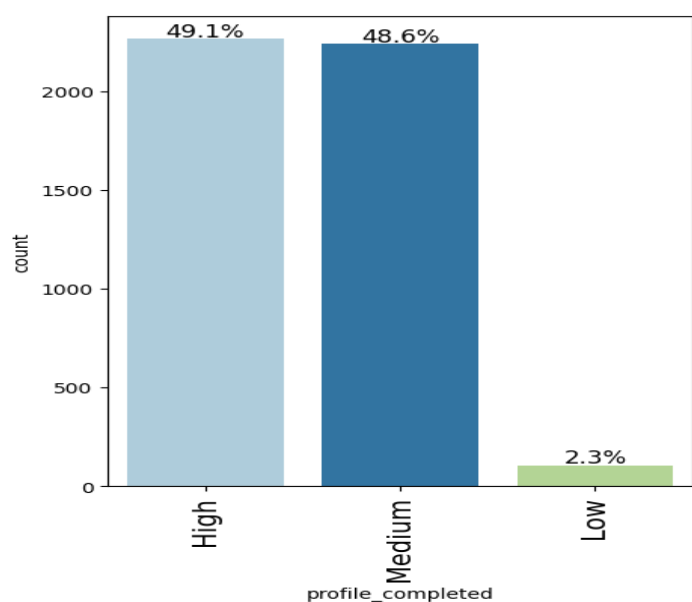


Image 11

Image shows only around half of the persons visited website have completed their profile as it is around 49.10%.

Observations on last activity

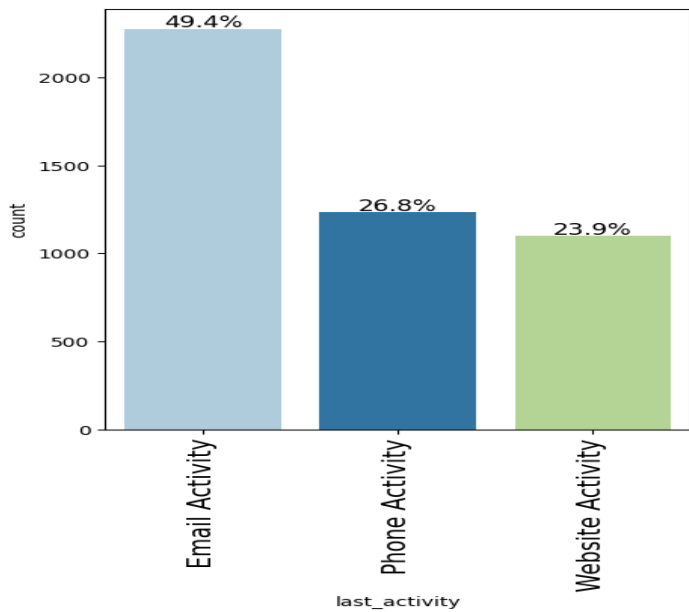


Image 12

Above image shows mostly interaction between the lead and Extraalearn is through Email (49.4%) followed by the phone (26.8%) and website (23.9%).

Observations on print_media_type1

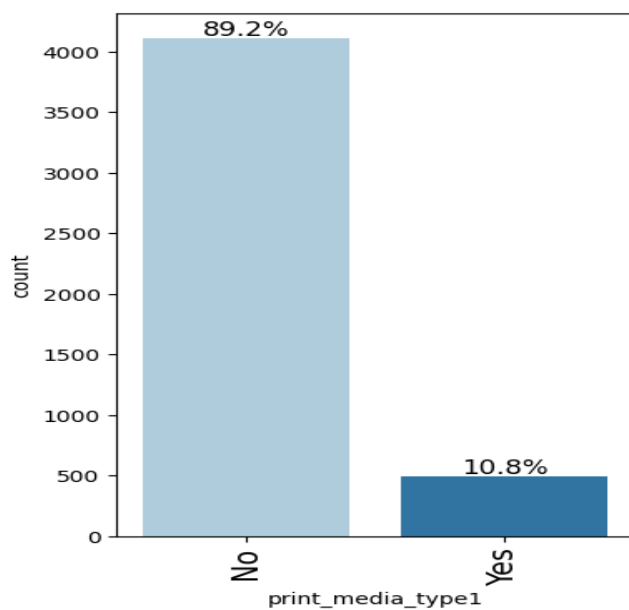


Image 13

Image shows around 11% of the lead has seen ad in the Newspaper.

Observations on print media type2

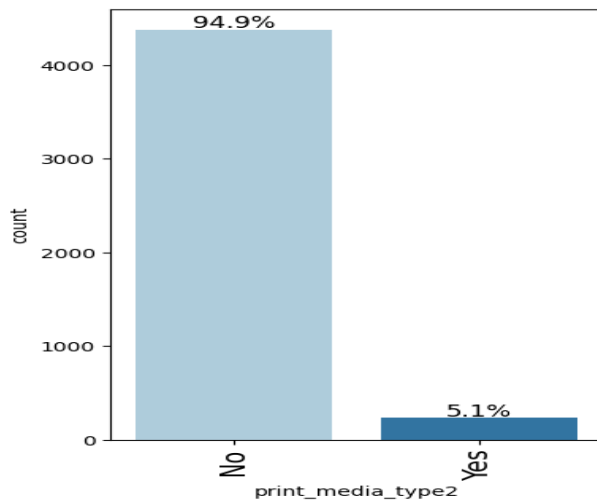


Image 14

Image shows around 5% of the lead has seen ad in the Magazine.

Observations on room type reserved

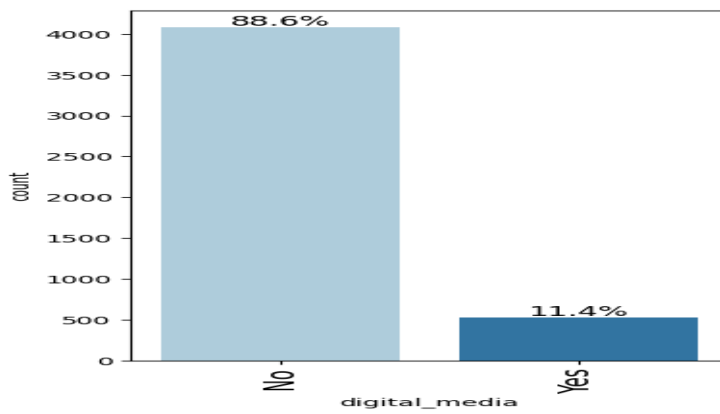


Image 15

Image shows around 11.5% of the lead has seen ad in the digital media.

Observations on educational channels

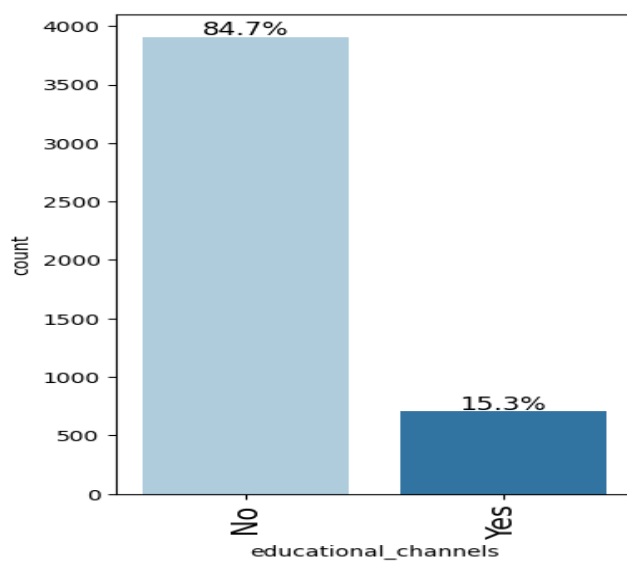


Image 16

Image shows around 15% of the lead have heard about the Extraalearn on Educational channels or on educational website.

Observations on referral

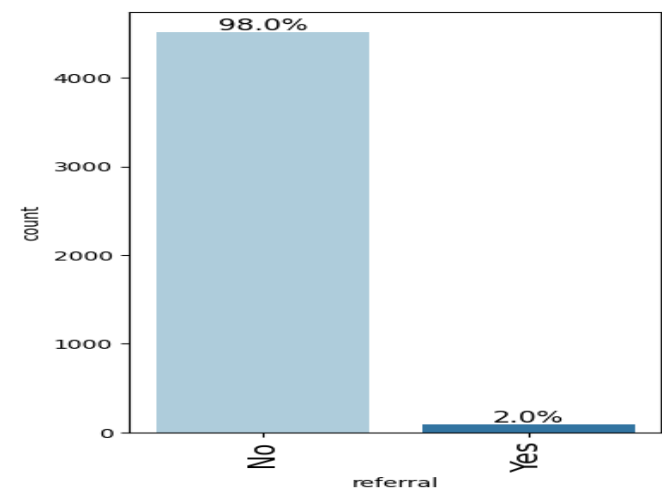


Image 17

Image shows around only 2% of the lead have heard about the Extraalearn through references.

