ExtraaLearn EdTech Project

Potential Customers Prediction – Classification and Hypothesis Testing

Monday, Sept. 11, 2023

Presentation By: Ayomikun C. Adeniran

Contents / Agenda

- Business Problem Overview and Solution Approach
- Data Overview
- EDA Results - Univariate and Multivariate
- Data Preprocessing
- Model Performance Summary
- Conclusion and Recommendations

Business Problem Overview and Solution Approach

Objective

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

Business Problem Overview and Solution Approach

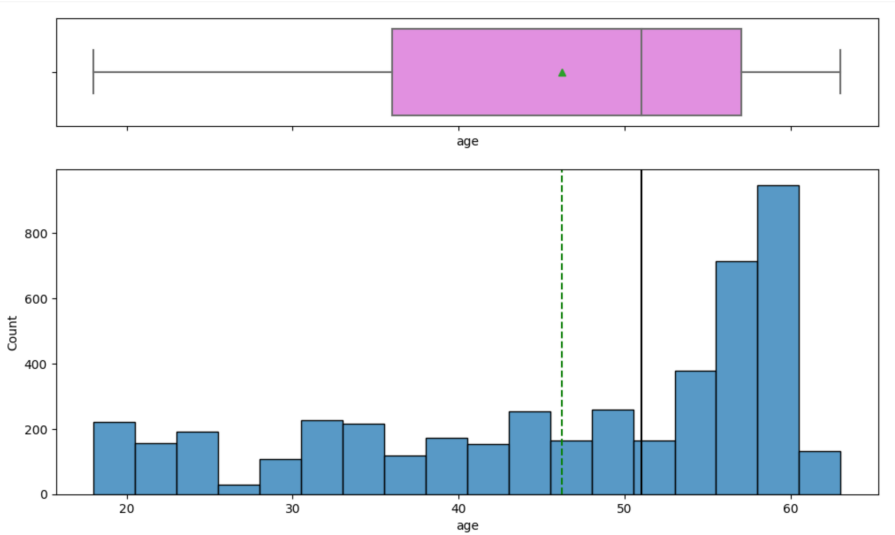
Solution Approach

- Exploratory Data Analysis (Univariate and Multivariate)
- Data Preprocesssing
- Build several ML classification models
- Analyze model performance to help identify which leads are more likely to convert to paid customers.
- Conclude and make business recommendations.

Data Overview

- The data has 4612 rows and 15 columns.
- The dataset has 5 numerical columns and 10 categorical columns. There are no missing values. There are no duplicated values.
- The average age of customers is 46 and the median age is 51. All customers are between the ages 18 and 63.
- On average, potential customers visit the website about 3-4 times.
- The average time spent on the website is roughly 12 minutes.
- The average website visitor views 3 pages per visit.

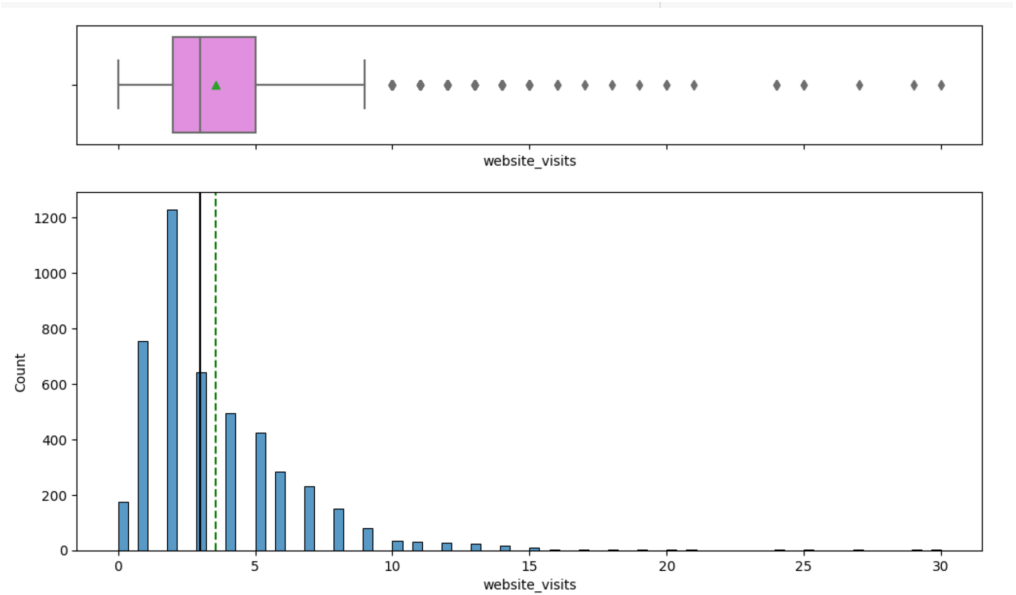
EDA Results



The age is skewed to the left. Most potential customers are in their mid to late 50s and early 60s. This makes sense as this age range likely either:

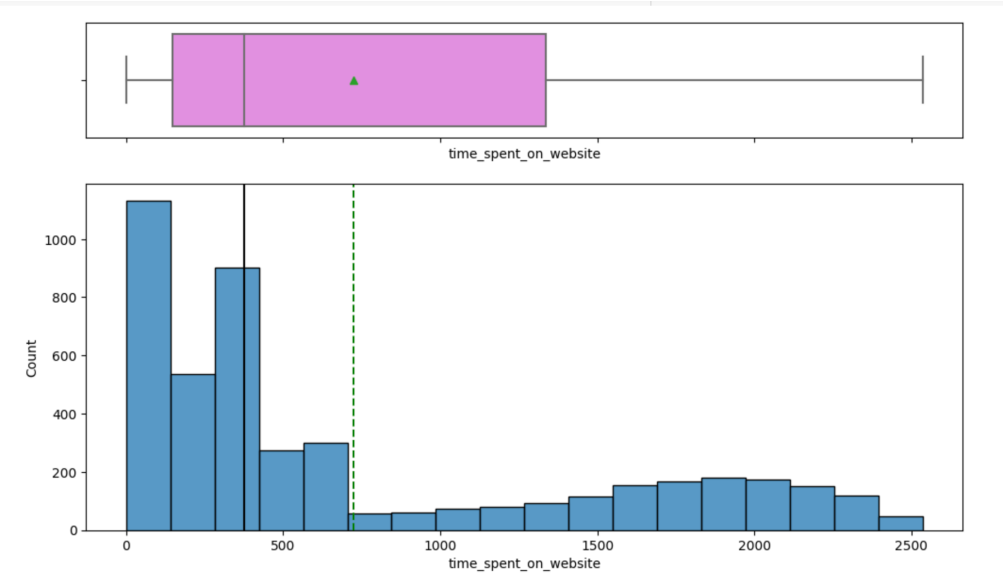
- have children who are in college and are purchasing this for them, and/or
- are themselves seeking to reskill or upskill (since it has been a while they got their education/skills and things have likely changed the most for this category of people in most sectors)

