ExtraaLearn Project

Shreya Saxena

8/7/24

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

- Problem

Develop a ML model that can predict which leda are more likely to convert to paid customers for ExtraaLearn

- Initial Recommendations
  - Come with ways to advertise it to older age groups- preferably in the late 40s
  - Make the website more easy to read since leads tend to spend more time on the website
  - Partner up with companies since more leads with jobs visit us
  - Newspaper, Magazine, references, and other digital platforms are not useful in advertisements. Figure out another way to advertise us

2. Data Overview

Data Overview

Shape of the data

- 4612 rows
- 15 columns

Data Types

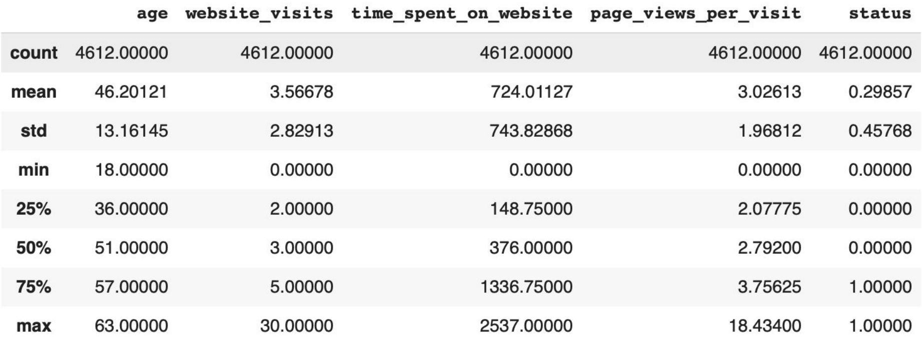
- ID: object
- Age: integer
- Current occupation: object
- First interaction: object
- Profile completed: object
- Website visits: integer
- Time spent on website: integer
- Page views per visit: float
- Last activity: object
- Print media type 1: object
- Print media type 2: object
- Digital media: object
- Educational channels: object
- Referral: object
- Status: integer

Data Overview

How many duplicate entries are in the data?

- 0

Statistical Summary of the data



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

Data Overview

Unique Values

- ID
  - Each ID appears once
  - 4612 unique ID in total
- Current occupation
  - Has 3 unique values
  - Professional: 2616 occurrences
  - Unemployed: 1441 occurrences
  - Student: 555 occurrences
- First interaction
  - Has 2 unique values
  - Website: 2542 occurrences
  - Mobile App: 2070 occurrences
- Profile completed
  - Has 3 unique values
  - High: 2264 occurrences
  - Medium: 2241 occurrences
  - Low: 107 occurrences
- Last activity
  - Has 3 unique values
  - Email activity: 2278 occurrences
  - Phone activity: 1234 occurrences
  - Website activity: 1000 occurrences
- Print media type 1
  - Has 2 unique values
  - No: 4115 occurrences
  - Yes: 497 occurrences

Data Overview

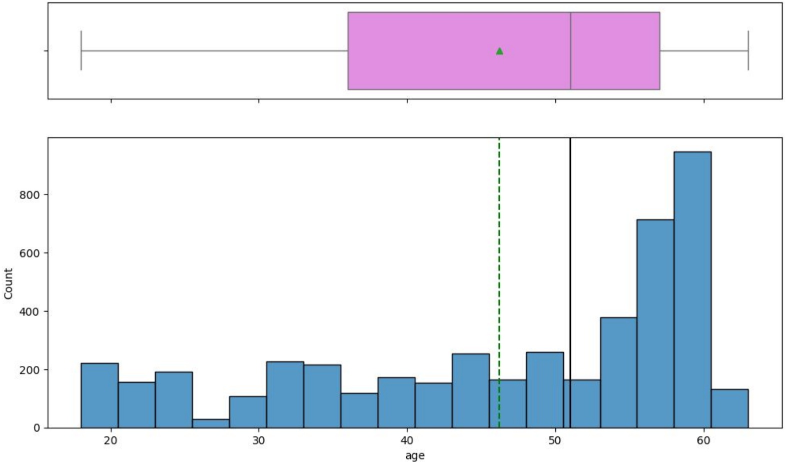
Unique Values (continued)

- Print media type 2
  - Has 2 unique values
  - No: 4379 occurrences
  - Yes: 233 occurrences
- Digital media
  - No: 4085 occurrences
  - Yes: 527 occurrences
- Educational channels
  - Has 2 unique values
  - No: 3907 occurrences
  - Yes: 705 occurrences
- Referral
  - Has 2 unique values
  - 4519 occurrences
  - 93 occurrences

3. EDA Results

Univariate Analysis

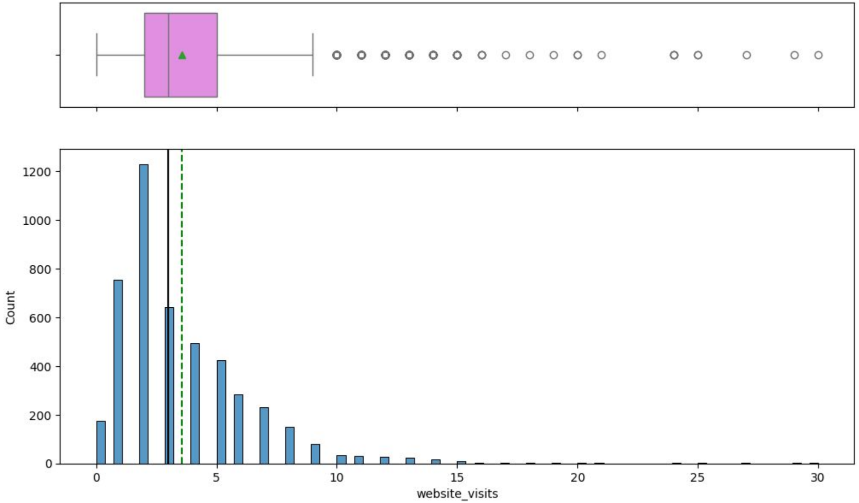
Observations on age



- Bimodal Distribution
- Ranges between late teens - mid 60s
- Median age is around early 50s
- No outliers
- More aged in the 50s and 60s