Observations on status

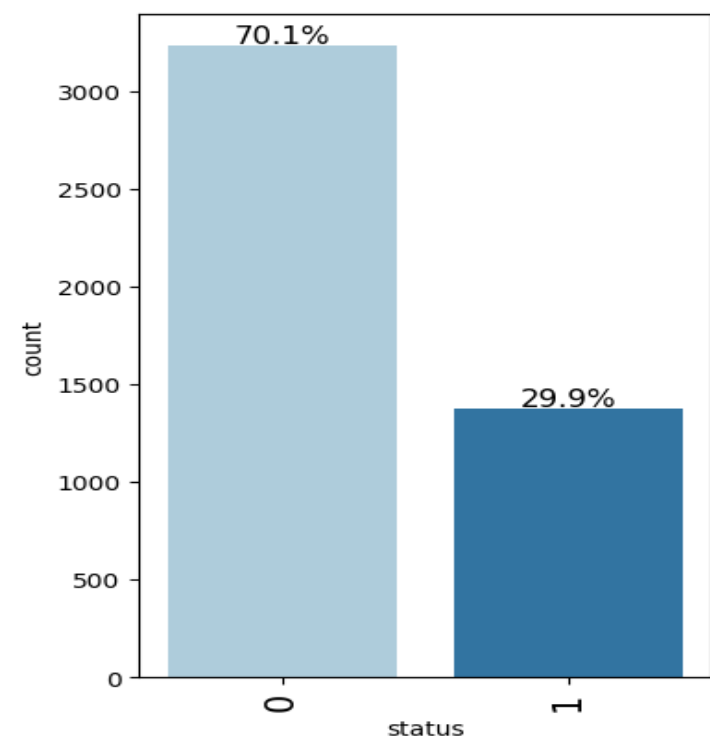


Image 18

Image shows only around 3/10 of the lead was converted to paid customers.

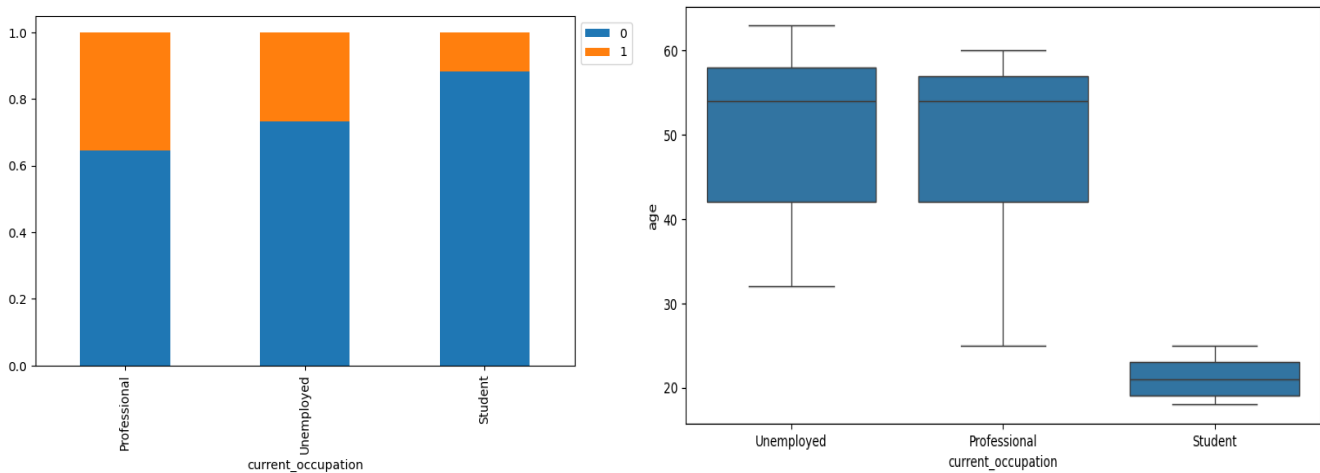
Bivariate Analysis



Image 19

Above image shows there is not a very strong co relation between the given variable.

1. 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.



	count	mean	std	min	25%	50%	75%	max
current_occupation								
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

Image 20

Above image shows us that mostly the Professionals have the tendency to become a paid customer followed by the unemployed and then the students.

Above image explains the age distribution, and the time they spend on the web.

2. 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?

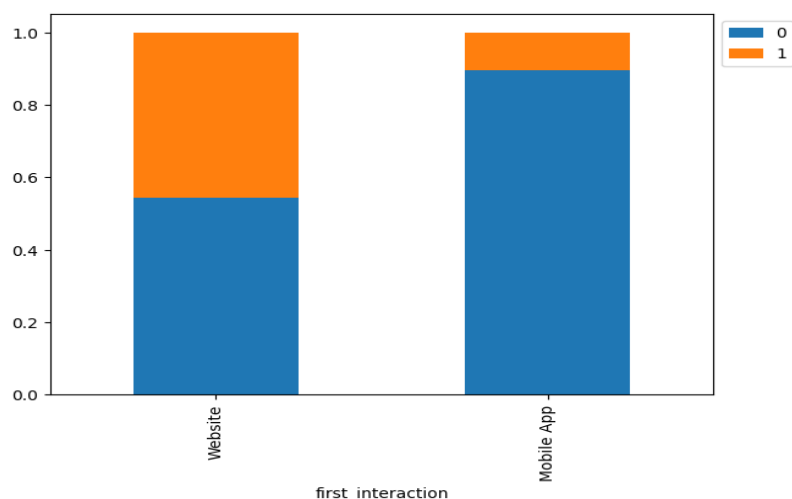


Image 21

Above image shows mostly the Leads interacted with the help of website are most likely impacting the leads.

The above image shows the relationship between the first interaction and the conversion rates for the leads who uses website and the mobile app.

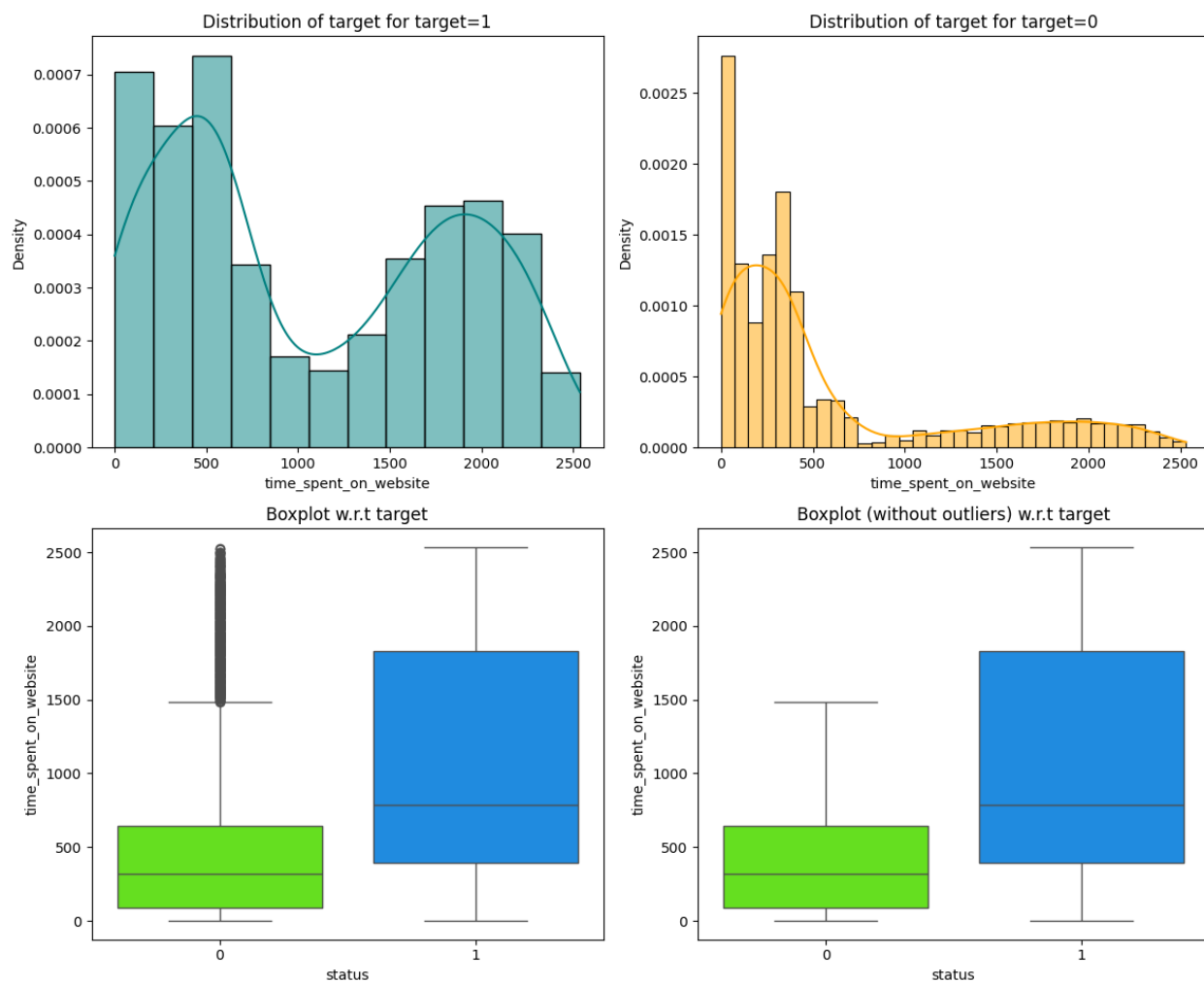


Image 22

Above image shows the relationship between the lead spend time on the website and their conversion to become the paid subscriber.