EDA Results



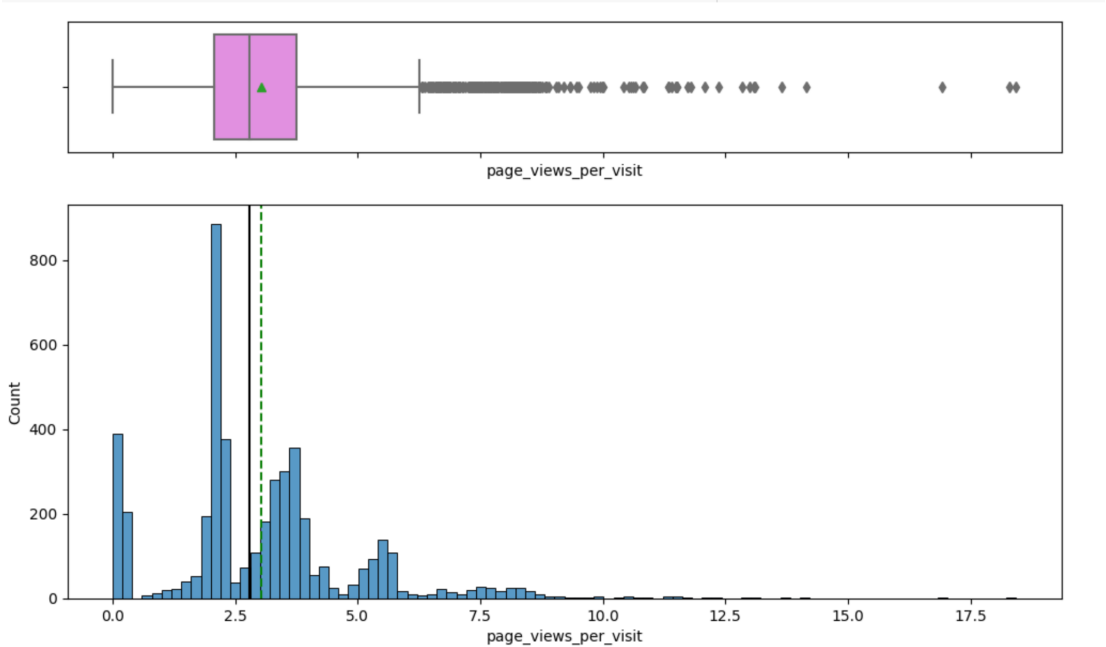
- The website visits is skewed right. Most people have visited the website 15 times or less.
- There are about 17 outliers who have visited the company's website more than 15 times.
- Of all the potential customers, 174 of them have never visited the website. So majority of leads have visited the website.

EDA Results



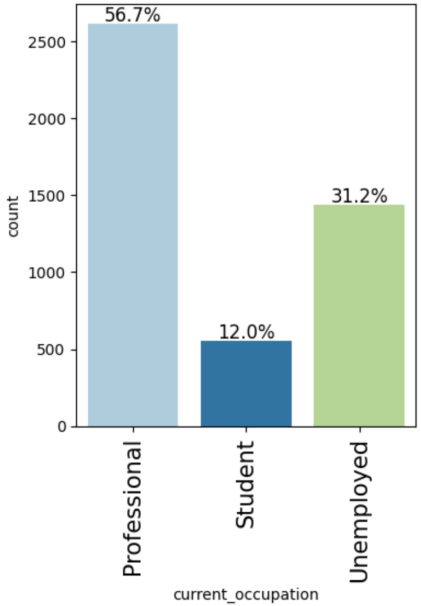
- The time spent on the website is highly skewed to the right. Most people spend less than 500 secs (i.e. 8.33 mins) on the website.

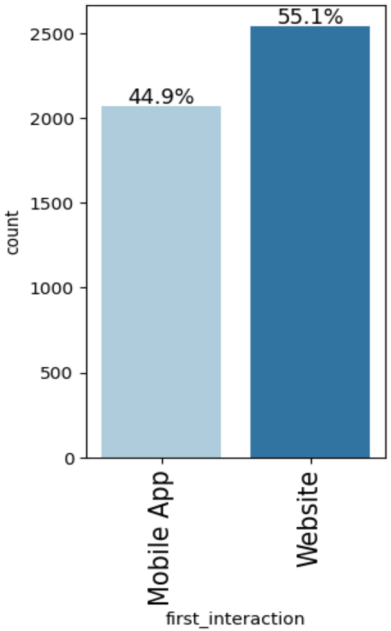
EDA Results



- The number of pages viewed per visit is highly skewed to the right.
- There are a lot of outliers too.
- Most people view roughly 1 to 8 pages per visit.

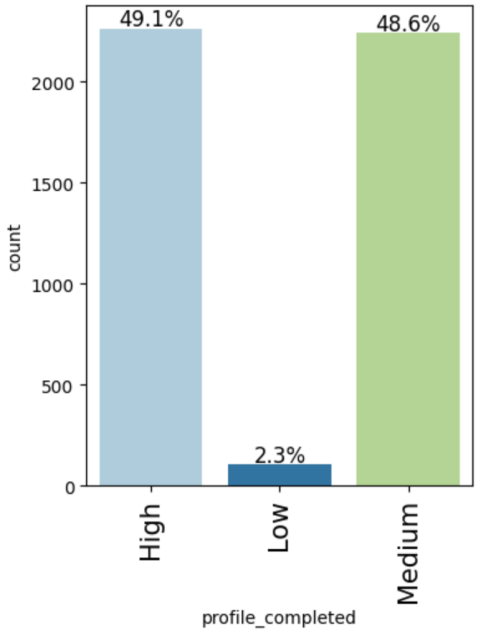
EDA Results

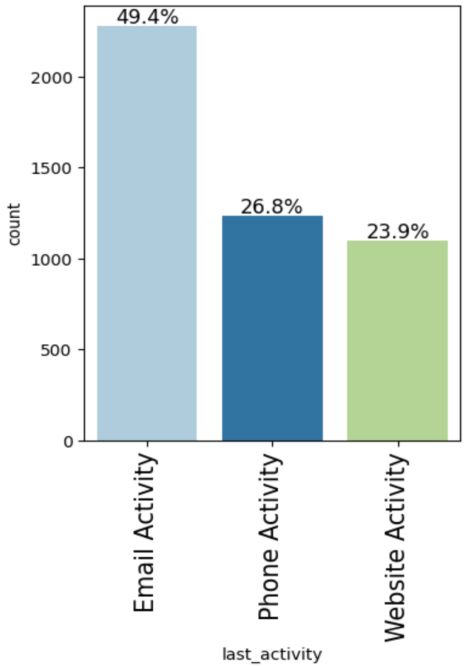




- A lot of potential leads are working professionals.
- A majority of leads make their first interaction via the company's website.