Univariate Analysis

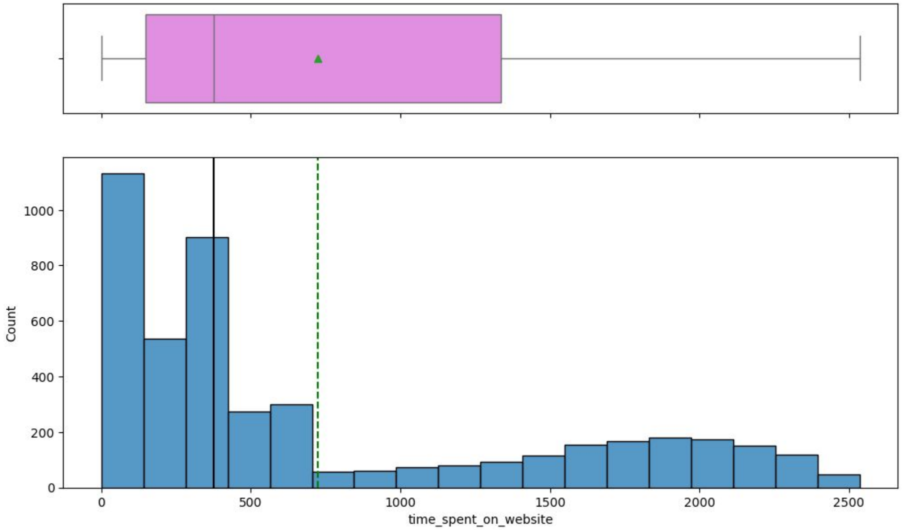
Observations on website visits



- Skewed to the right
- Leads visit website anywhere between 0-10 times, with some outliers from 10-30 minutes
- The median is about 3 times
- Most visit 1-3 times

Univariate Analysis

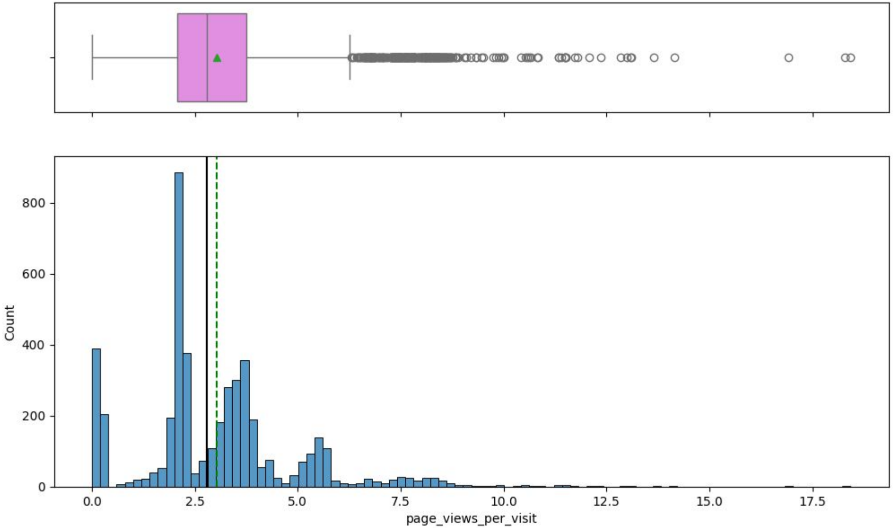
Observations on number of time spent on website



- Bimodal distribution
- Leads spent anywhere between 0-2500 minutes on website
- No outliers
- The median is about 250 minutes
- Most spent from 0-500 minutes

Univariate Analysis

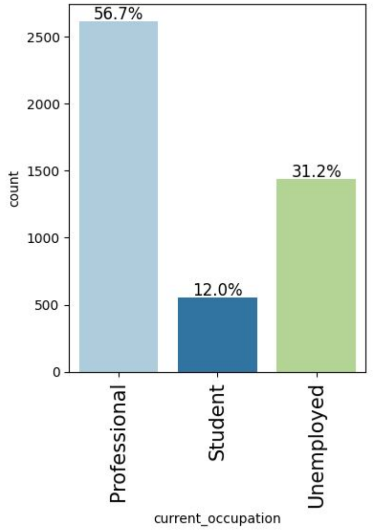
Observations on number of page views per visit



- Asymmetrical shape
- 0-6 page visits per visit, with outliers from 6-17.5 visits
- Median is about 3 views per visit
- Most of the time, the page views per visit are are 0, 2, and around 3

Univariate Analysis

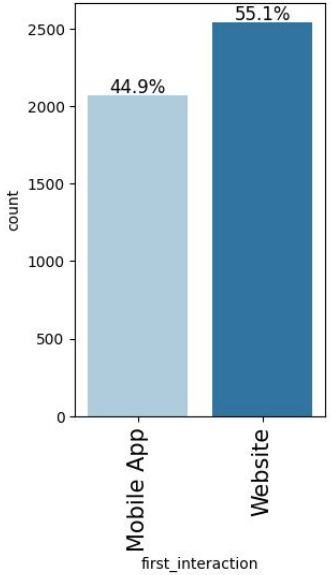
Observations on current occupation



- Most leads have professional jobs
- Less than half of them are unemployed
- A very few of them are students

Univariate Analysis

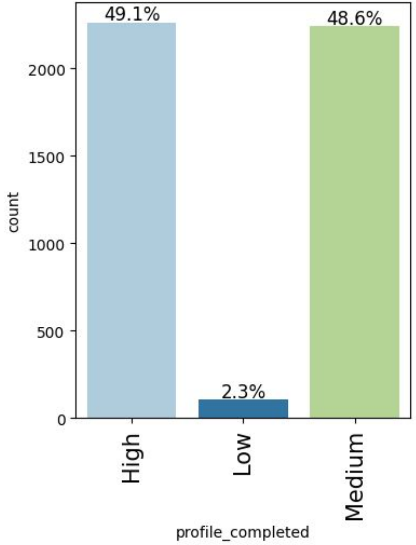
Observations on number of first interaction



- First interactions are more on the website than the mobile app

Univariate Analysis

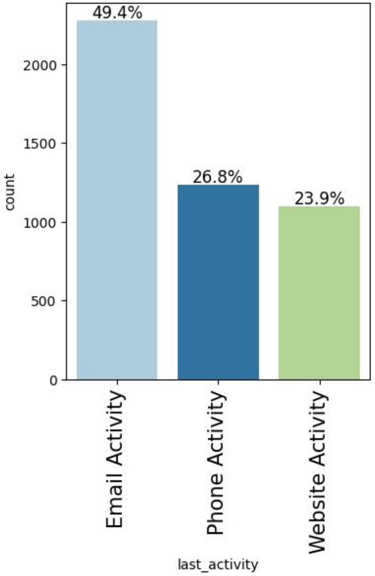
Observations on profile completed



- Most are high or medium
- Very few are low

Univariate Analysis

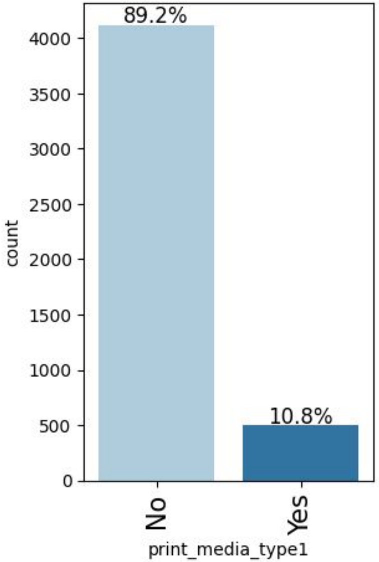
Observations on last activity



- Most last activity was done through email
- The rest are either through phone or the website