As lead spend time on the website the higher the more likely they will subscribe the paid subscription of the web.

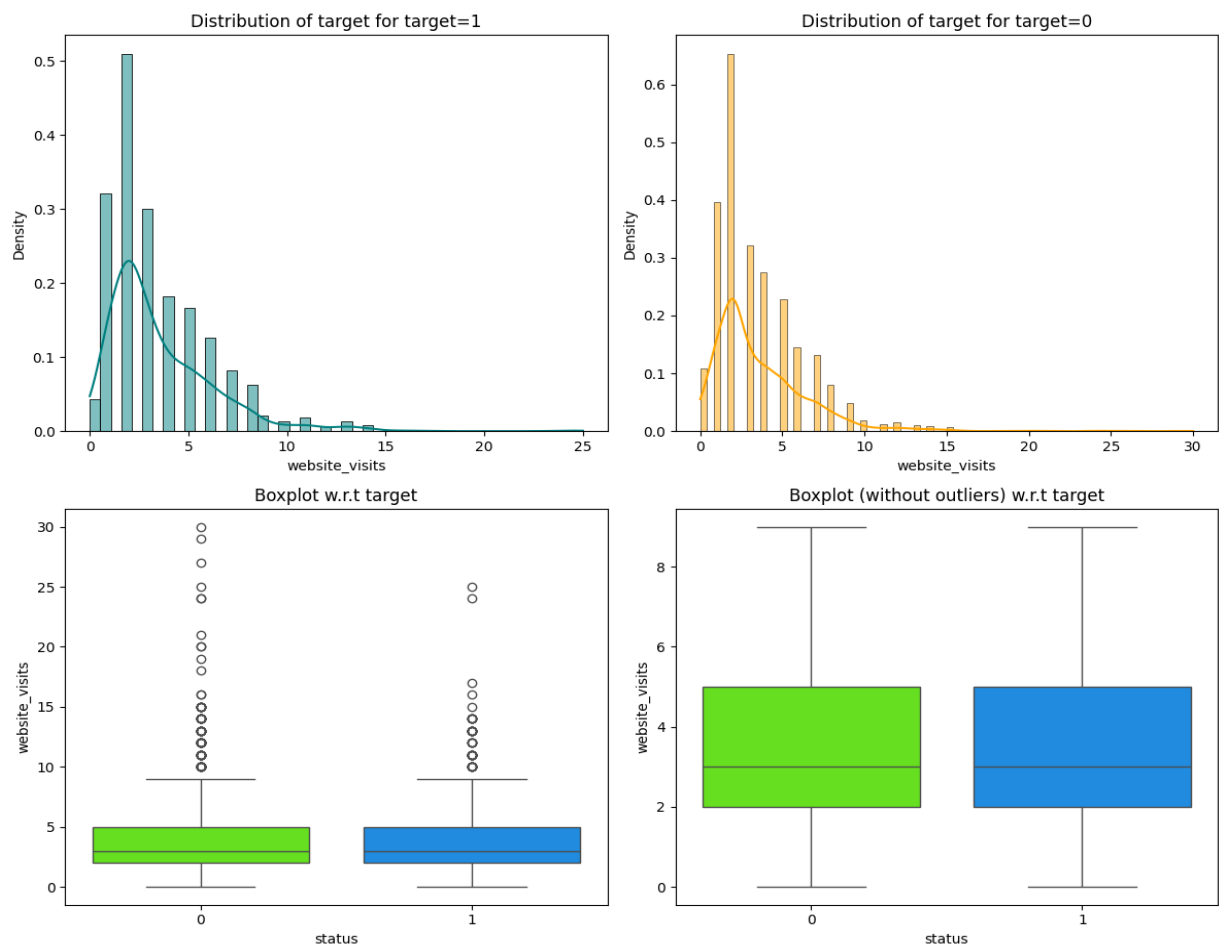
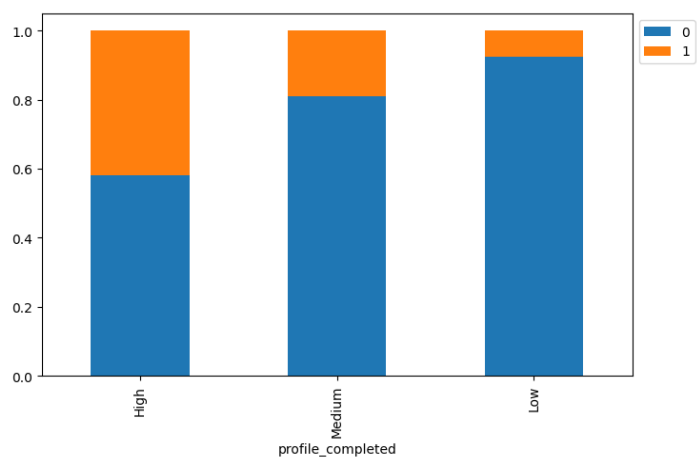


Image 23

Above image shows the relationship between the lead visited on the website and their conversion to become the paid subscriber.

As lead visited website more frequently the higher the more likely they will subscribe the paid subscription of the web.

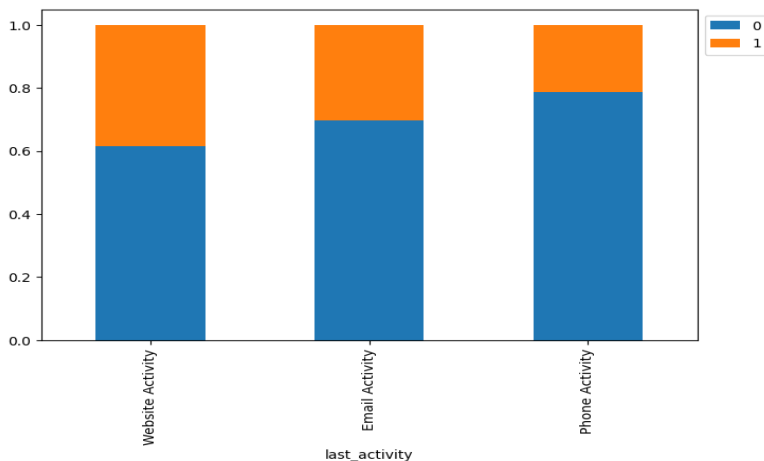


status	0	1	All
profile_completed			
All	3235	1377	4612
High	1318	946	2264
Medium	1818	423	2241
Low	99	8	107

Image 24

Above image shows the percentage of profile filled by the lead on mobile or on the website.

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



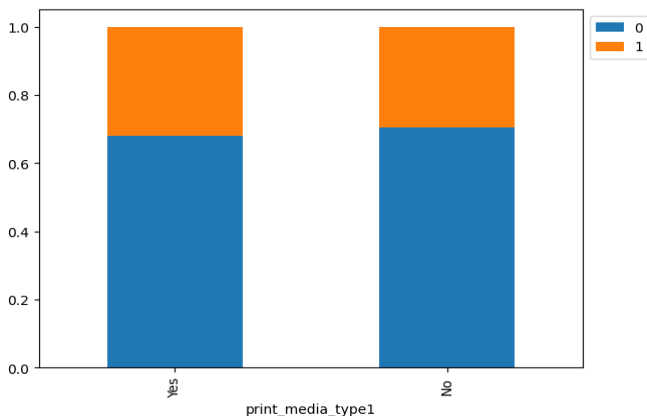
status	0	1	All
last_activity			
All	3235	1377	4612
Email Activity	1587	691	2278
Website Activity	677	423	1100
Phone Activity	971	263	1234

Image 25

Above image shows the leads who seeks details about programme as we found more likely the leads visited by email are more then the others like website and phone activity.

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

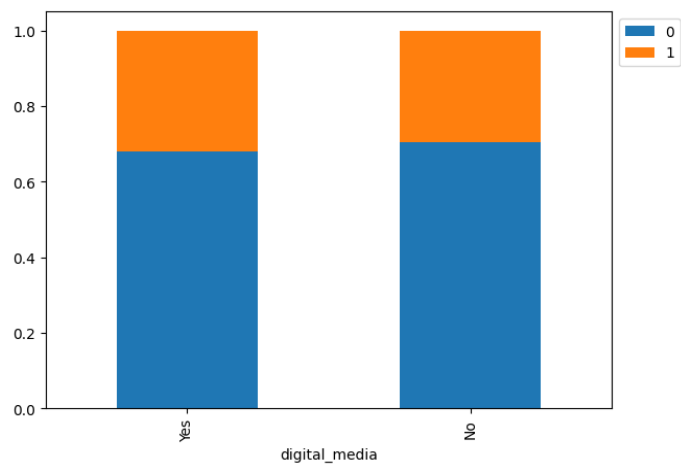
The leads they get from the educational channels (Digital Media) has the high conversion rate



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

Above image shows the conversion rate of leads who see the ad on Newspaper.