EDA Results

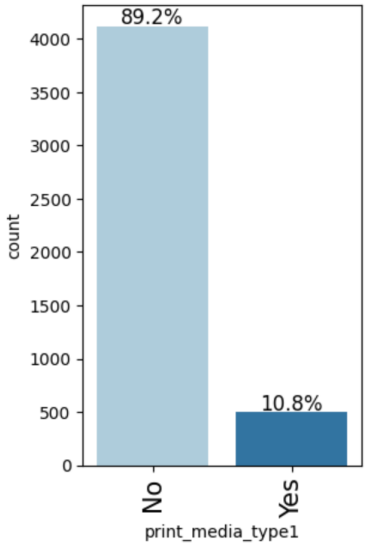


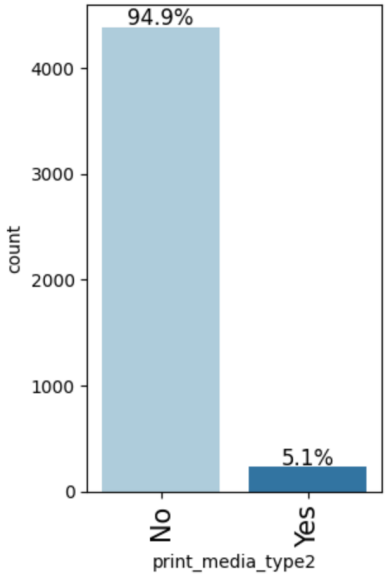


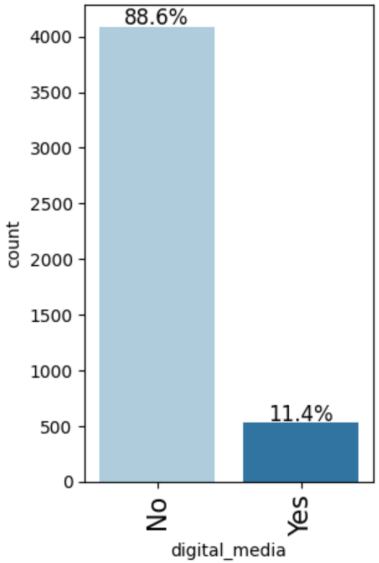
- Slightly less than 50% of leads had high profile completion. And a similar percentage had medium profile completion. So about 98% of leads had medium or high profile completion.

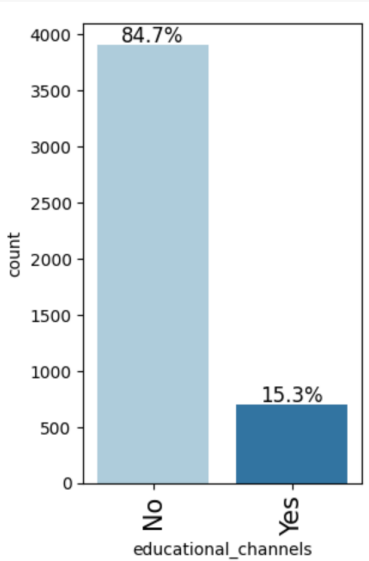
- Seems like majority of leads ultimately used email interaction later on as against the initial slant towards website activity.

EDA Results



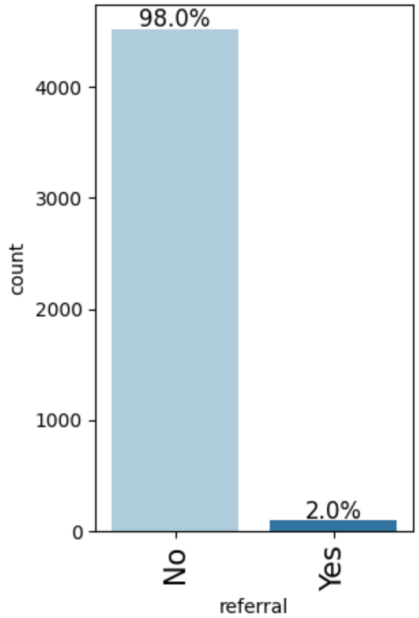


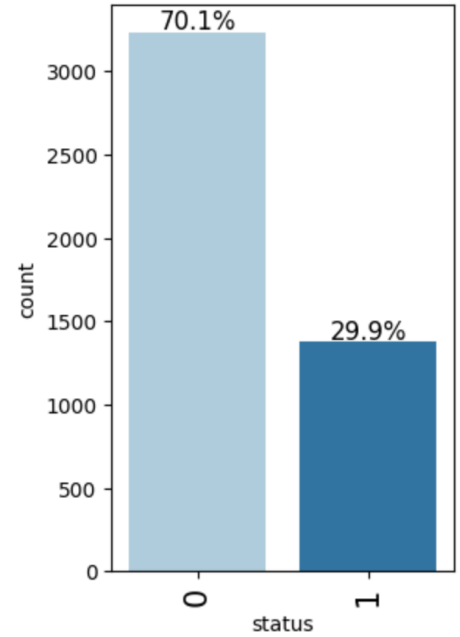




- Majority of people didn't get to know about the company through marketing channels.
- Of all the channels, the educational channels seems to have been the more popular means as compared to the rest.

EDA Results





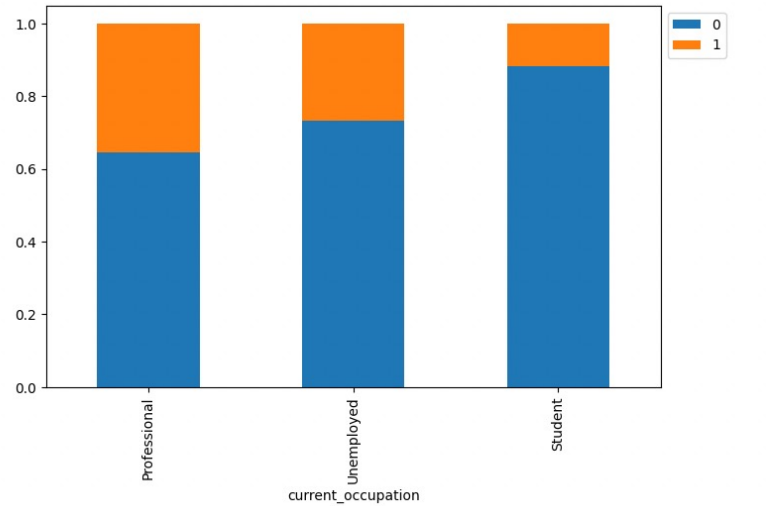
- Only a very small percentage (2%) of people got to know about the company via referrals.
- About 30% of potential leads ended up being converted while 70% of them were not.

EDA Results

- There is no significant correlation between most features.
- There is some slight correlation between the time spent on the website and status (i.e. if the lead is converted to a paid customer or not).

