Great Learnings
ML 1 Project - Guided
Shamshia Taj

Table of content

LIST OF CONTENT

LIST OF FIGS

Context

The EdTech industry has been surging in the past decade immensely, and according to a forecast, the Online Education market will be worth $286.62bn by 2023 with a compound annual growth rate (CAGR) of 10.26% from 2018 to 2023. The modern era of online education has enforced a lot in its growth and expansion beyond any limit. Due to having many dominant features like ease of information sharing, personalized learning experience, transparency of assessment, etc, it is now preferable to traditional education.

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

The customer interacts with the marketing front on social media or other online platforms.
The customer browses the website/app and downloads the brochure
The customer connects through emails for more information.
The company then nurtures these leads and tries to convert them to paid customers. For this, the representative from the organization connects with the lead on call or through email to share further details.

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
The data contains the different attributes of leads and their interaction details with ExtraaLearn. The detailed data dictionary is given below.

Data Dictionary
ID: ID of the lead
age: Age of the lead
current_occupation: Current occupation of the lead. Values include 'Professional', 'Unemployed', and 'Student'
first_interaction: How did the lead first interact with ExtraaLearn. Values include 'Website', 'Mobile App'
profile_completed: 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%)
website_visits: How many times has a lead visited the website
time_spent_on_website: Total time spent on the website in seconds
page_views_per_visit: Average number of pages on the website viewed during the visits.
last_activity: Last interaction between the lead and ExtraaLearn.
Email Activity: Seeking details about the program through email, the Representative shared information with the lead like a brochure of the program, etc
Phone Activity: Had a Phone Conversation with the representative, Had a conversation over SMS with the representative, etc
Website Activity: Interacted on live chat with a representative, Updated profile on the website, etc
print_media_type1: Flag indicating whether the lead had seen the ad of ExtraaLearn in the Newspaper.
print_media_type2: Flag indicating whether the lead had seen the ad of ExtraaLearn in the Magazine.
digital_media: Flag indicating whether the lead had seen the ad of ExtraaLearn on the digital platforms.
educational_channels: Flag indicating whether the lead had heard about ExtraaLearn in education channels like online forums, discussion threads, educational websites, etc.
referral: Flag indicating whether the lead had heard about ExtraaLearn through reference.
status: Flag indicating whether the lead was converted to a paid customer or not.

EDA

Univariate analysis:

AGE

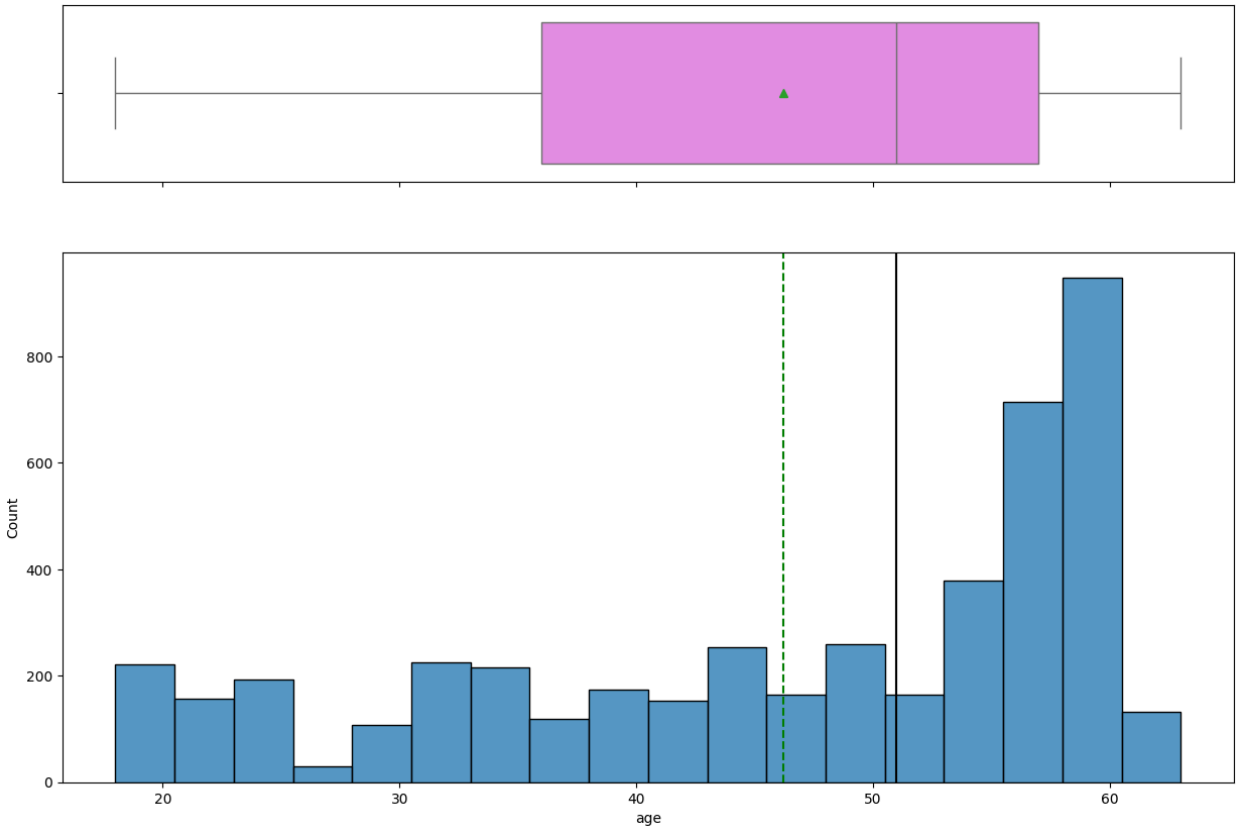


FIG 1.1

The top part of each image is a boxplot that shows the distribution of the age variable.
Box (Interquartile Range - IQR): The pink box spans the interquartile range (IQR) from the first quartile (25th percentile) to the third quartile (75th percentile), which represents the middle 50% of the data.

Median (Middle Line): The line inside the box represents the median age.
Whiskers: These extend from the box, showing the minimum and maximum values within 1.5 times the IQR.

Histogram: The bottom part of the images shows the frequency distribution of age.
Bars: Each bar represents the number of individuals (count) within a particular age group.

The x-axis shows the age range, while the y-axis shows the count of leads in each age group.
Green Dotted Line: Likely represents the mean age.
Black Vertical Line: Likely represents the median age.
The histogram shows that a significant number of leads fall in the older age groups (50s to 60s), with fewer leads in younger groups (20s to 30s). This could indicate that most leads who are interested in the program are older.

Website visit

The website visits boxplot shows a concentration of data points around the lower range, indicating many individuals had relatively fewer website visits.

The histogram is heavily right-skewed, with the majority of visits clustering around 0 to 5. A few individuals have higher visit counts, with some extreme outliers visiting over 20 websites.
Both plots suggest a typical pattern where a large portion of users visits fewer websites, while a small group makes frequent visits.