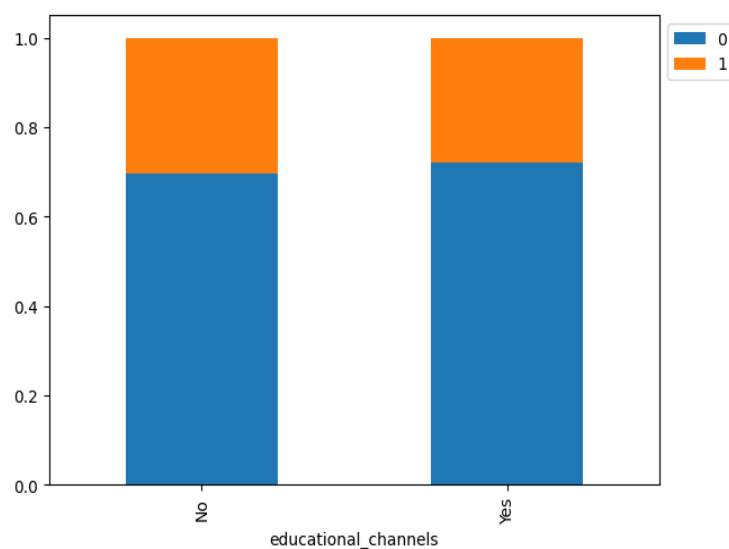
Image 26



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

Image 27

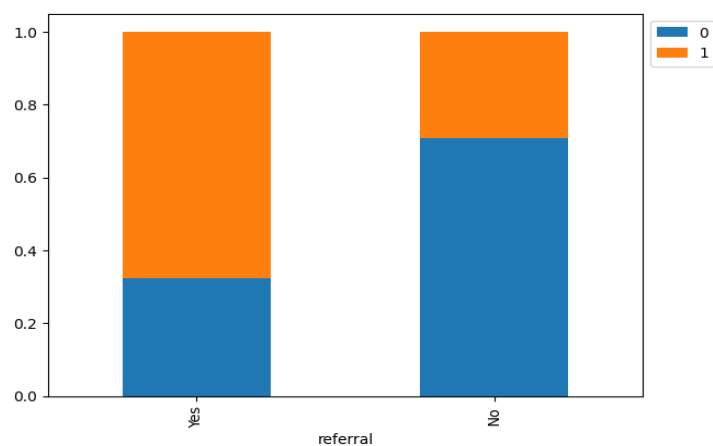
Above image shows the conversion rate of leads who see the ad on Magazine.



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

Image 28

Above image shows the conversion rate of leads who heard about the Extraalearn on education channels.

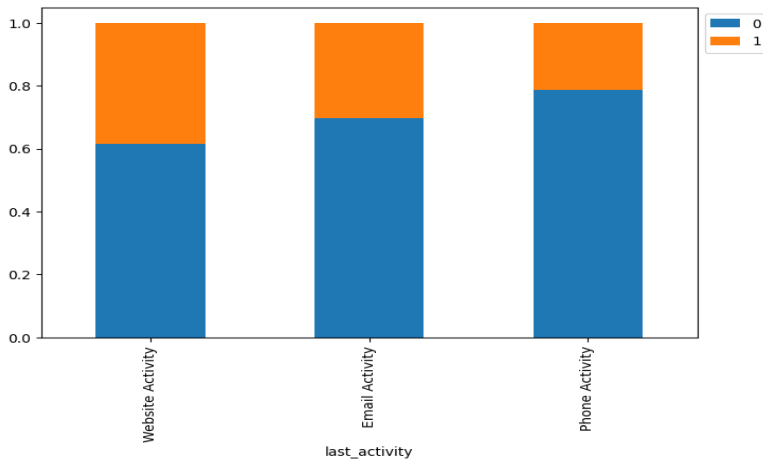


status	0	1	All
referral			
All	3235	1377	4612
No	3205	1314	4519
Yes	30	63	93

Image 29

Above Image shows the status of the leads who learn about the Extraalearn through the referral's.

5 People browsing the website or 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 increase the chances of conversion?



Above image shows the leads uses the website have more likely to become a paid customer.

Data Preprocessing

Outlier Check

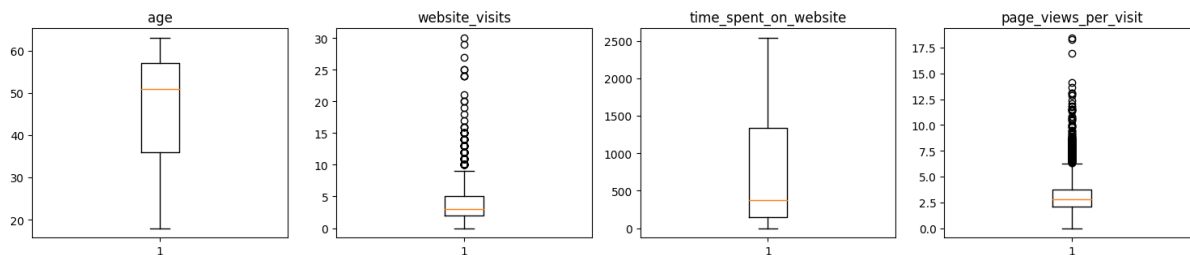


Image 30

This image shows the outliers present in our data set.

Data Preparation for modelling

```
Shape of Training set : (3459, 16)
Shape of test set : (1153, 16)
Shape of Training set : (3459,)
Shape of test set : (1153,)
Percentage of classes in training set:
status
0    0.70107
1    0.29893
Name: proportion, dtype: float64
Percentage of classes in test set:
status
0    0.70252
1    0.29748
```

Image 31

Above image shows we have taken 3459 rows for training set and 1153 rows for test set.

Model Building

Logistic Regression (with Statsmodel)

Dep. Variable:	status	No. Observations:	3459
Model:	Logit	Df Residuals:	3442
Method:	MLE	Df Model:	16
Date:	Thu, 08 Aug 2024	Pseudo R-squ.:	0.3553
Time:	21:09:28	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

Image 32

Above image shows the performance of our model we have built with data available.

Checking Logistic Regression model performance on training set

	Accuracy	Recall	Precision	F1
0	0.82249	0.65377	0.72532	0.68769

Image 33

Above image shows that our recall matrix for Logistic Regression which is around 65% so it is not underfitting.

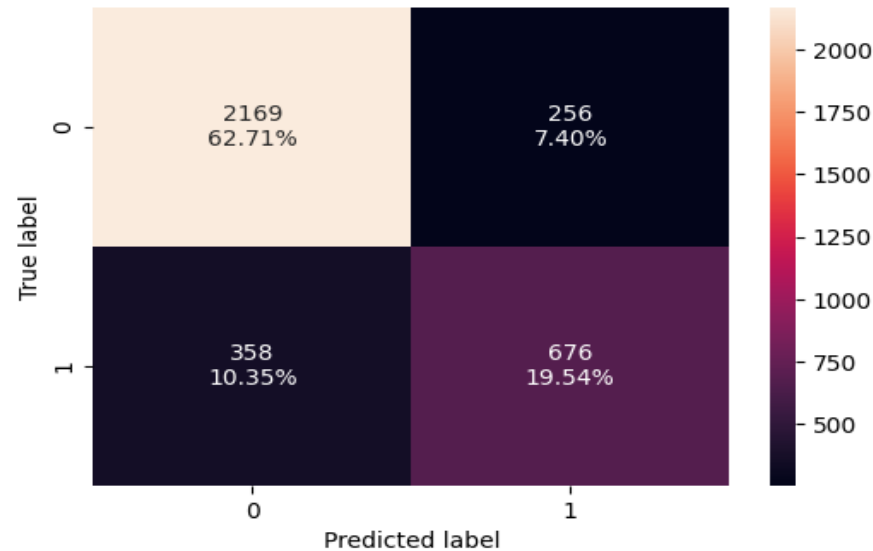


Image 34

Confusion matrix on training set by Logistic regression.

Checking Logistic Regression model performance on test set

	Accuracy	Recall	Precision	F1
0	0.80399	0.61224	0.69307	0.65015

Image 35

The above image shows Recall value for the test matrix is 61% is compatible with the training data shows the model is not underfitting for Logistic Regression.

But we need improvement in the model performance.