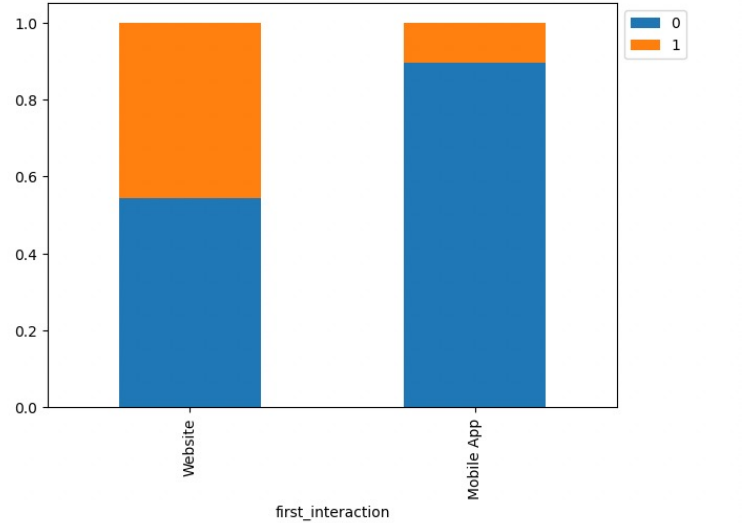
EDA Results

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



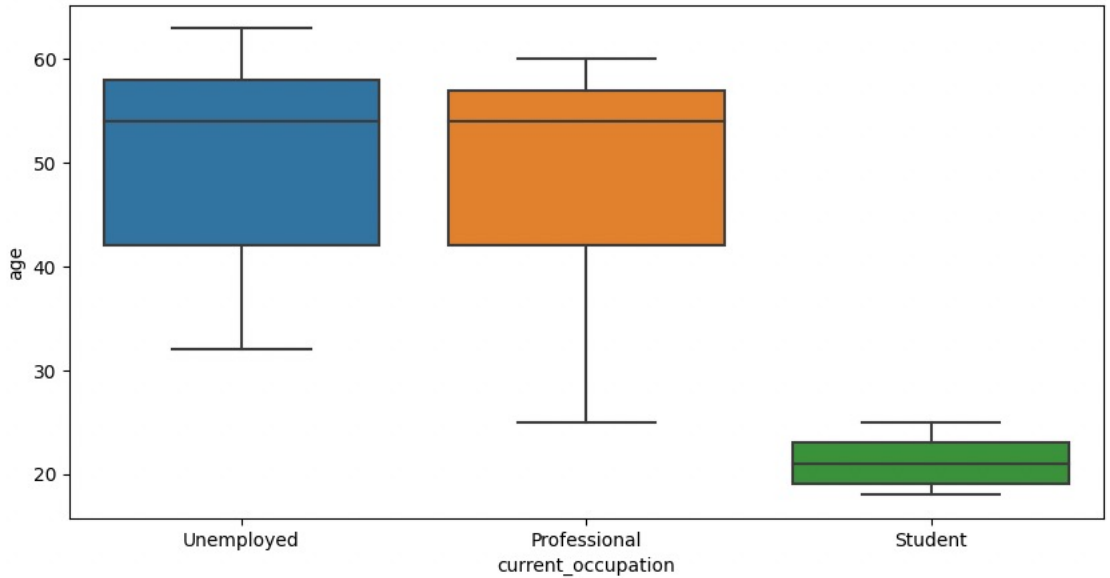
A higher proportion of working professionals seem to be converted to paid customers than unemployed people and students. This makes sense since professionals are generally able to afford to pay for the program since they are working, while unemployed people and students generally are not.

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



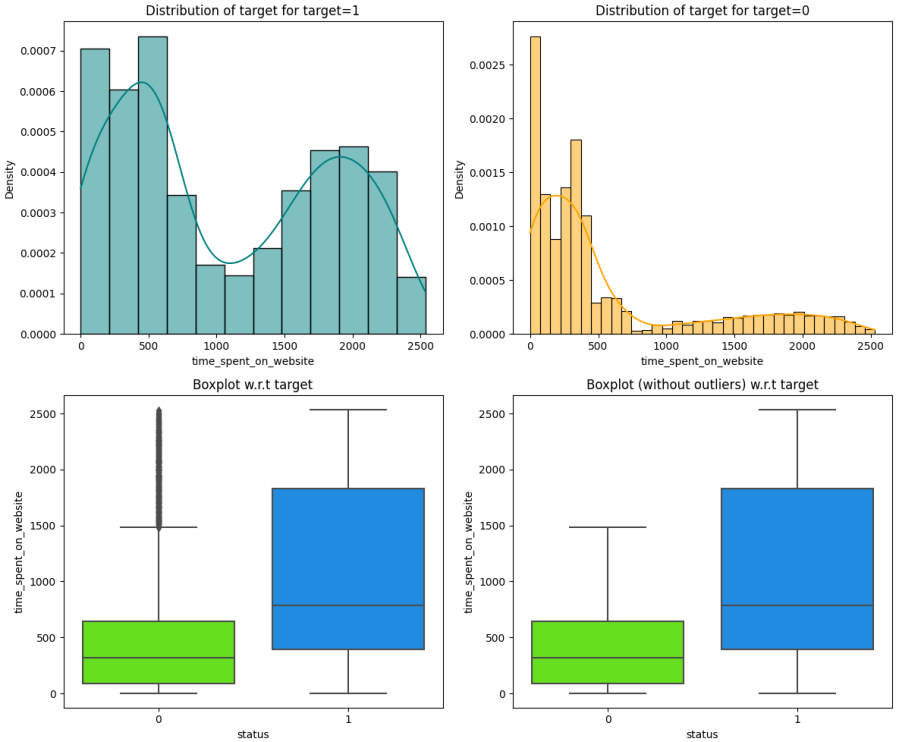
The proportion of leads converted is higher for those whose first interaction was via the website than those whose first interaction was via mobile app.