Univariate Analysis

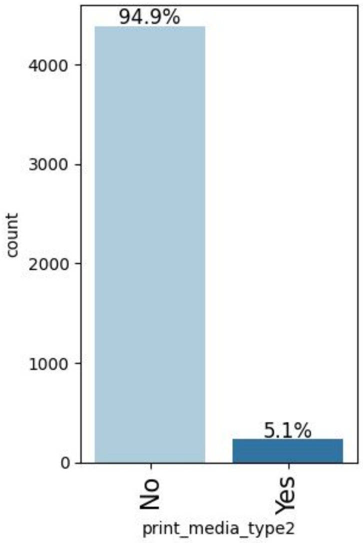
Observations on print media type 1



Most leads had not seen the ad of ExtraaLearn in the Newspaper

Univariate Analysis

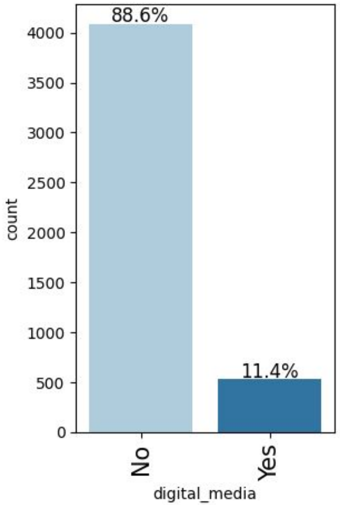
Observations on print media type 2



Most leads had not seen the ad of ExtraaLearn in the Magazine

Univariate Analysis

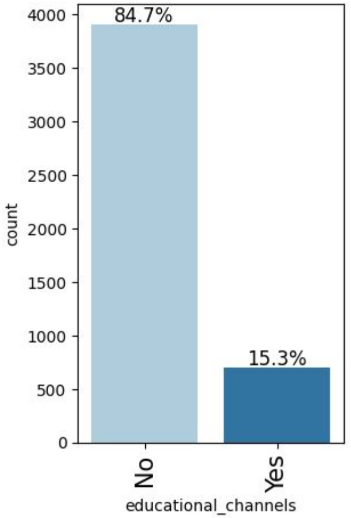
Observations on digital media



Most leads had not seen the ad of ExtraaLearn on the digital platforms

Univariate Analysis

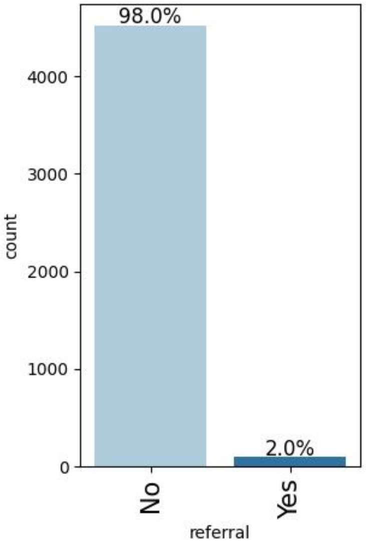
Observations on educational channels



Most leads had not heard about ExtraaLearn in the education channels like online forums, discussion threads, educational websites, etc.

Univariate Analysis

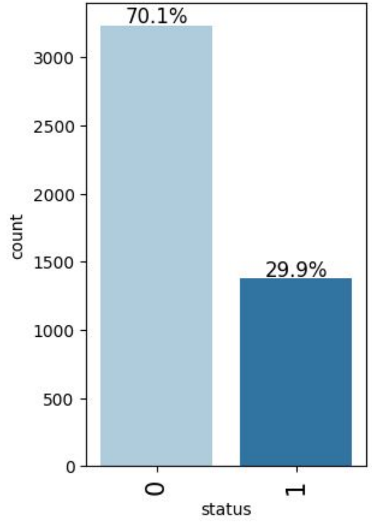
Observations on referral



Most leads had not heard about ExtraaLearn through reference.

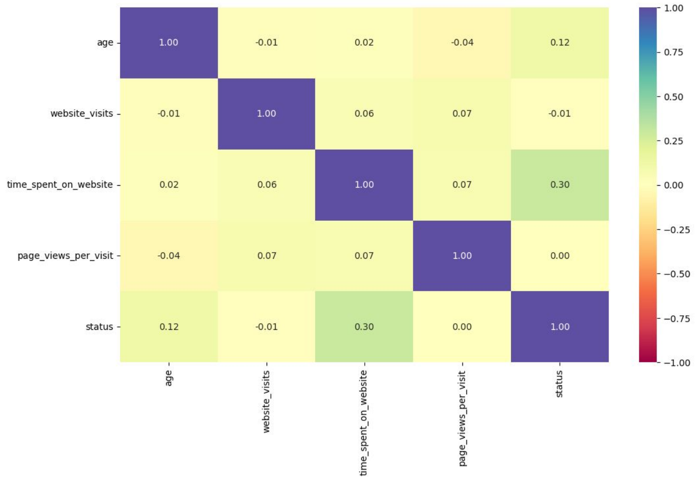
Univariate Analysis

Observations on status



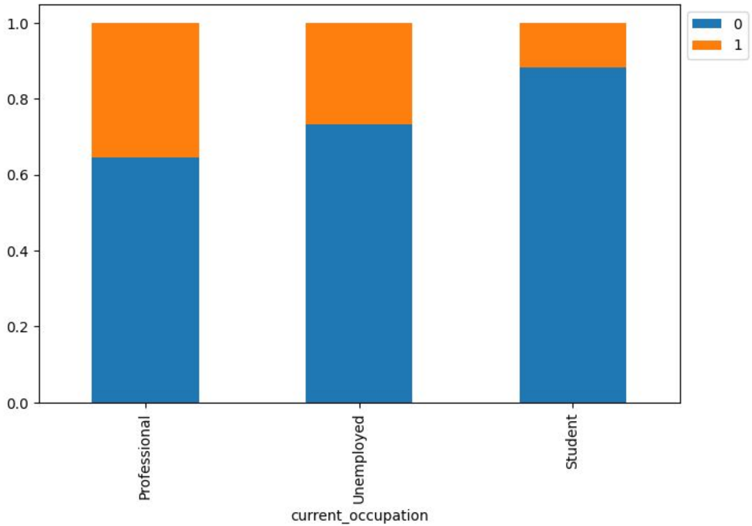
Most leads were not converted to a paid customer

Bivariate Analysis



Bivariate Analysis

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



Bivariate Analysis

Age can be a good factor to differentiate between such leads.