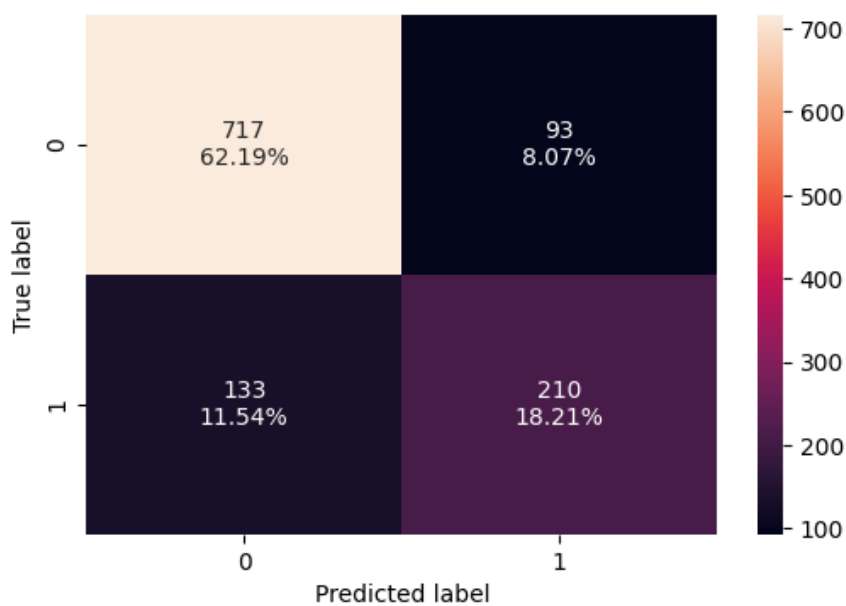


Image 36

Confusion matrix on test set by Logistic regression.

Naïve - Bayes Classifier

Checking Naive - Bayes Classifier performance on training set

	Accuracy	Recall	Precision	F1
0	0.78925	0.76886	0.61868	0.68564

Image 37

Above image shows our recall value is improved now with the Naïve bayes classifier as it increases and its 76% with the training data.

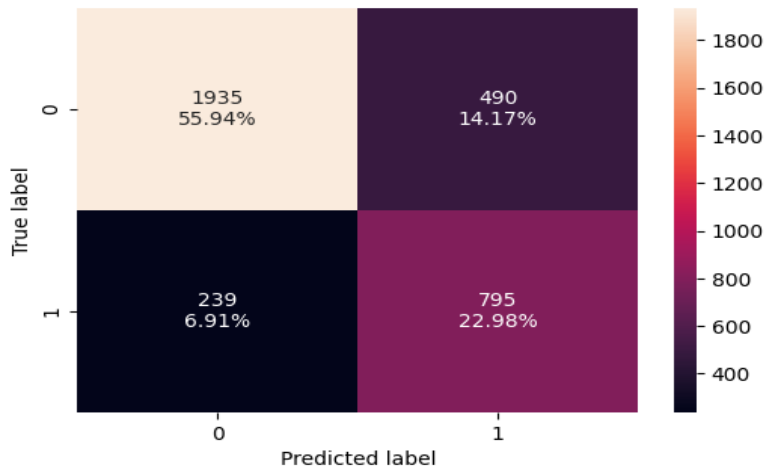


Image 38

Confusion matrix on training set by Naive Bayes Classifier.

Checking Naive - Bayes Classifier performance on test set

	Accuracy	Recall	Precision	F1
0	0.77971	0.75802	0.60325	0.67183

Image 39

Above image shows our recall score for the test data has increased to 75% and it is compatible with the training data so our model is not underfitting for Naïve Bayes Classifier.

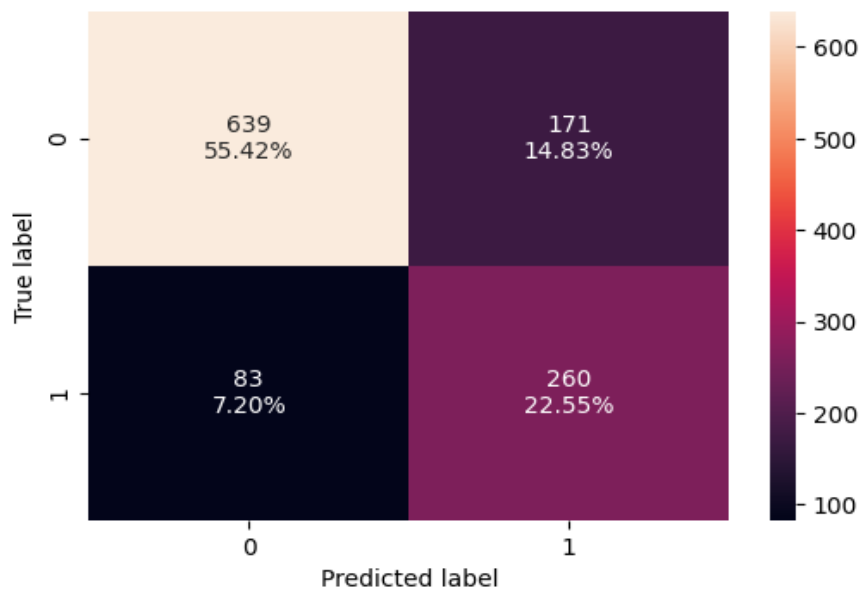


Image 40

Confusion matrix on test set by Naive Bayes Classifier.

KNN Classifier (K = 3)

Checking KNN Classifier performance on training set

	Accuracy	Recall	Precision	F1
0	0.88927	0.77466	0.84227	0.80705

Image 41

Above image shows our recall value has increase up to 77% now and Precision is around 84%, and F1 score is around 81% for the training set for KNN Classifier.

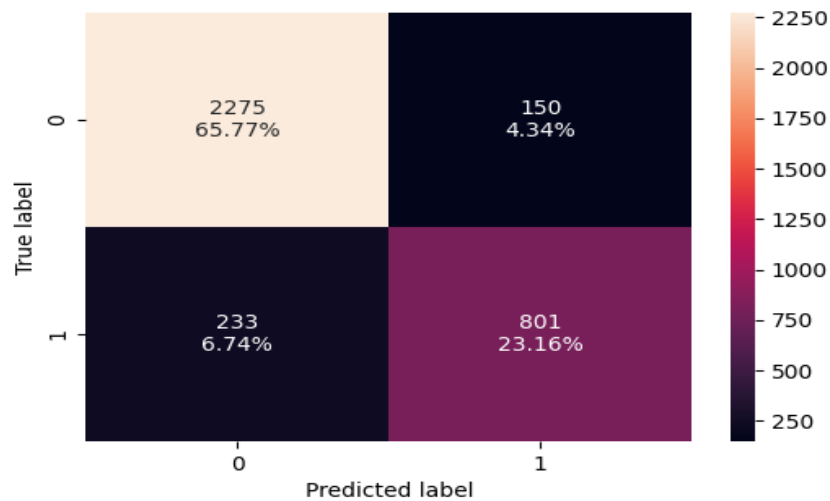


Image 42

Confusion matrix on training set by KNN Classifier.

Checking KNN Classifier performance on test set

	Accuracy	Recall	Precision	F1
0	0.88927	0.77466	0.84227	0.80705

Image 43

Above image shows our recall value has increase upto 77% now and Precision is around 84%, and F1 score is around 81% for the test set for KNN Classifier.

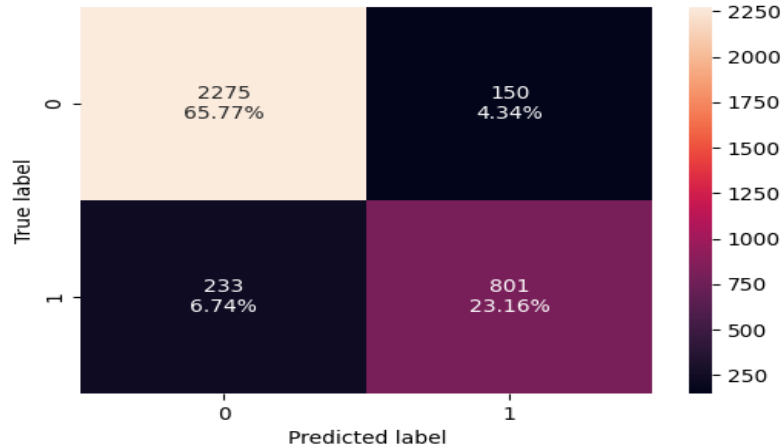


Image 44

Confusion matrix on test set by KNN Classifier.

Decision Tree Classifier

Checking Decision Tree Classifier performance on training set

	Accuracy	Recall	Precision	F1
0	0.76670	0.70406	0.59235	0.64339

Image 45

Above image shows our recall value is up to 76% now and Precision is around 59%, and F1 score is around 64% for the training set for Decision Tree Classifier.