EDA Results



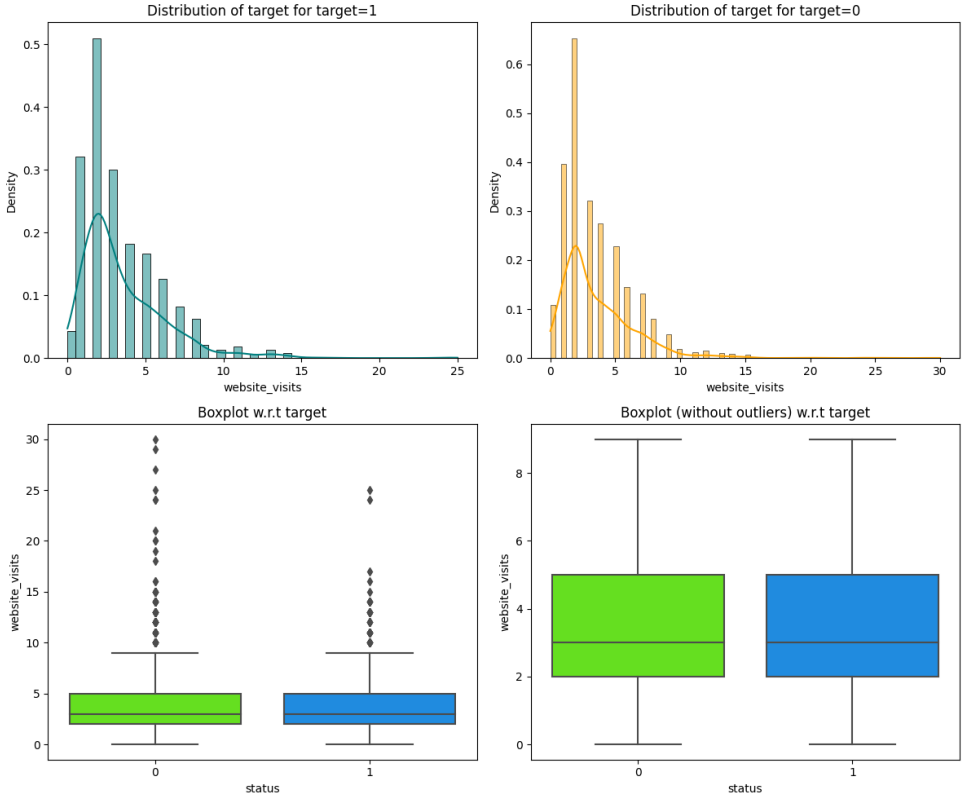
Unemployed leads (between 32 to 63 with mean 50) and Professional (between 25 to 60 with mean 50) leads have similar age distribution while students are significantly younger (approximately 18 to 25 with mean 21), as expected.

EDA Results



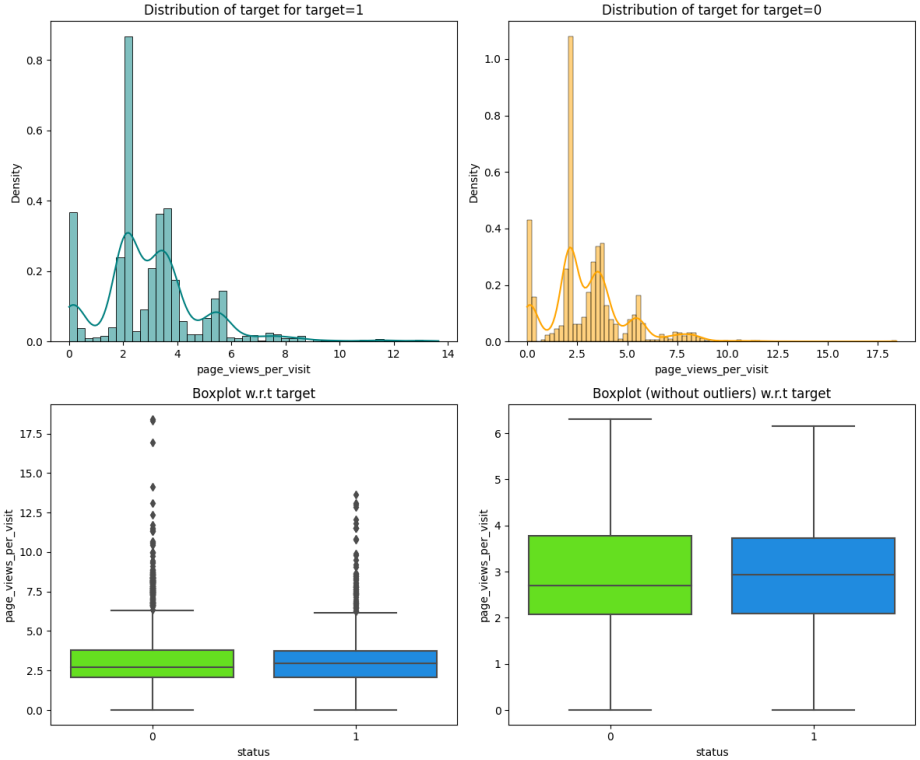
Those leads that become paid customers were generally those whose median website time was 2.5 times that of those who remained unconverted.

EDA Results



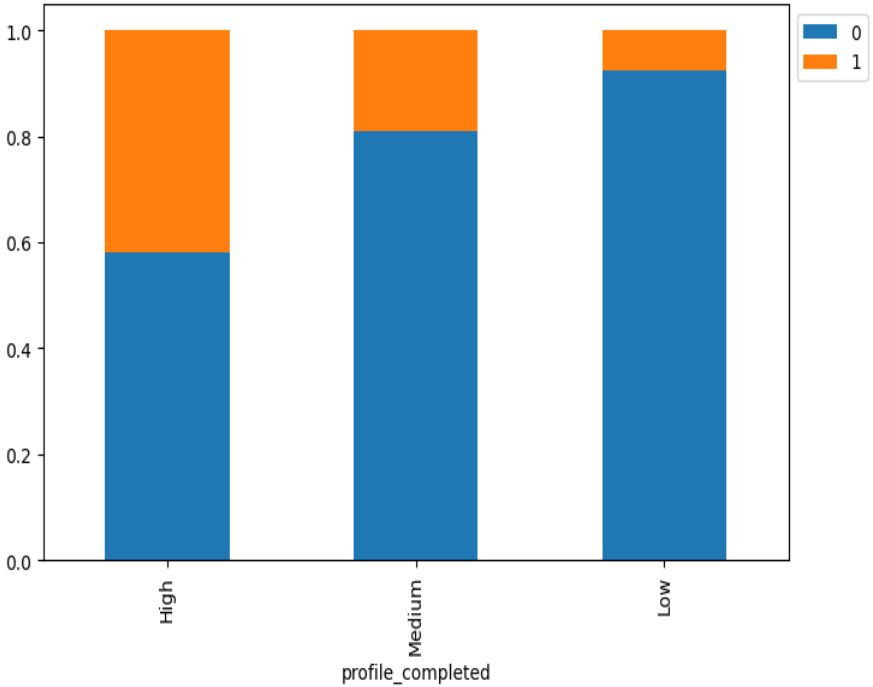
If we remove outliers, there is no significant difference between converted and unconverted leads in terms of number of website visits.