Statistical Summary (current occupation and age)

Professional:

- Count: 2616
  - This indicates that there are 2616 individuals categorized as professionals in the dataset.
- Mean Age: 49.35
  - The average age of professionals is approximately 49.35 years.
- Standard Deviation (Std): 9.89
  - The age variation among professionals is about 9.89 years.
- Min Age: 25
  - The youngest professional is 25 years old.
- 25th Percentile (25%): 42
  - 25% of professionals are aged 42 or younger.
- Median (50%): 54
  - The median age for professionals is 54 years.
- 75th Percentile (75%): 57
  - 75% of professionals are aged 57 or younger.
- Max Age: 60
  - The oldest professional is 60 years old.

Bivariate Analysis

Statistical Summary (continued)

Student:

- Count: 555
  - There are 555 students in the dataset.
- Mean Age: 21.14
  - The average age of students is approximately 21.14 years.
- Standard Deviation (Std): 2.00
  - The age variation among students is about 2.00 years.
- Min Age: 18
  - The youngest student is 18 years old.
- 25th Percentile (25%): 19
  - 25% of students are aged 19 or younger.
- Median (50%): 21
  - The median age for students is 21 years.
- 75th Percentile (75%): 23
  - 75% of students are aged 23 or younger.
- Max Age: 25
  - The oldest student is 25 years old.

Bivariate Analysis

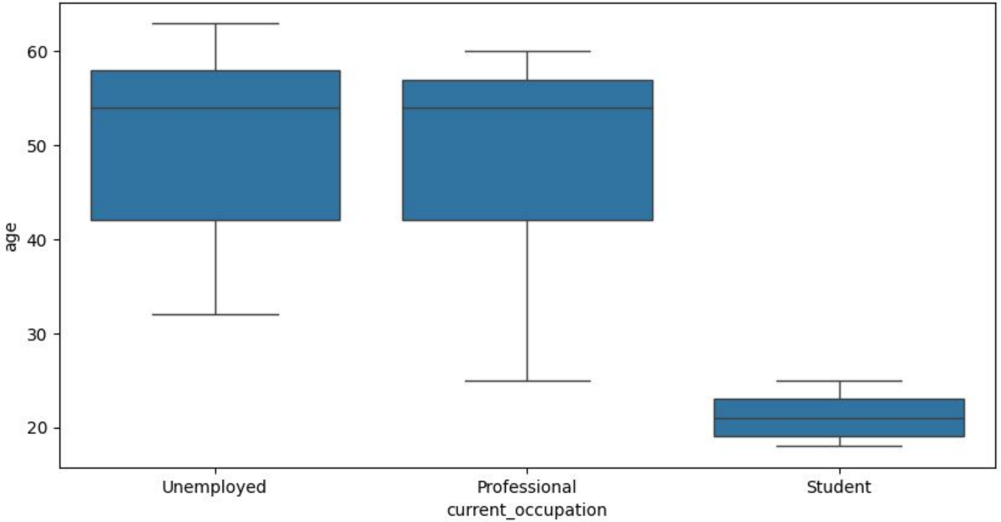
Statistical Summary (continued)

Unemployed:

- Count: 1441
  - There are 1441 individuals categorized as unemployed in the dataset.
- Mean Age: 50.14
  - The average age of unemployed individuals is approximately 50.14 years.
- Standard Deviation (Std): 9.99
  - The age variation among unemployed individuals is about 9.99 years.
- Min Age: 32
  - The youngest unemployed individual is 32 years old.
- 25th Percentile (25%): 42
  - 25% of unemployed individuals are aged 42 or younger.
- Median (50%): 54
  - The median age for unemployed individuals is 54 years.
- 75th Percentile (75%): 58
  - 75% of unemployed individuals are aged 58 or younger.
- Max Age: 63
  - The oldest unemployed individual is 63 years old.

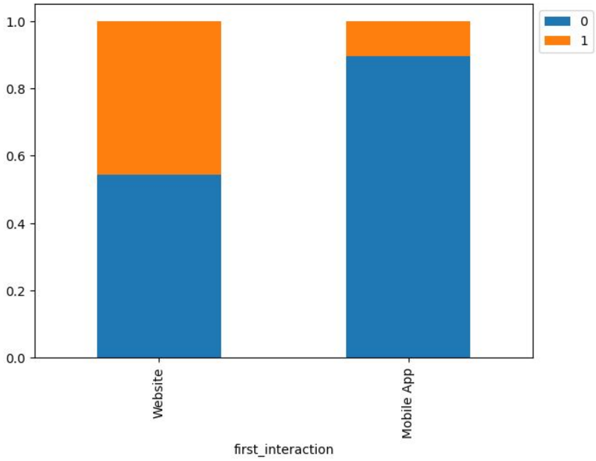
Bivariate Analysis

Boxplot that shows the statistical summary of current occupation and age:



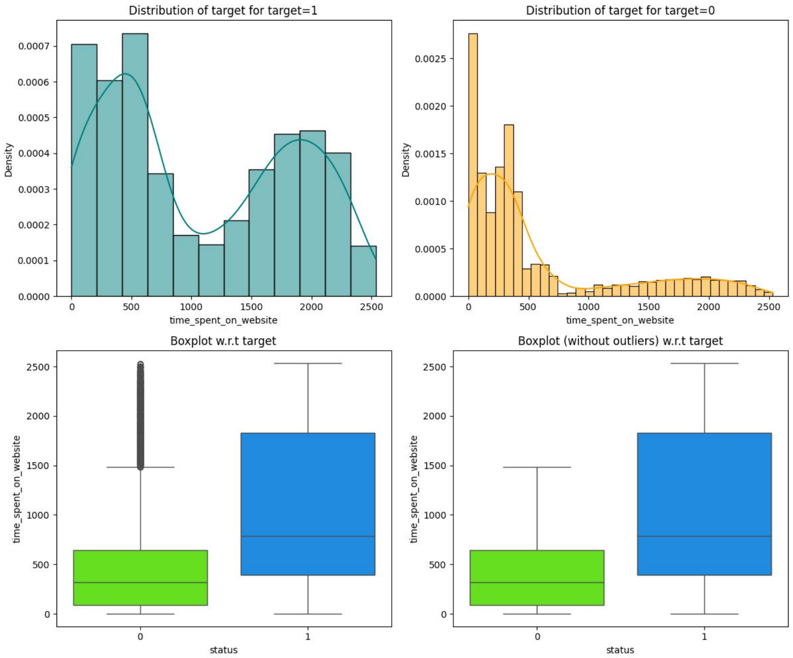
Bivariate Analysis

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



Bivariate Analysis

Time spend on website vs the different categories of status variable:



Bivariate Analysis

Median values of time spent on website on each status:

Status 0 (Not Converted)