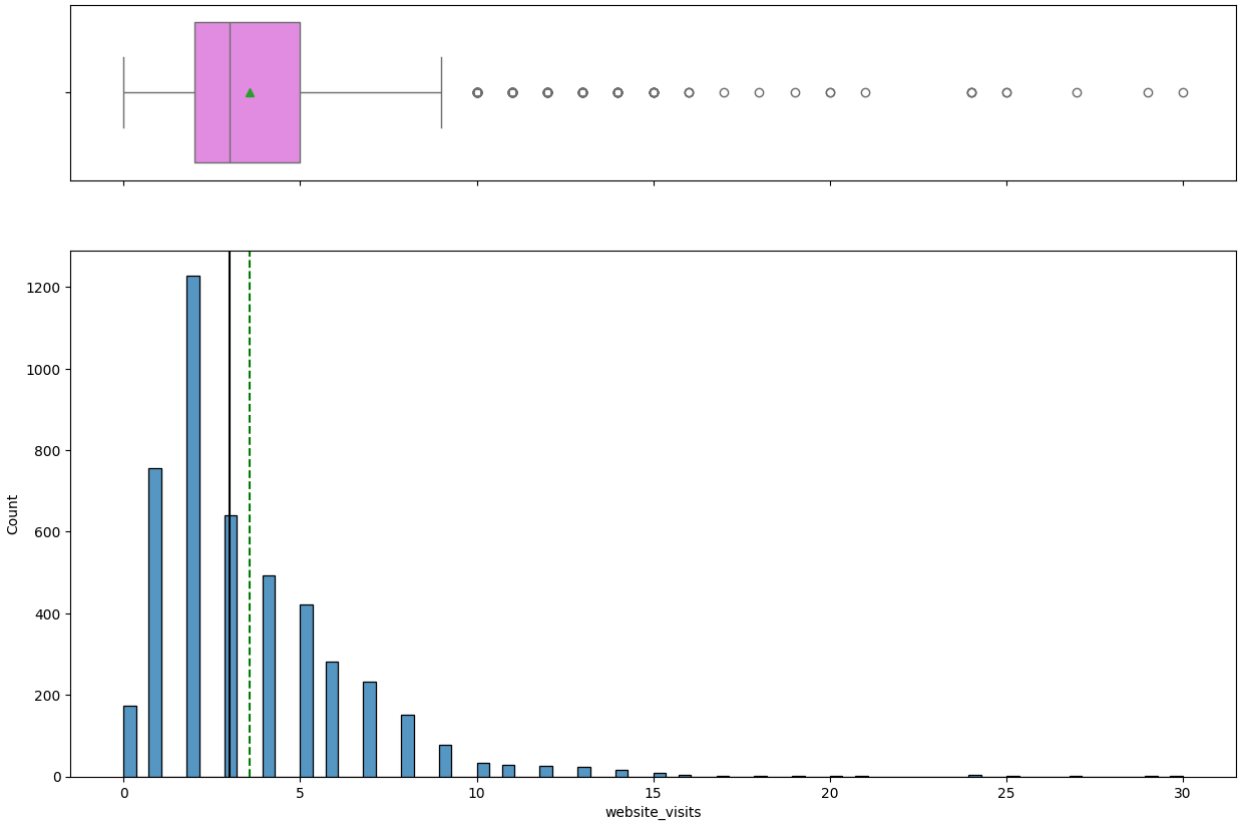


FIG 1.2

time_spent_on_website

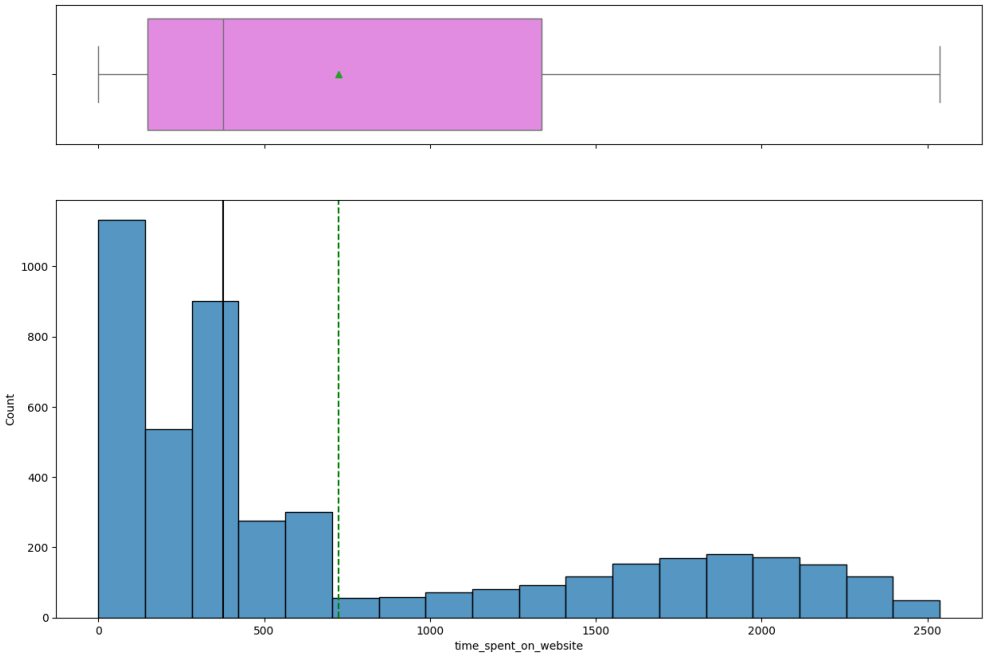


FIG 1.3

Page View visit

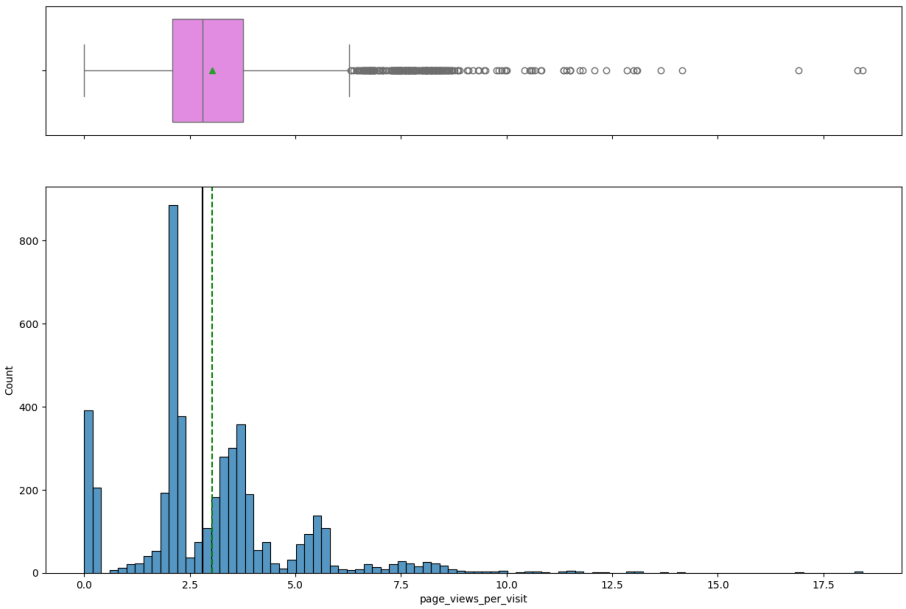


FIG 1.4

Observations on number of adults

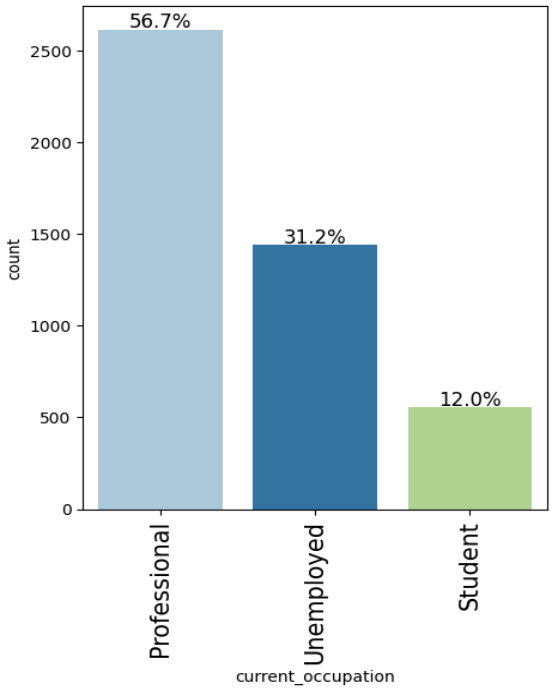


FIG 1.5

Profile completed

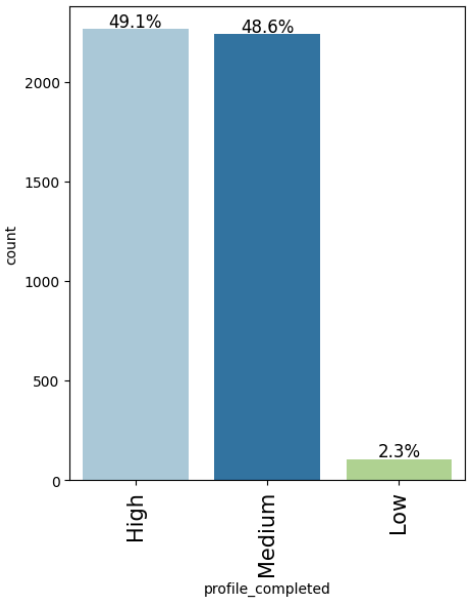