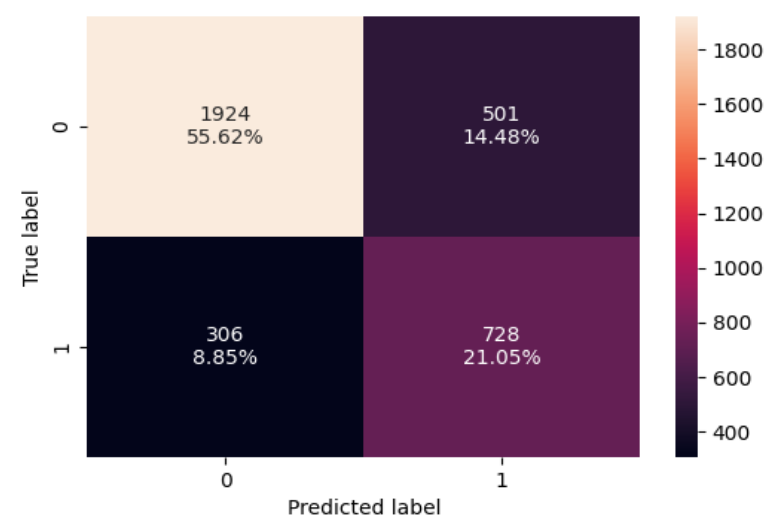


Image 46

Confusion matrix on training set by Decision Tree Classifier.

Checking Decision Tree Classifier performance on test set

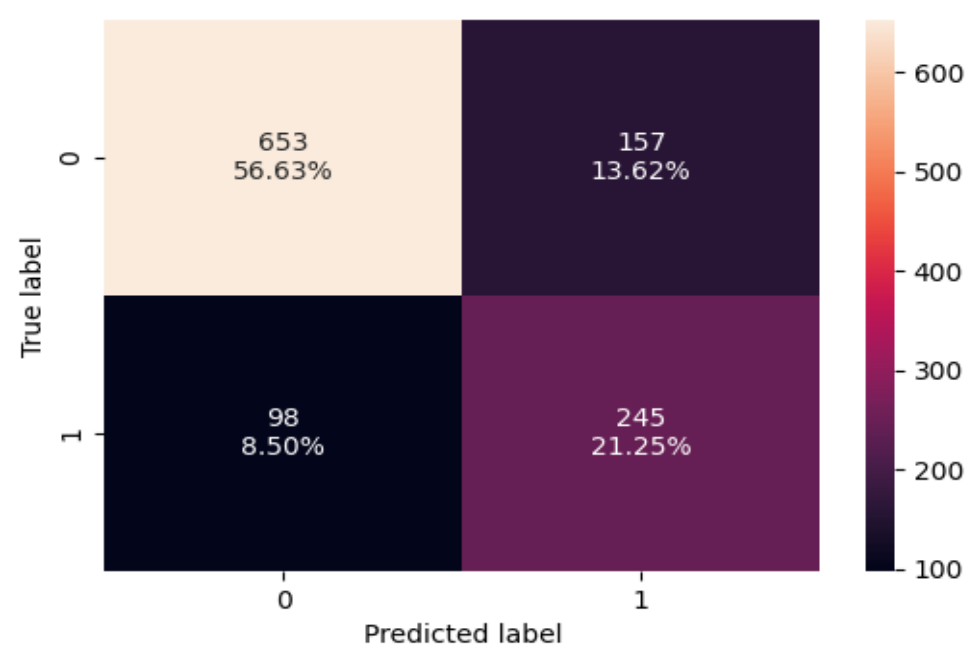


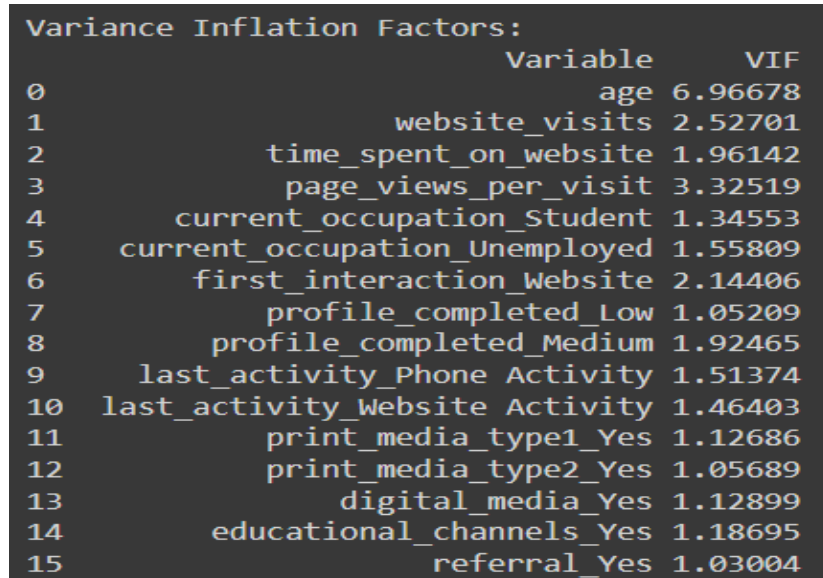
Image 47

Confusion matrix on test set by Decision Tree Classifier.

Model Performance Improvement

Logistic Regression (deal with multicollinearity, remove high p-value variables, determine optimal threshold using ROC curve)

Logistic Regression - Dealing with Multicollinearity



Variance Inflation Factors:		
	Variable	VIF
0	age	6.96678
1	website_visits	2.52701
2	time_spent_on_website	1.96142
3	page_views_per_visit	3.32519
4	current_occupation_Student	1.34553
5	current_occupation_Unemployed	1.55809
6	first_interaction_Website	2.14406
7	profile_completed_Low	1.05209
8	profile_completed_Medium	1.92465
9	last_activity_Phone Activity	1.51374
10	last_activity_Website Activity	1.46403
11	print_media_type1_Yes	1.12686
12	print_media_type2_Yes	1.05689
13	digital_media_Yes	1.12899
14	educational_channels_Yes	1.18695
15	referral_Yes	1.03004

Image 48

Above Image shows we have deal with the Logistic Regression from Multicollinearity.

Dealing with high p-value variables

```

Optimization terminated successfully.
  Current function value: 0.393259
  Iterations 7
Dropping column website_visits with p-value: 0.6479479510695756
Optimization terminated successfully.
  Current function value: 0.393289
  Iterations 7
Dropping column digital_media_Yes with p-value: 0.5827477567779233
Optimization terminated successfully.
  Current function value: 0.393332
  Iterations 7
Dropping column educational_channels_Yes with p-value: 0.5567480874402153
Optimization terminated successfully.
  Current function value: 0.393382
  Iterations 7
Dropping column print_media_type2_Yes with p-value: 0.3388431424365996
Optimization terminated successfully.
  Current function value: 0.393513
  Iterations 7
Dropping column page_views_per_visit with p-value: 0.32588281893475446
Optimization terminated successfully.
  Current function value: 0.393653
  Iterations 7
Dropping column age with p-value: 0.28978634672139825
Optimization terminated successfully.
  Current function value: 0.393816
  Iterations 7
Dropping column print_media_type1_Yes with p-value: 0.1535692252215647
Optimization terminated successfully.
  Current function value: 0.394108
  Iterations 7
Dropping column referral_Yes with p-value: 8.577120192655911e-05

```

Image 49

Above Image shows we have deal with the high p-value variables.

Training the Logistic Regression model again with only the significant features

```

Optimization terminated successfully.
  Current function value: 0.394108
  Iterations 7

      Logit Regression Results
=====
Dep. Variable:      status    No. Observations:      3459
Model:              Logit    Df Residuals:        3449
Method:             MLE      Df Model:            9
Date:              Thu, 08 Aug 2024    Pseudo R-squ.:      0.3539
Time:              21:24:12    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
=====

```

Image 50

This is our final model as we have only significant features available in this model.

Determining optimal threshold using ROC Curve

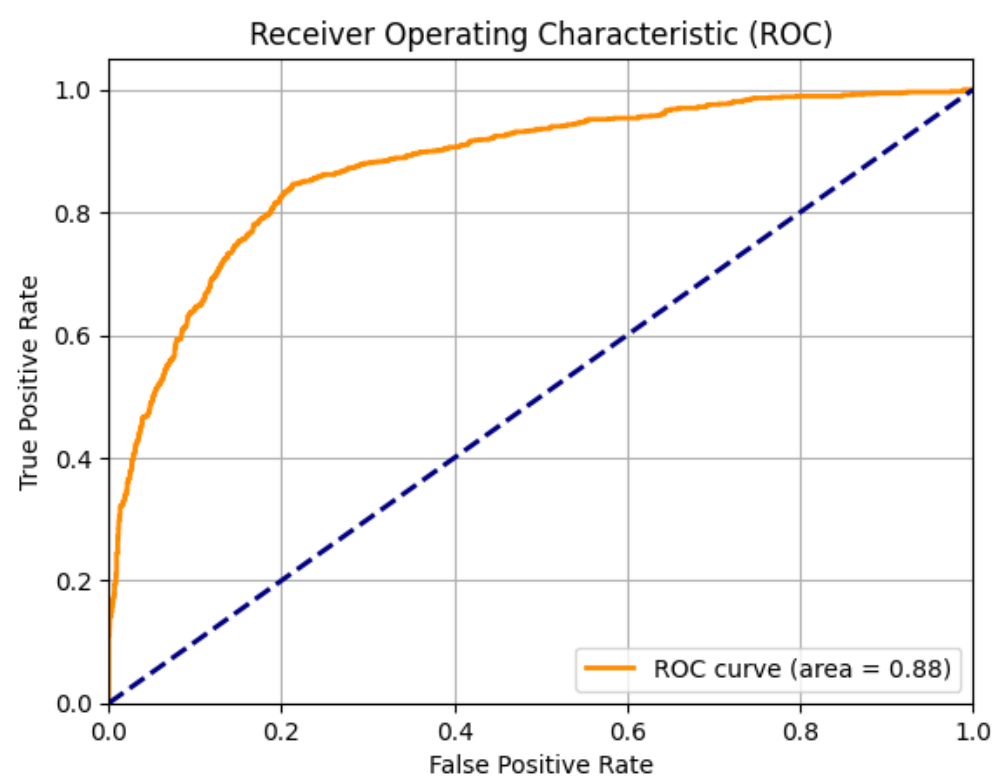


Image 52

Above Image shows the ROC curve which is tiled around the curve.