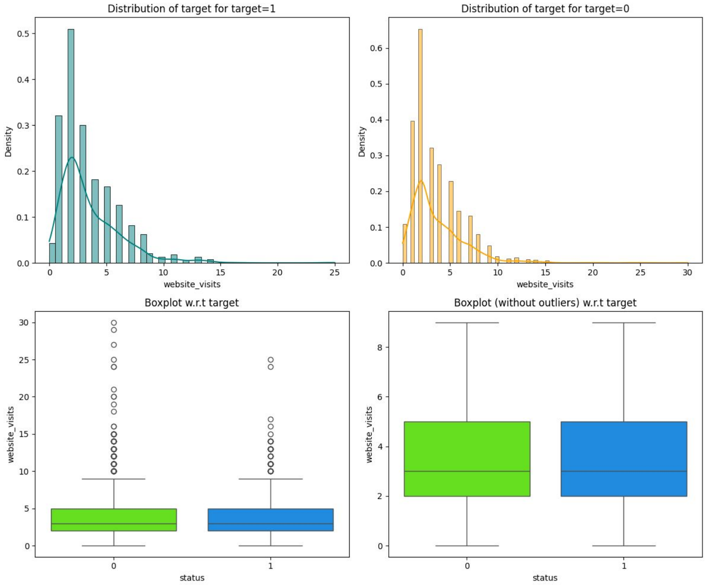
- Median time spent on website: 317 minutes

Status 1 (Converted)

- Median time spent on website: 789.00 minutes

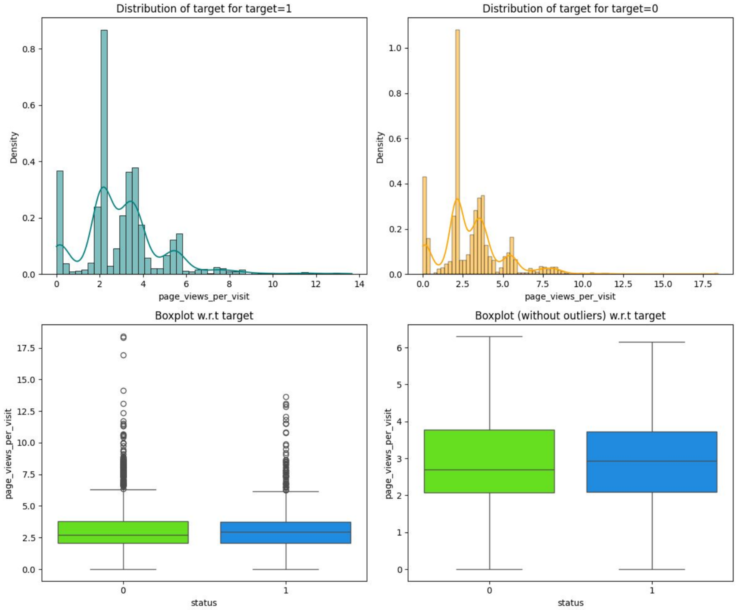
Bivariate Analysis

Time spent on website vs the different categories of status variable:



Bivariate Analysis

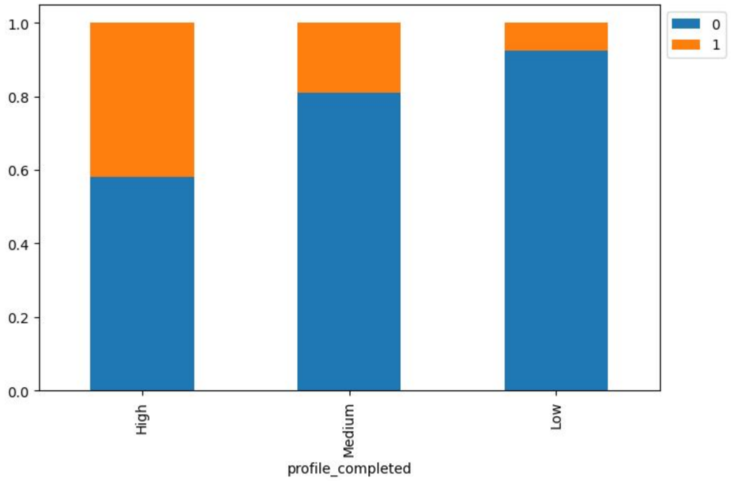
Page views per visit vs the different categories of status variable:



Bivariate Analysis

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

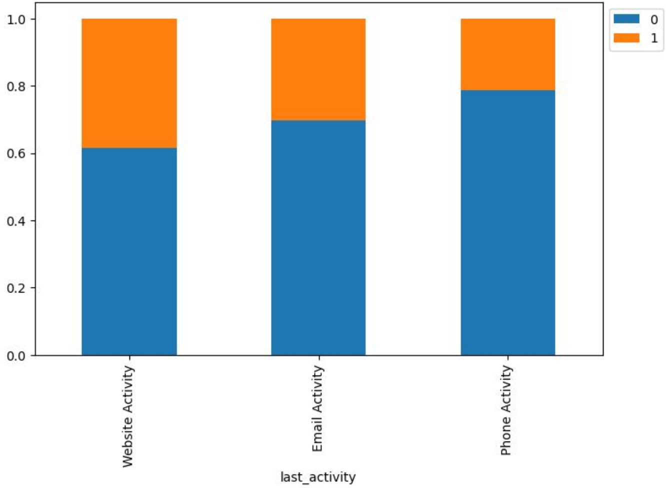
Profile completed vs status



Bivariate Analysis

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

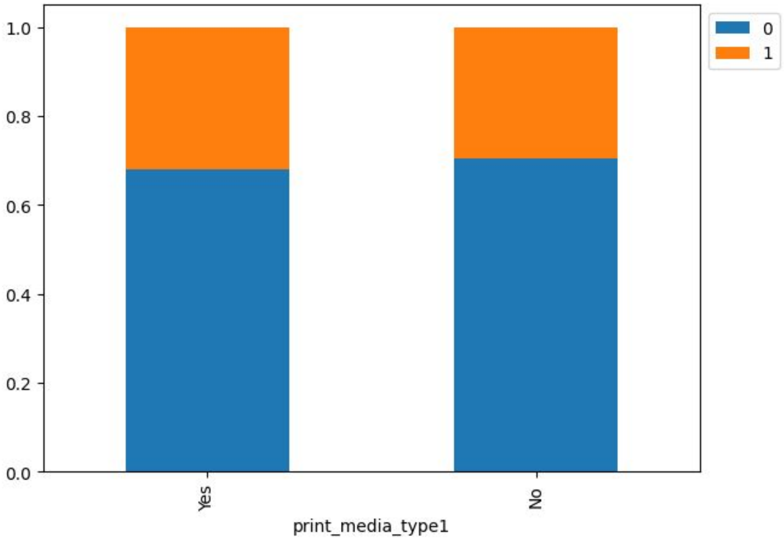
Last activity vs status



Bivariate Analysis

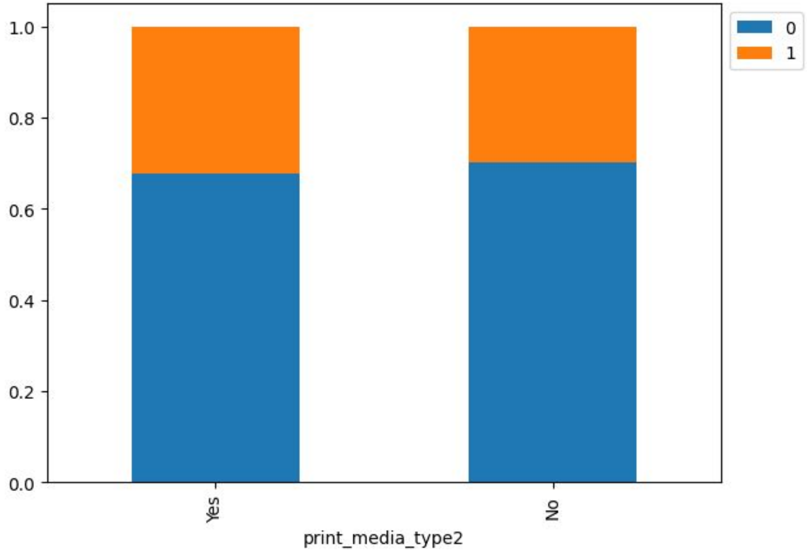
Let's see how advertisement and referrals impact the lead status:

Print media type 1 and status



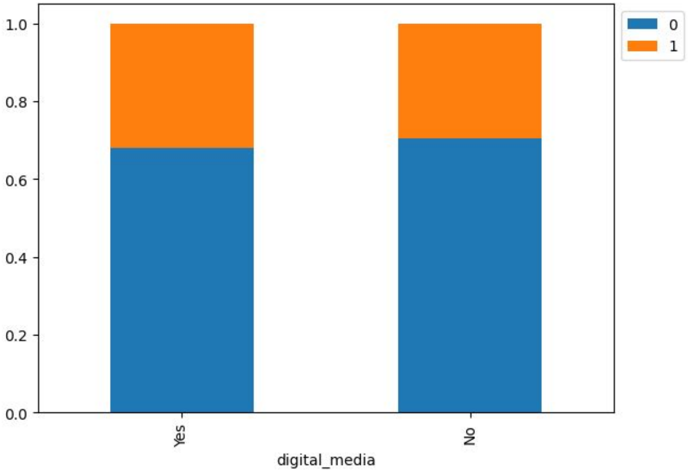
Bivariate Analysis

Print media type 2 and status



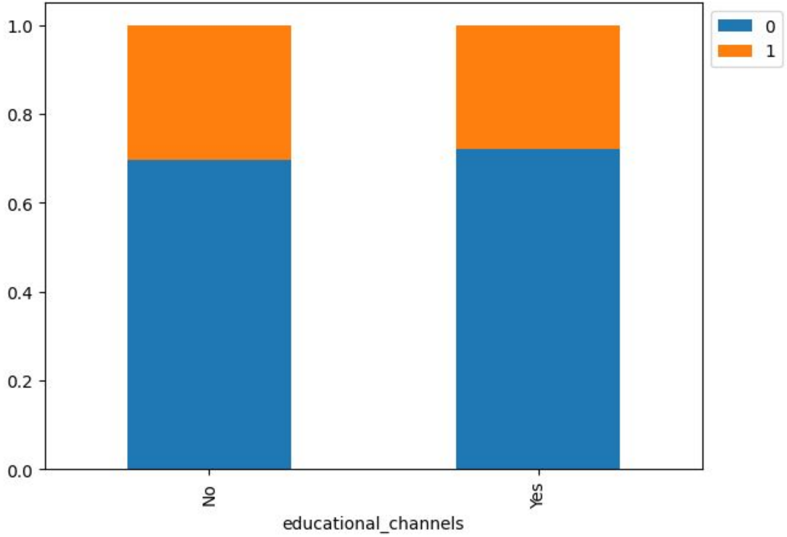
Bivariate Analysis

Digital media and status



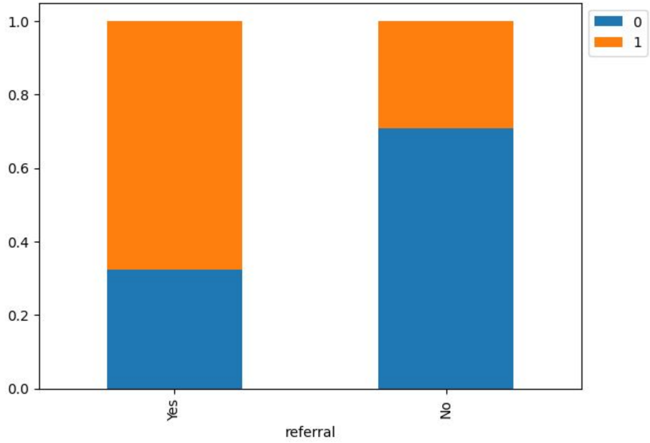
Bivariate Analysis

Educational channels and status

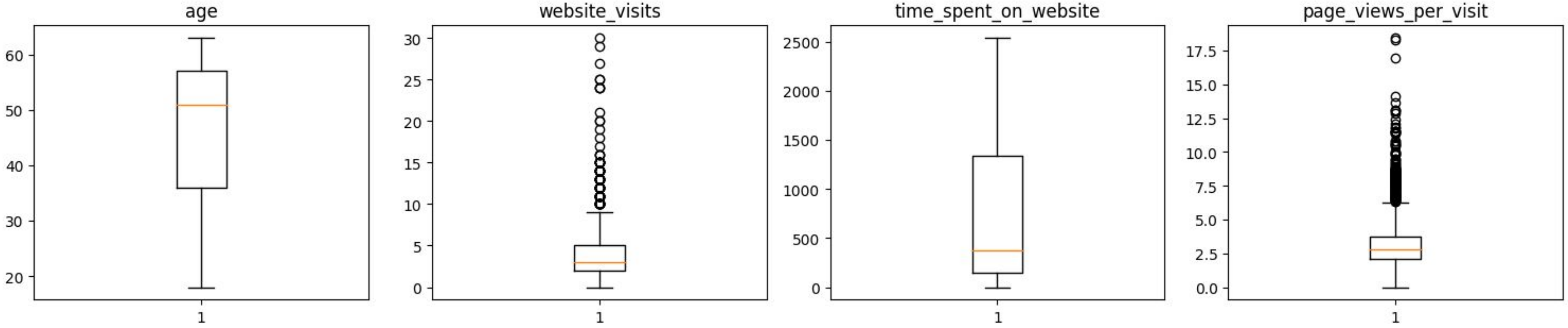


Bivariate Analysis

Referral and status



Outlier Check



4. Model Building & Performance Summary

Data Preparation and Splitting for modeling

Data Separation:

- Target variable status is separated from the feature set.
- Features are stored in X and the target in Y.

- Categorical features in X are converted to numerical values using one-hot encoding.
- drop_first=True removes the first level of each categorical feature to avoid multicollinearity.

Train-Test Split:

- The data is split into training (70%) and testing (30%) sets.
- random_state=1 ensures reproducibility of the split.