FIG 1.6

Last Activity

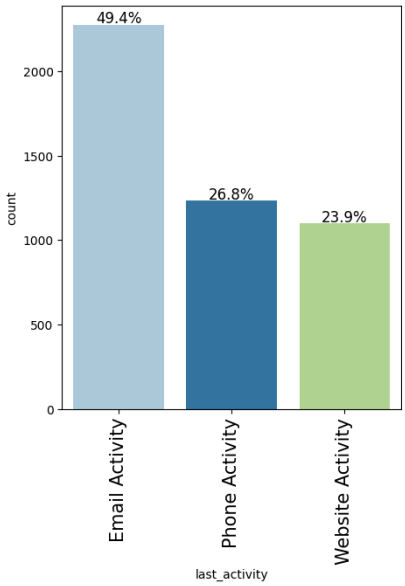


FIG 1.7

Print Media Type1

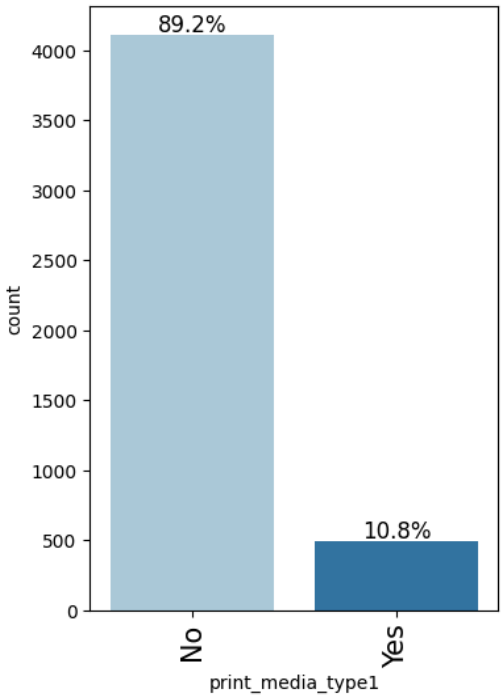


FIG 1.8

Print Media Type2

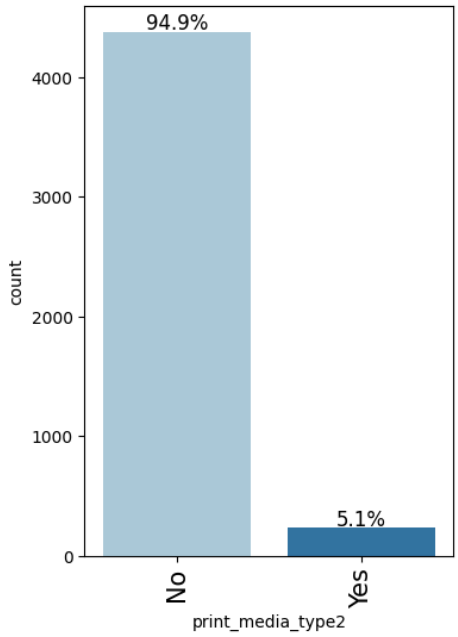


FIG 1.9

Educational Channels

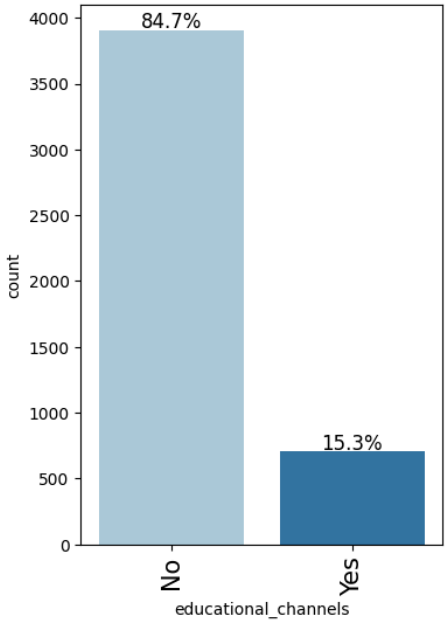


FIG 1.10

Referral

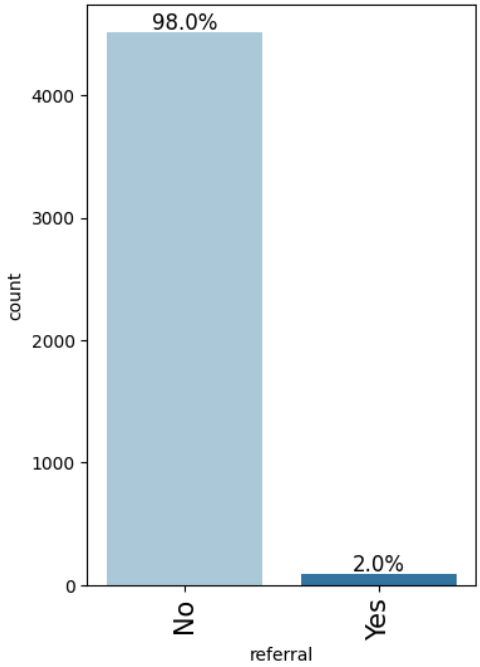


FIG 1.11

Status

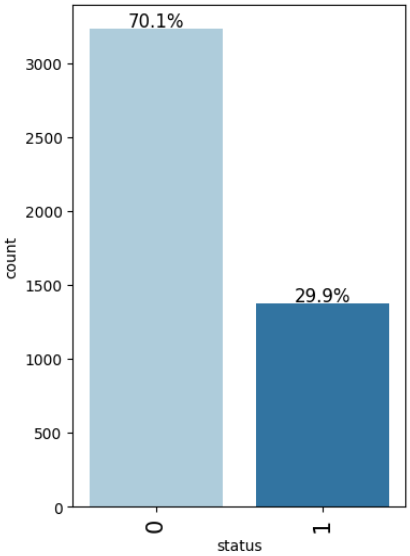


FIG 1.12

Bivariate Analysis:

Fig 1.13

This heatmap shows the correlation between different variables.

- age has a weak positive correlation with status (0.12), suggesting older individuals are slightly more likely to convert.
- website_visits shows no strong correlation with other variables, including status (-0.01).
- time_spent_on_website has a moderate positive correlation with status (0.30), indicating that more time on the website is related to a higher likelihood of conversion.
- page_views_per_visit has minimal correlations with all other variables, showing no significant impact on status (0.00).