Checking tuned Logistic Regression model performance on training set

	Accuracy	Recall	Precision	F1
0	0.80457	0.84623	0.62859	0.72135

Image

Image 53

Above image shows our recall value has increase upto 84% now and Precision is around 62%, and F1 score is around 72% for the training set for tuned Logistic Regression.

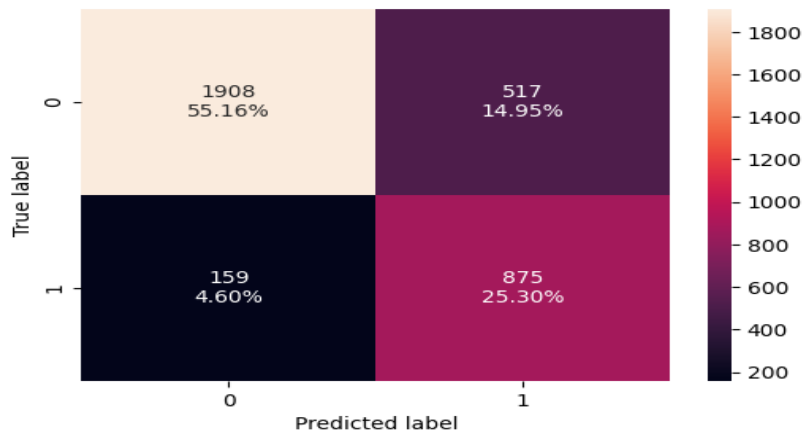


Image 54

Confusion matrix on training set for tuned Logistic Regression.

Checking tuned Logistic Regression model performance on test set

	Accuracy	Recall	Precision	F1
0	0.79098	0.83090	0.60897	0.70284

Image 55

Above image shows our recall value has increase upto 83% now and Precision is around 60%, and F1 score is around 70% for the test set for tuned Logistic Regression.

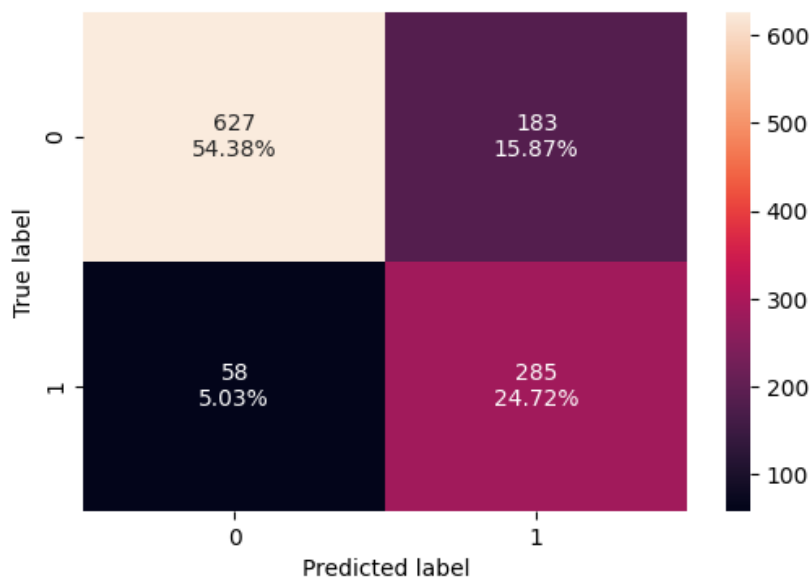


Image 56

Confusion matrix on test set for tuned Logistic Regression.

KNN Classifier (different values of K)

KNN Classifier Performance Improvement using different k values

```

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

```

Image 57

Above Image shows the different values with the different K values.

Checking tuned KNN model performance on training set

	Accuracy	Recall	Precision	F1
0	0.85256	0.68182	0.79571	0.73437

Image 58

Above image shows our recall value is 68% now and Precision is around 79%, and F1 score is around 73% for the training set for tuned KNN Model.

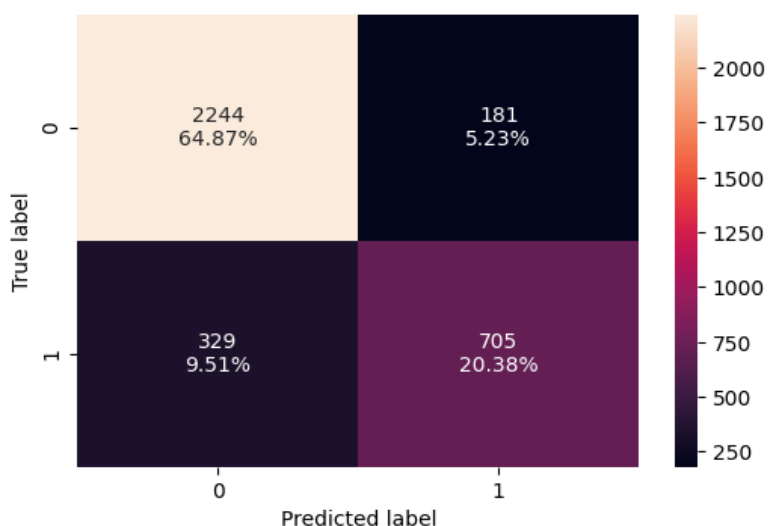


Image 59

Confusion matrix on training set for tuned KNN Model.

Checking tuned KNN model performance on test set

	Accuracy	Recall	Precision	F1
0	0.83695	0.65889	0.76094	0.70625

Image 60

Above image shows our recall value is 65% now and Precision is around 76%, and F1 score is around 70% for the test set for tuned KNN Model.

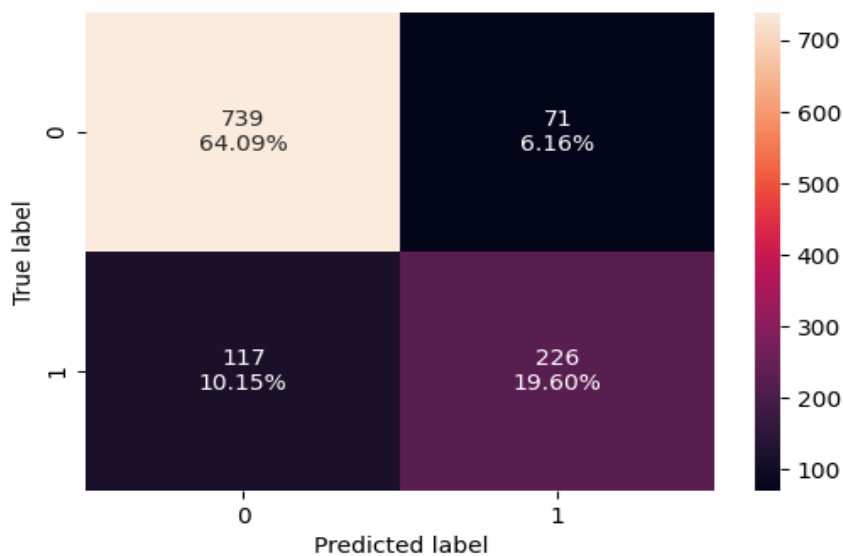


Image 61

Confusion matrix on test set for tuned KNN Model.

Decision Tree Classifier (pre-pruning)

Checking tuned Decision Tree Classifier performance on training set

	Accuracy	Recall	Precision	F1
0	0.77566	0.61315	0.62772	0.62035

Image 62

Above image shows our recall value is 61% now and Precision is around 63%, and F1 score is around 62% for the training set for tuned Decision Tree.