Data Preparation and Splitting for modeling

Dataset Shapes:

- Prints the shape of the training and testing sets.
- Training set shape: (3228, 4627)
- Testing set shape: (1384, 4627)

Class Distribution:

- Displays the distribution of target classes (status) in both training and testing sets.
- Training set:
  - Class 0: 70.415%
  - Class 1: 29.585%
- Testing set:
  - Class 0: 69.509%
  - Class 1: 30.491%

Building Classification Models

Model evaluation criterion

- Possible wrong predictions (false positives/negatives)
- Importance of the cases

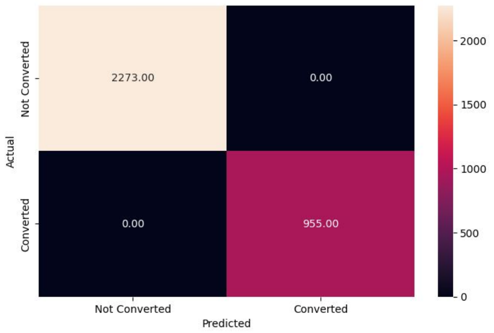
Reducing losses

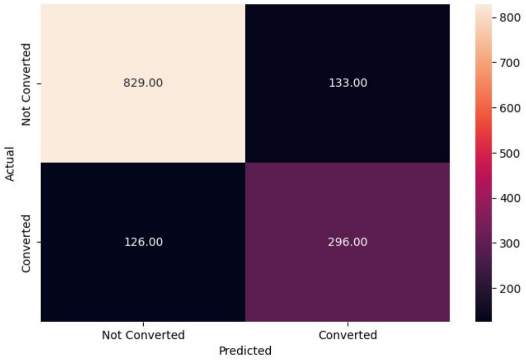
- Maximize recall ensures fewer leads are missed

Made functions for Model Performance Evaluation

- metrics_score

Decision Tree





Observations on Training Data:

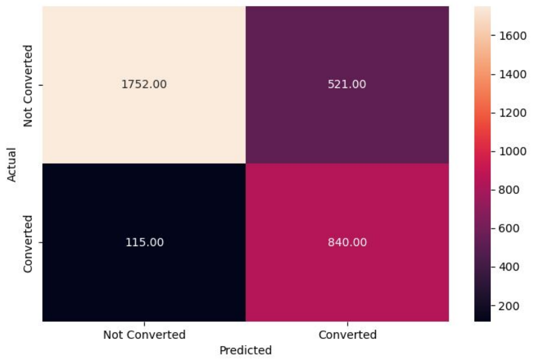
- Precision, Recall, F1-Score: Perfect scores of 1.00 indicate that the model performs exceptionally well on the training data.
- Accuracy: 100%

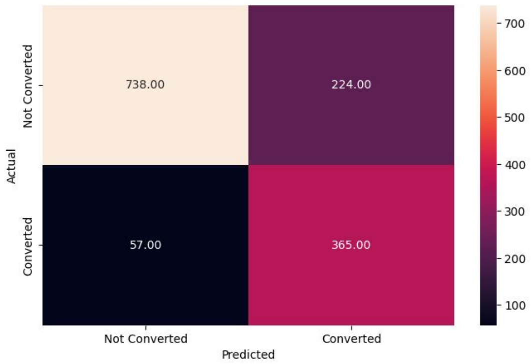
Observations on Test Data:

- Precision, Recall, F1-Score for Class 0: 0.87, 0.86, 0.86, indicating good performance for predicting non-converted leads.
- Precision, Recall, F1-Score for Class 1: 0.69, 0.70, 0.70, indicating moderate performance for predicting converted leads.
- Accuracy: 81%, which is significantly lower than the training accuracy, suggesting the model is overfitting.

Decision Tree - Hyperparameter Tuning

In order to reduce the overfitting of the Decision Tree Model, we used hyperparameters using GridSearchCV.





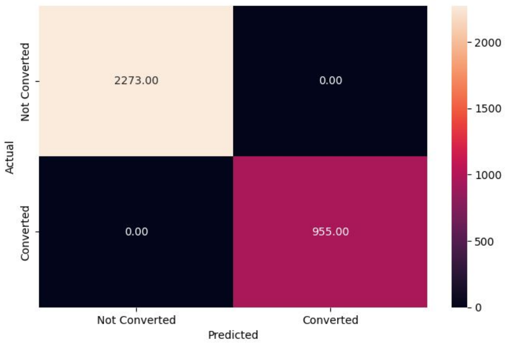
Observations on Training Data:

- Precision, Recall, F1-Score for Class 0: 0.94, 0.77, 0.85.
- Precision, Recall, F1-Score for Class 1: 0.62, 0.88, 0.73.
- Accuracy: 80%, which shows improvement in reducing overfitting.

Observations on Test Data:

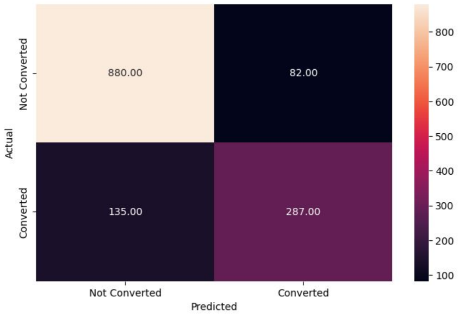
- Precision, Recall, F1-Score for Class 0: 0.93, 0.77, 0.84.
- Precision, Recall, F1-Score for Class 1: 0.62, 0.86, 0.72.
- Accuracy: 80%, indicating a balanced performance between training and testing datasets

Random Forest Classifier Model



Training Data Performance:

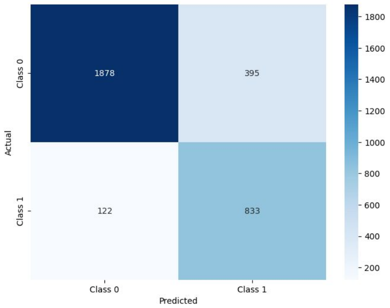
- Precision, Recall, and F1-Score: The model performs perfectly on the training data with precision, recall, and F1-scores all equal to 1.00 for both classes.
- Accuracy: The accuracy is 1.00, indicating that the model has learned the training data extremely well.

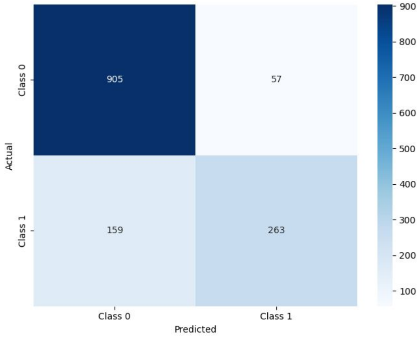


Testing Data Performance:

- Precision, Recall, and F1-Score:
  - Class 0 (negative class): The model performs quite well with a precision of 0.87, recall of 0.91, and F1-score of 0.89.
  - Class 1 (positive class): The model's performance is less strong with a precision of 0.78, recall of 0.68, and F1-score of 0.73.
- Accuracy: The overall accuracy is 0.84, which is good but shows a drop compared to the training accuracy.
- Macro Average: The macro average F1-score is 0.81, indicating the average performance across both classes.
- Weighted Average: The weighted average F1-score is 0.84, reflecting the model's overall performance considering the class distribution.