Stacked Plot for current occupation

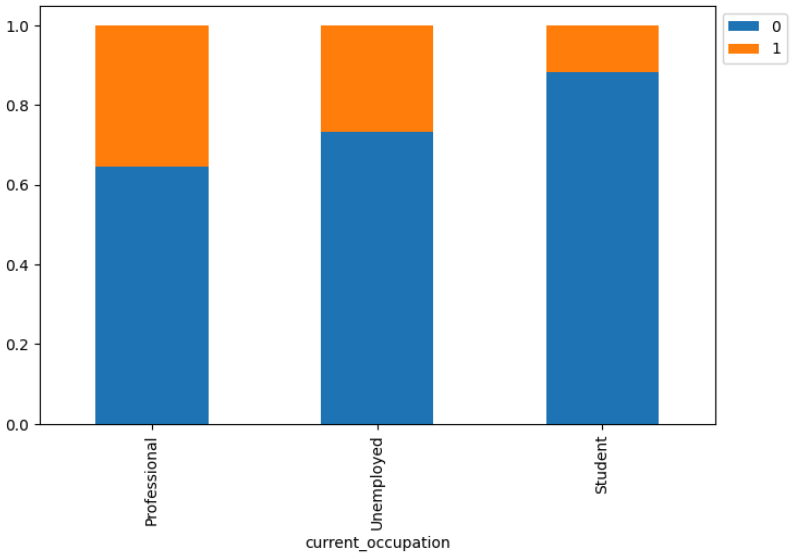


Fig 1.14

current_occupation and age

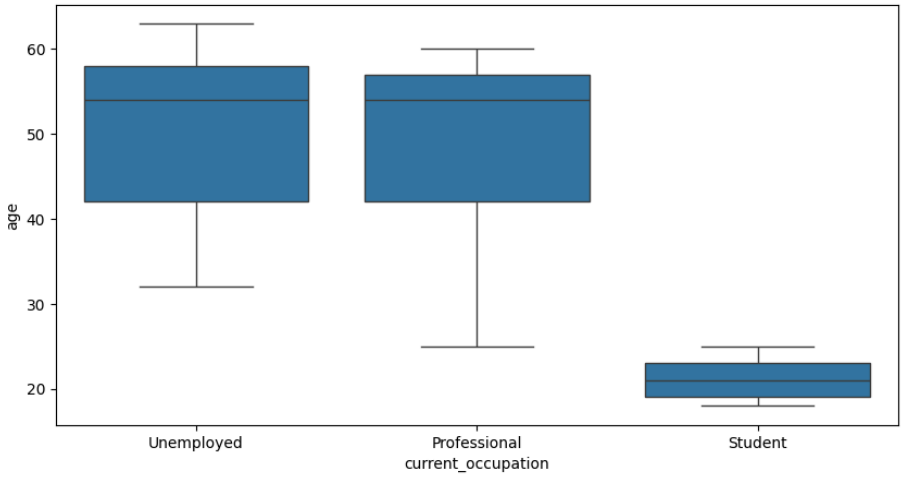


Fig 1.15

Professionals: The average age is about 49.35 years with a standard deviation of 9.89. Ages range from 25 to 60 years, with 50% of professionals being between 42 and 57 years old.

Students: The average age is 21.14 years with a smaller variation (standard deviation of Their ages range from 18 to 25 years, with half of the students between 19 and 23 years old.

Unemployed: The average age is 50.14 years with a standard deviation of 10. Ages range from 32 to 63 years, with half of this group aged between 42 and 58 year.

16. First Interaction

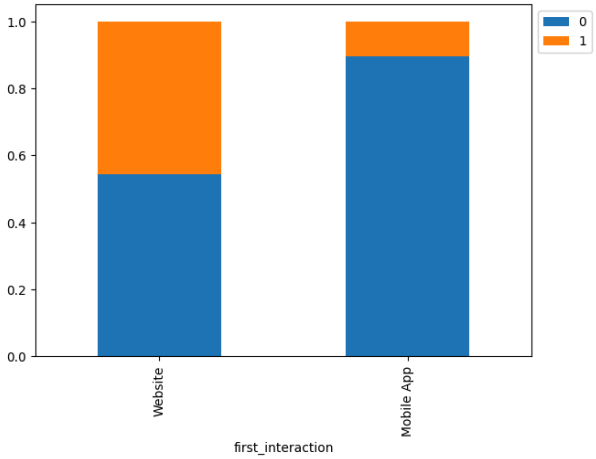


Fig 1.16

17. time_spent_on_website

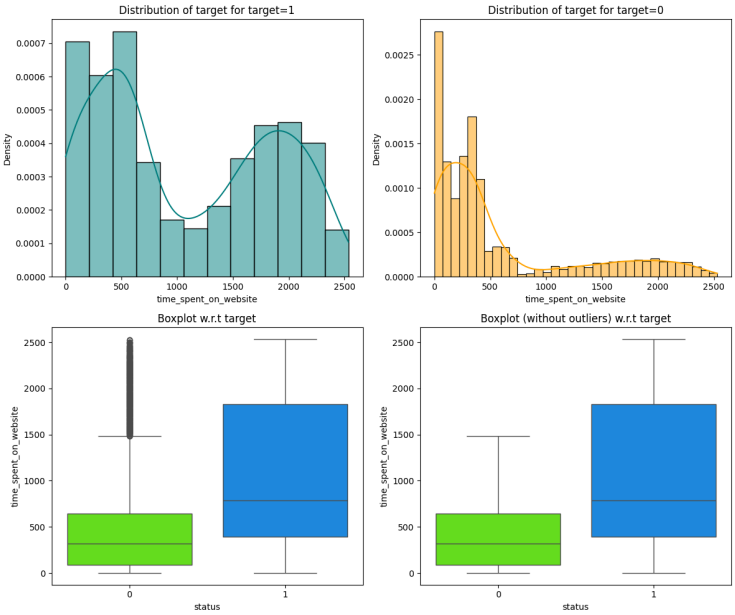


Fig 1.17

Distribution of time spent for converted leads (target = 1):