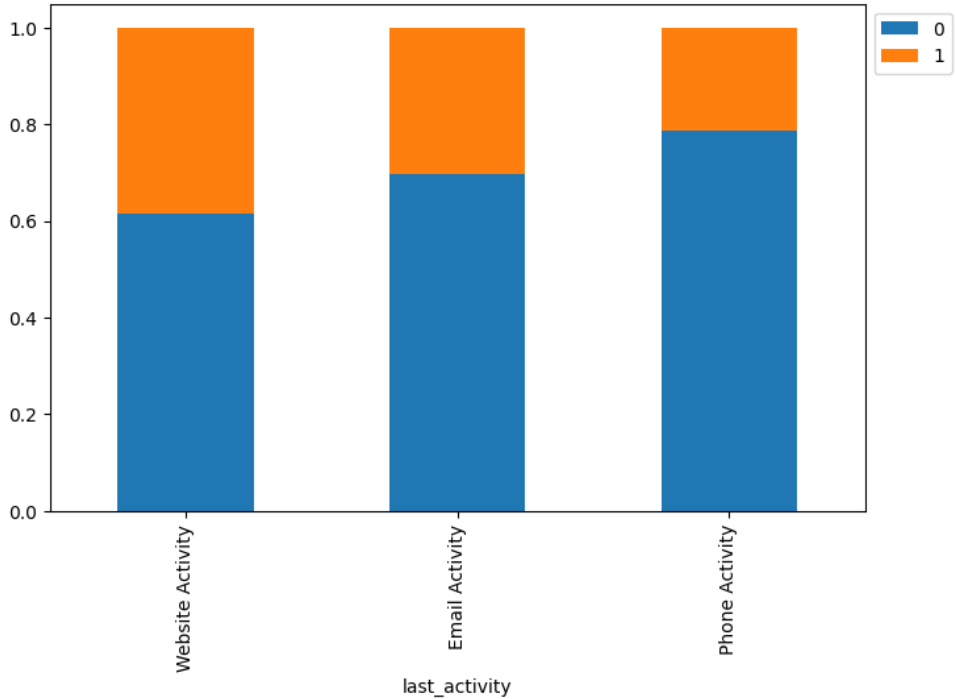
EDA Results



If we remove outliers, there is no significant difference between converted and unconverted leads in terms of number of page views per visit.

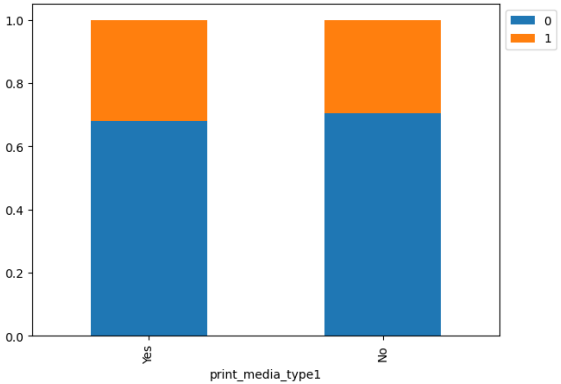
EDA Results

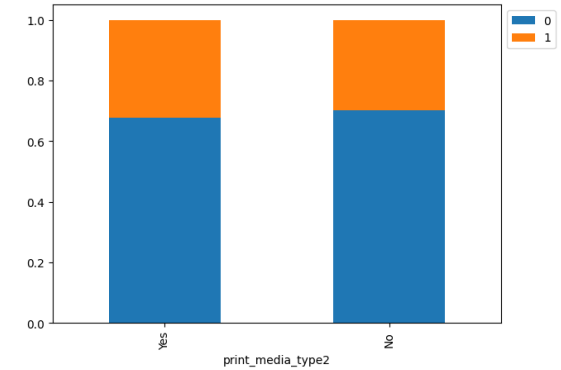


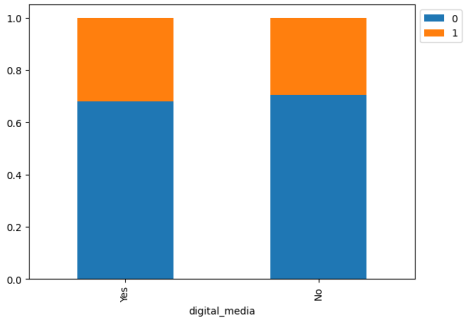


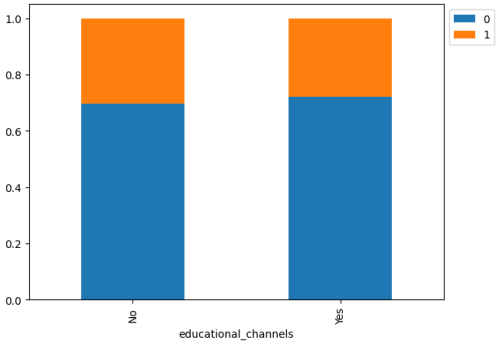
- As we would expect, those leads who had high profile completion were more likely to be converted (roughly about 40%). On the other hand, those with medium profile completion had 20% conversion rate and those with low profile completion had less than 10% conversion rate.
- If the last activity was on the website, a lead was more likely to be converted than if their last activity was via email or phone communication.

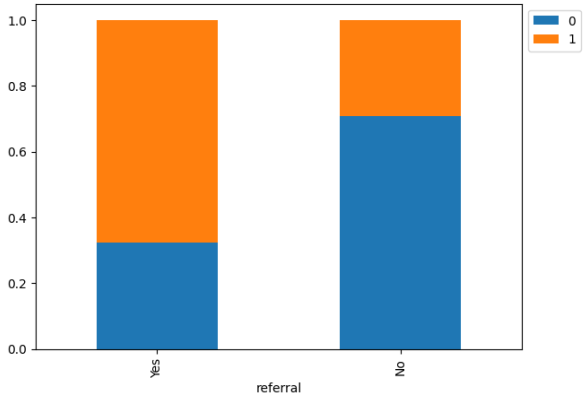
EDA Results





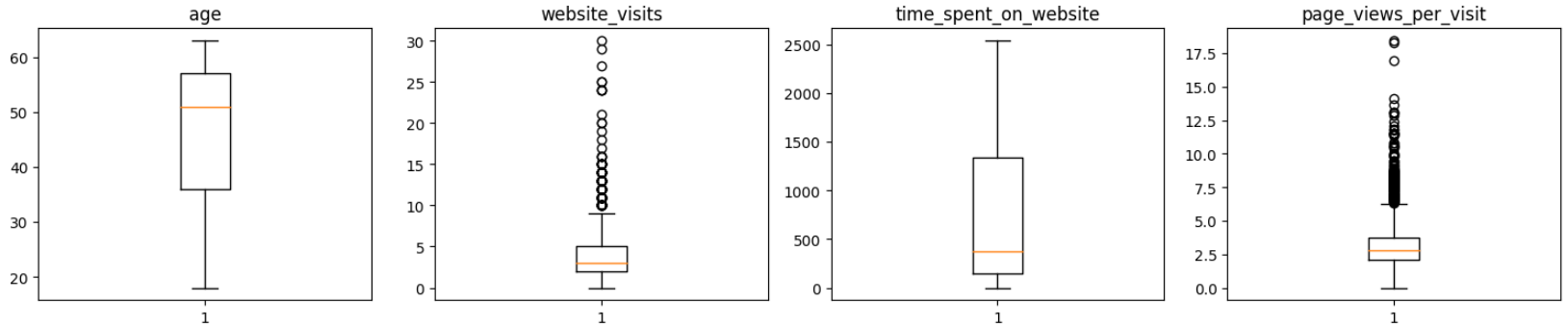






- In terms of advertisements (via various media channels), the proportion of leads converted was NOT significantly different from the unconverted ones.
- However, in terms of referrals, the proportion of those who were referred by someone they knew were more than twice likely to be converted to paid customers than those who were unreferred.