Random Forest Classifier - Hyperparameter Tuning





Training data performance

- Precision, Recall, and F1-Score: Performing quite well on the training data with precision, recall, and F1-scores for Class 0 - 0.94, 0.83, and 0.88. For Class 1, the scores are slightly lower with a precision of 0.68, recall of 0.87, and F1-score of 0.76. Indicates that while the model is good at identifying true positives for both classes, it struggles more with precision for Class 1, leading to more false positives.
- Accuracy: The overall accuracy of 0.84 suggests that the model has learned the training data well but is not overfitting, as it does not achieve perfect accuracy. This balance is often desirable to avoid overfitting.
- Macro and Weighted Averages: The macro and weighted averages indicate balanced performance across both classes, with slightly better performance on Class 0. The macro average recall of 0.85 shows that the model is fairly consistent in identifying both classes.

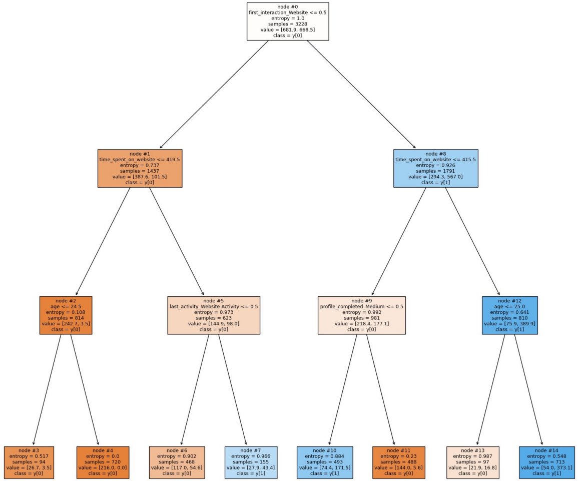
Testing data performance

- Precision, Recall, and F1-Score: On the testing data, the model maintains good performance for Class 0 with precision of 0.85, recall of 0.94, and F1-score of 0.89. For Class 1, precision is high at 0.82, but recall drops to 0.62, resulting in an F1-score of 0.71. This indicates that while the model is still good at predicting true positives for Class 1, it misses more actual instances of Class 1 compared to the training data.
- Accuracy: The overall accuracy on the testing data is 0.84, consistent with the training data accuracy. This shows that the model generalizes well to new data and is not overfitting.
- Macro and Weighted Averages: The macro average recall is slightly lower at 0.78, reflecting the drop in recall for Class 1. The weighted averages remain consistent, indicating balanced performance.

5. Feature Importance & Summary

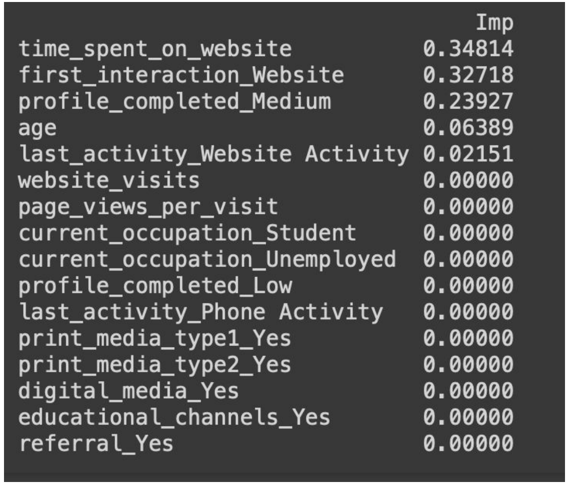
Feature importance of tuned decision tree

1. Root Split: The first interaction being on the website (first_interaction_Website <= 0.5) is the most significant split
2. Important Features:
   - ○ time_spent_on_website: high importance in predicting conversions.
   - ○ profile_completed and last_activity_Website Activity also play significant roles
3. Class Distribution:
   - ○ Blue Leaves: Represent converted leads (Class 1).
   - ○ Orange Leaves: Represent not converted leads (Class 0).
   - ○ Darker colors indicate a higher number of observations in the leaf.
4. Model Insights:
   - ○ Low entropy nodes indicate homogenous class distributions.
   - ○ Clear decision rules show that less time spent on the website often leads to not converted status.
5. Interpretability: The tree visualization highlights the decision-making process, providing transparency into which features drive predictions.

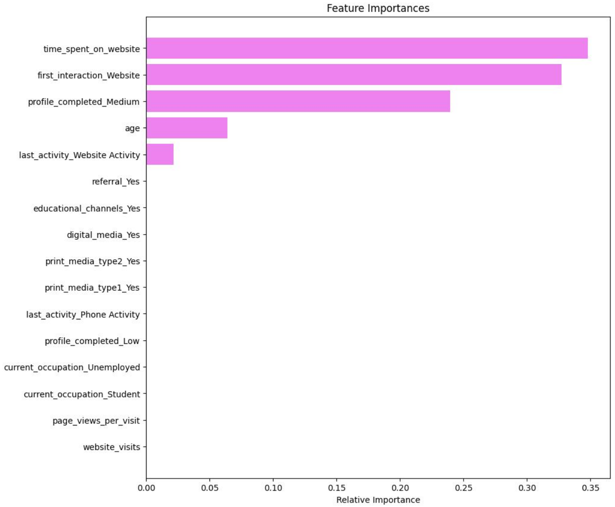


Feature importance of tuned decision tree

- Time spent on website, first interaction website, profile completed medium, age, and last activity website activity are all top features
- Whereas, features like website visits, page views per visit, and current occupation have no importance



```
                                      Imp
time_spent_on_website             0.34814
first_interaction_Website         0.32718
profile_completed_Medium          0.23927
age                               0.06389
last_activity_Website Activity 0.02151
website_visits                    0.00000
page_views_per_visit              0.00000
current_occupation_Student        0.00000
current_occupation_Unemployed     0.00000
profile_completed_Low             0.00000
last_activity_Phone Activity      0.00000
print_media_type1_Yes             0.00000
print_media_type2_Yes             0.00000
digital_media_Yes                 0.00000
educational_channels_Yes          0.00000
referral_Yes                      0.00000
```



Feature importance of tuned random forest

- Similar to the decision tree model, time spent on website, first_interaction_website, profile_completed, and age are the top four features that help differentiate between not converted and converted leads.
- Unlike the decision tree, the random forest gives some importance to other factors such as the occupation, page_views_per_visit. Therefore, the random forest is the better model since this shows that the random forest is giving importance to more factors in comparison to the decision tree.