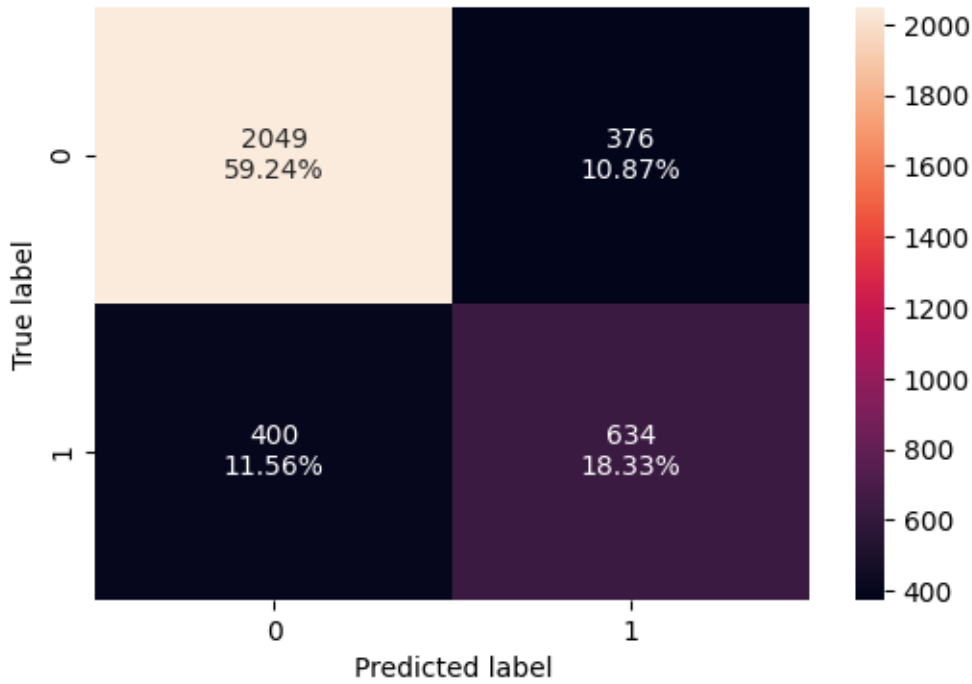


Image 63

[Confusion matrix on training set for tuned Decision Tree Classifier.](#)

Confusion matrix on training set for tuned odel.

Checking tuned Decision Tree Classifier performance on test set

	Accuracy	Recall	Precision	F1
0	0.79358	0.64431	0.65579	0.65000

Image 64

Above image shows our recall value is 64% now and Precision is around 65%, and F1 score is around 65% for the test set for tuned Decision Tree Classifier.

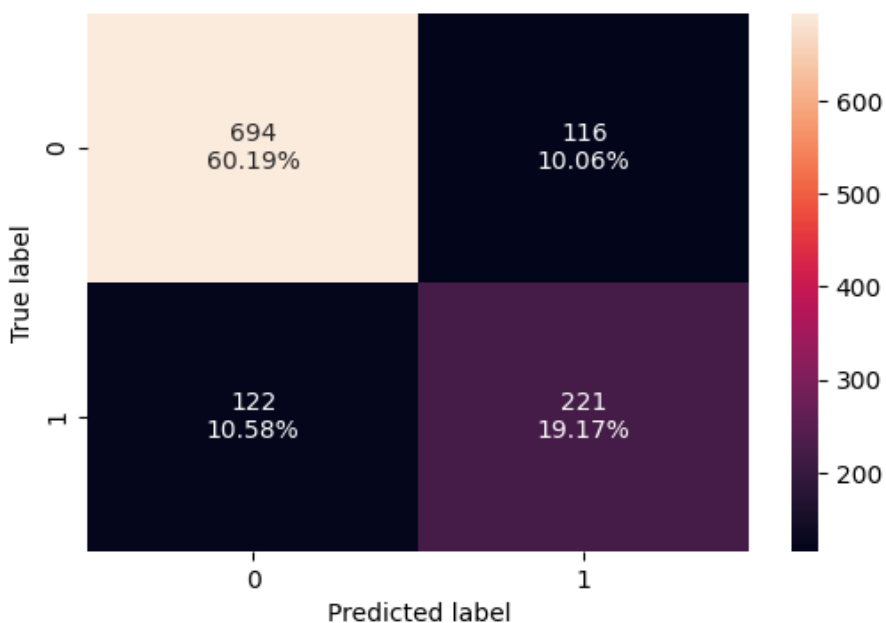


Image 65

Confusion matrix on test set for tuned Decision Tree Classifier.

Visualizing the Decision Tree

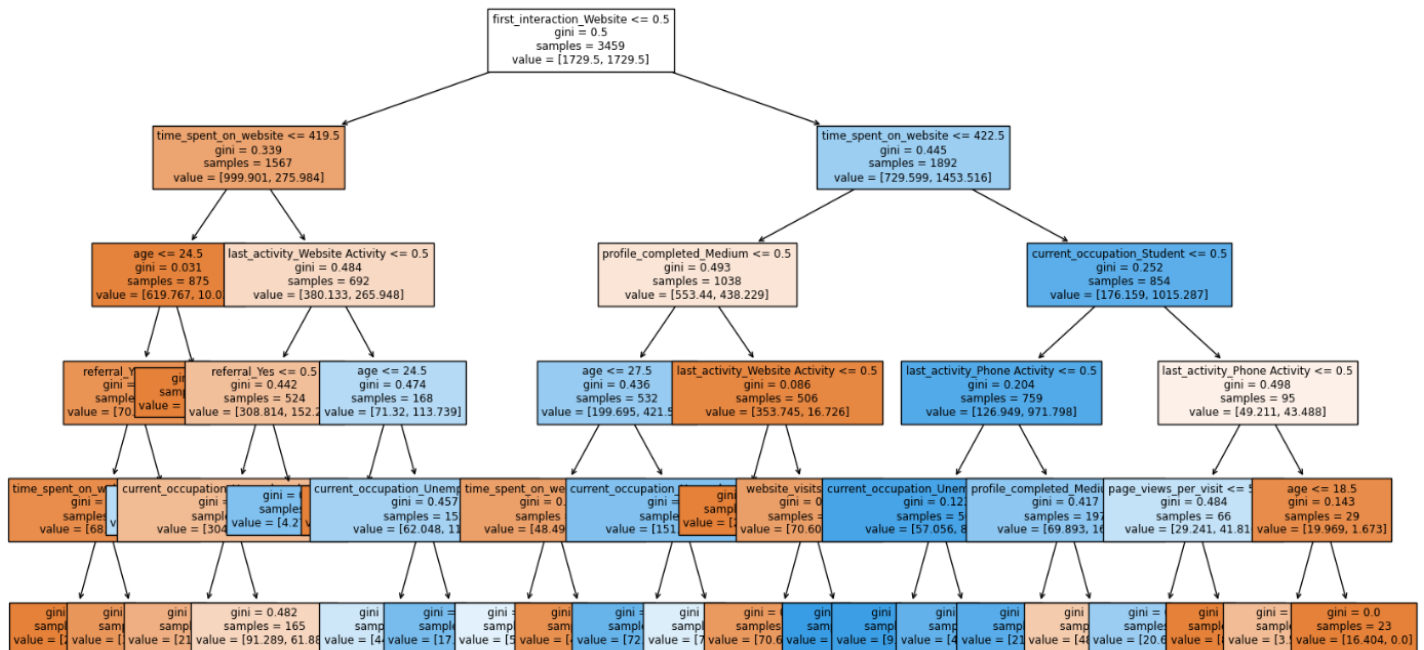


Image 66

Image showing the visualizing the Decision Tree.

Model Performance Comparison and Final Model Selection

	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	0.76670	0.77566
Recall	0.65377	0.84623	0.76886	0.77466	0.68182	0.70406	0.61315
Precision	0.72532	0.62859	0.61868	0.84227	0.79571	0.59235	0.62772
F1	0.68769	0.72135	0.68564	0.80705	0.73437	0.64339	0.62035

Image 67

This is our final model with all the matrix Logistic Regression Base, Logistic Regression Tuned, Naïve Bayes Base, KNN Base, KNN Tuned, Decision Tree Base, Decision Tree Tuned.

For Logistic Regression Base our Recall Value is around 82%, Precision is around 72%, F1 Score is around 68%.

For Logistic Regression Tuned our Recall Value is around 84%, Precision is around 62%, F1 Score is around 72%.

For Naive Bayes Base our Recall Value is around 76%, Precision is around 61%, F1 Score is around 68%.

For KNN Base our Recall Value is around 77%, Precision is around 84%, F1 Score is around 80%.

For KNN Tuned our Recall Value is around 68%, Precision is around 79%, F1 Score is around 73%.

For Decision Tree Base our Recall Value is around 70%, Precision is around 59%, F1 Score is around 64%.

For Decision Tree Tuned our Recall Value is around 61%, Precision is around 62%, F1 Score is around 62%.

Actionable Insights and Recommendations

The time spend on the website gives plays a key role in identifying if a lead will be converted to a paid customer or not.

We observe as the lead spend more time on the web are more likely to be converted to a paid customer.

The leads fill their profile have higher chance to become a paid customer.

From the data set we have found out that our main focus should be on the older leads as they have the more chances to be converted.

So, in our marketing ads we should keep this factor our main target age category is older once.

Company should invest more on developing the website as we found that more the leads spend their time on the web more likely they have chances to be converted.