EDA Results – Outlier Check



- Both age and time spent on website have no outliers.
- Website visits have a few outliers.
- There are quite a lot of outliers with regards to page views per visit.

Model Building

Data Preparation for Modeling

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

Model Building

Training and Testing (using 70:30)

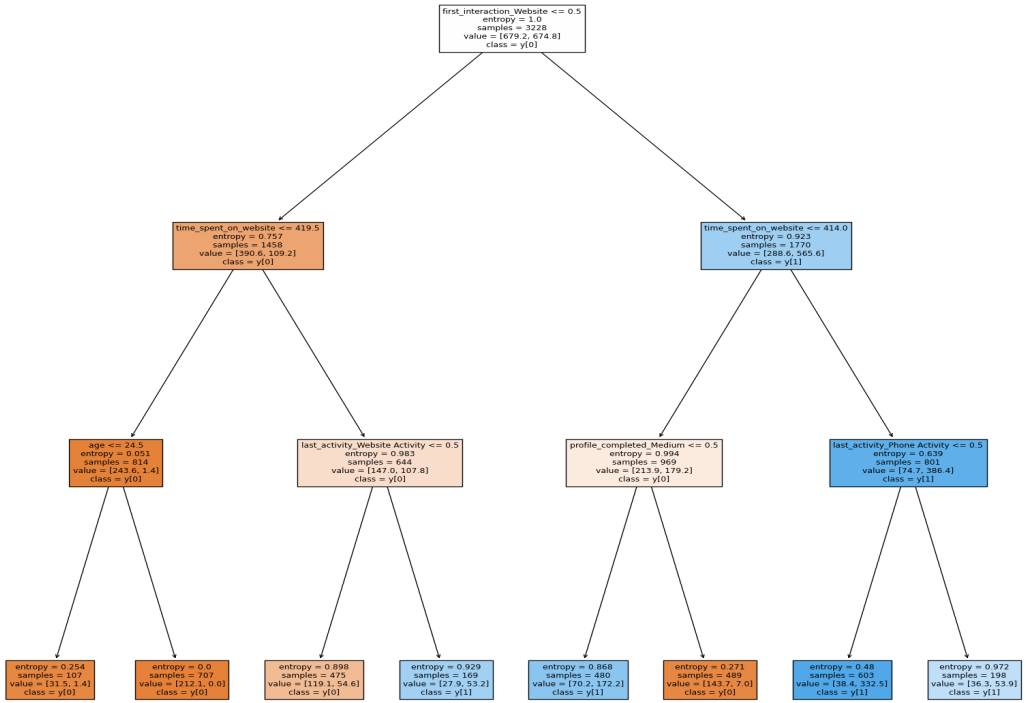
```
Shape of Training set : (3228, 16)

Shape of test set : (1384, 16)

Percentage of classes in training set: (0=not converted 1=converted)

0 0.70136

1 0.29864

Percentage of classes in test set:

0 0.70159

1 0.29841
```

Model Building

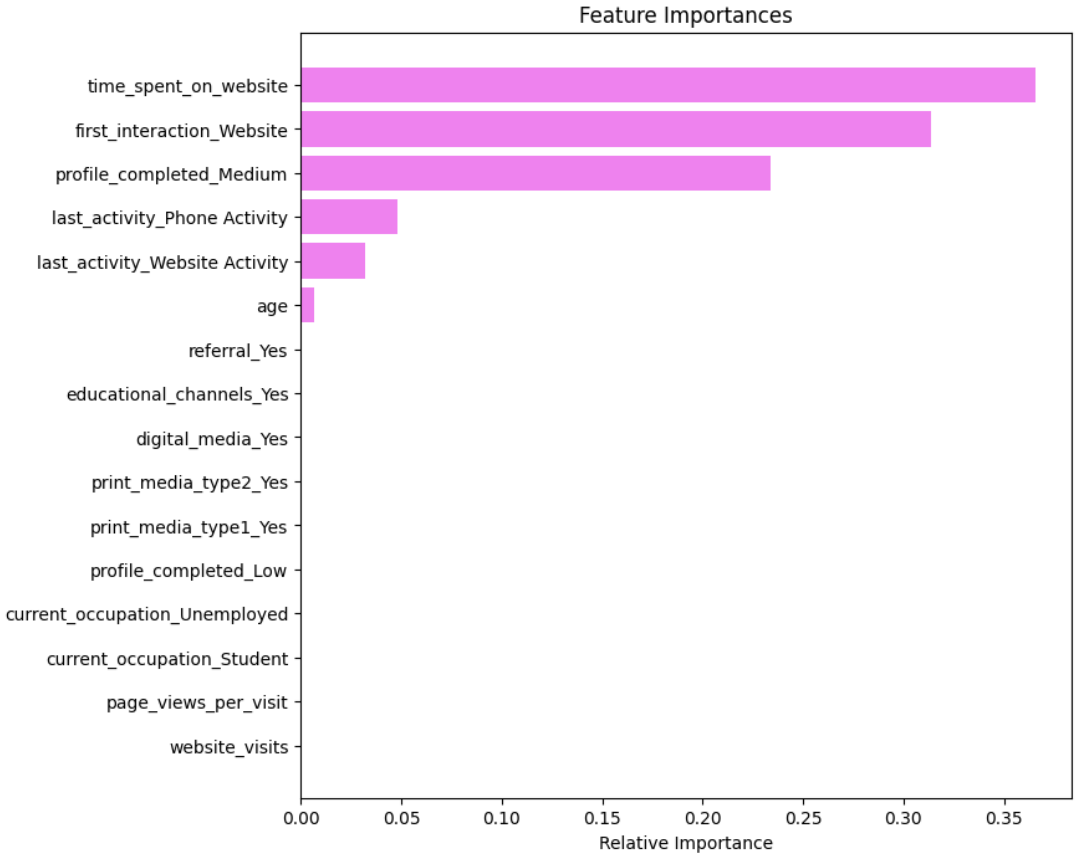
Two Models we build and tune:

- The Decision Tree
- The Random Forest

Feature Importance – Tuned Decision Tree



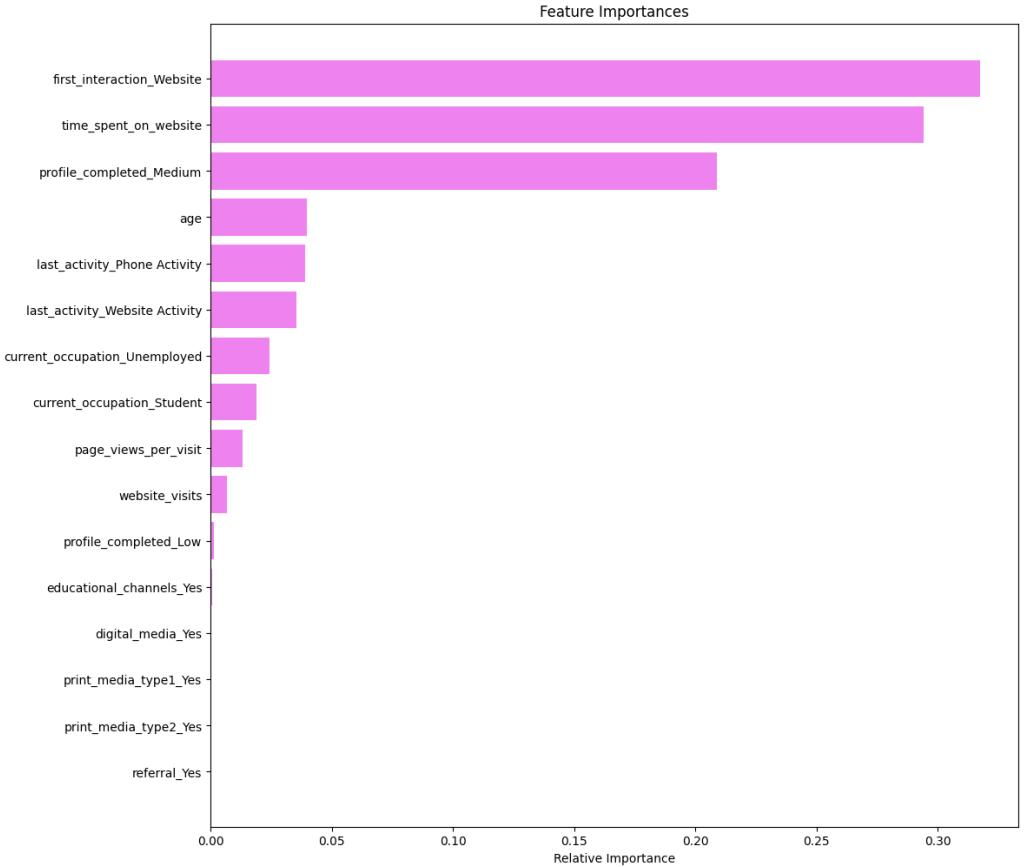
Feature Importance – Tuned Decision Tree



Observations:

- Time spent on the website and first_interaction_website are the most important features followed by profile_completed, age, and last_activity.
- The rest of the variables have no impact in this model, while deciding whether a lead will be converted or not.

Feature Importance – Tuned Random Forest



Observations:

•Similar to the decision tree model, **time spent on website, first_interaction_website, profile_completed, and age are the top four** that help distinguish between not converted and converted leads.

•Unlike the decision tree, **the random forest gives some importance to other variables like last_activity, current_occupation, page_views_per_visit, and website_visit as well.** This implies that the random forest is giving importance to more factors in comparison to the decision tree.

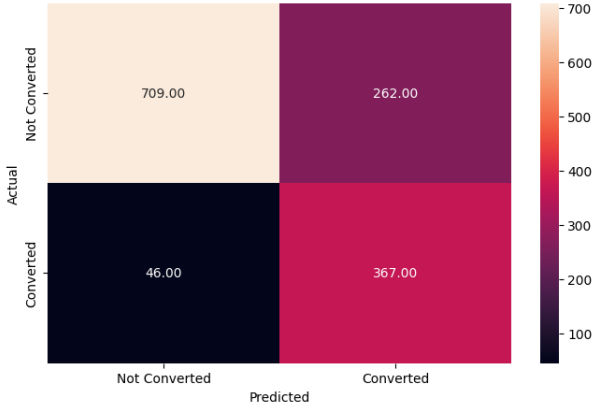
Model Performance Summary – Tuned Decision Tree

Performance on Training dataset

	precision	recall	f1-score	support
0	0.95	0.75	0.84	2264
1	0.60	0.91	0.72	964
accuracy			0.79	3228
macro avg	0.78	0.83	0.78	3228
weighted avg	0.85	0.79	0.80	3228

Performance on Test dataset

	precision	recall	f1-score	support
0	0.94	0.73	0.82	971
1	0.58	0.89	0.70	413
accuracy			0.78	1384
macro avg	0.76	0.81	0.76	1384
weighted avg	0.83	0.78	0.79	1384



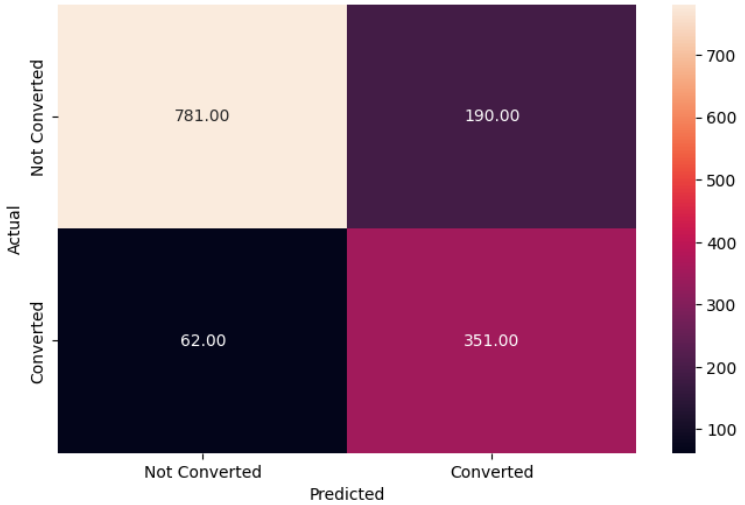
Model Performance Summary – Tuned Random Forest

Performance on Training dataset

	precision	recall	f1-score	support
0	0.94	0.82	0.88	2264
1	0.67	0.89	0.77	964
accuracy			0.84	3228
macro avg	0.81	0.85	0.82	3228
weighted avg	0.86	0.84	0.84	3228

Performance on Test dataset

	precision	recall	f1-score	support
0	0.93	0.80	0.86	971
1	0.65	0.85	0.74	413
accuracy			0.82	1384
macro avg	0.79	0.83	0.80	1384
weighted avg	0.84	0.82	0.82	1384



Model Performance Summary

- We should use the decision tree model since it outperformed the random forest model and gave more balanced metrics (and higher recall).
- We have been able to build a predictive model that can be used by the Edtech company to predict the leads who are likely to be converted (with recall score of 0.89) helping formulate marketing policies accordingly.

Business Recommendations:

- Most potential customers are in their mid to late 50s and early 60s. The company needs to target these customers more since they're more likely to be converted and age is a major factor in conversion.
- The company also needs to focus more on marketing via their educational channels since of all the media channels, this has the highest proportion of potential customers.
- The website outreach and activity needs to be prioritized since a lot of first timers contact the company via the web. However, ExtraaLearn also needs to ensure that their customer reps are more available to answer emails from leads who have already come in contact with the company.
- A lead is likely to be converted if they are slightly older, a working professional, have made first contact with the company via the website, mostly completed their online profile and spend ample time on the company's webpage. Using our model, this is the typical (projected) profile of a lead who is likely to become a paid customer.

APPENDIX