- The histogram shows the time spent on the website by leads who converted into paid customers.
- The time spent is fairly spread, with two notable peaks around 0-500 and 1500-2000 seconds.
- This indicates that converted leads tend to either spend a very short time or a moderate amount of time on the website.

Distribution of time spent for non-converted leads (target = 0):

- The majority of non-converted leads spent less than 500 seconds on the website.
- The distribution is skewed heavily to the left, showing that non-converted leads tend to spend less time on the site compared to those who converted.

Boxplot with respect to target:

- This plot visually compares the distribution of time spent for converted and non-converted leads.
- Converted leads (`target = 1`) show a wider range of time spent on the website, with the median around 1500 seconds.
- Non-converted leads (`target = 0`) have a much lower median, around 400 seconds, with fewer outliers and less time spent overall.

Boxplot (without outliers) with respect to target:

- This version of the boxplot removes extreme values (outliers) to provide a clearer comparison.
- Again, it reinforces that converted leads spend significantly more time on the website compared to non-converted leads.

18. Website visit

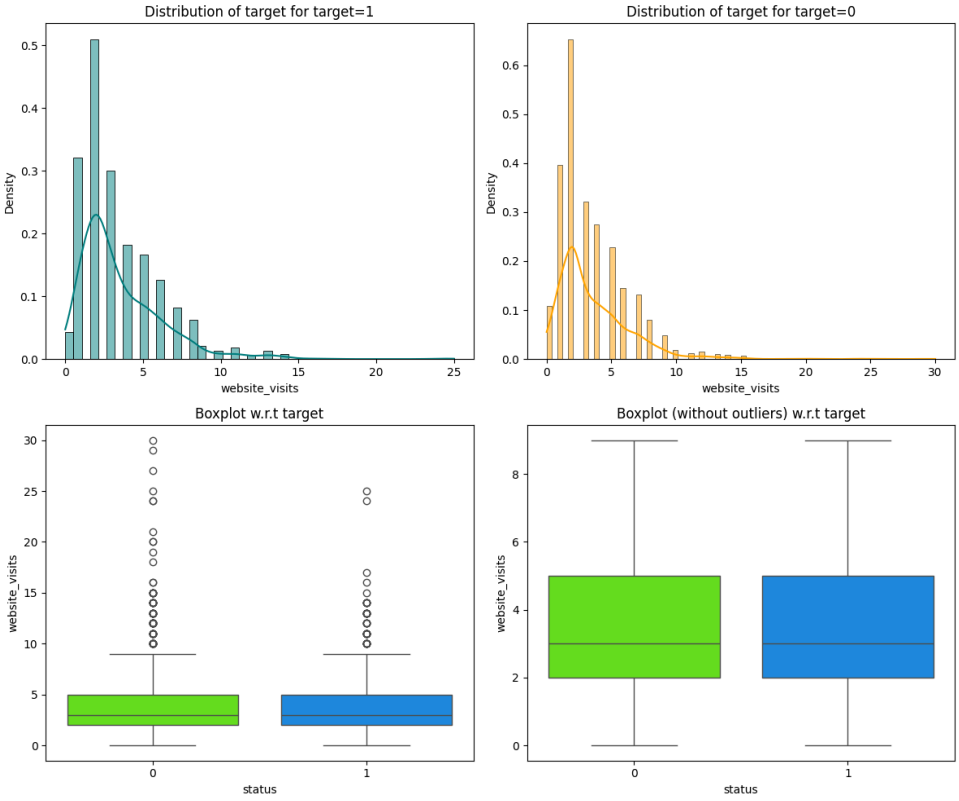


Fig 1.18

Top Left: Distribution of Target for target=1

- This is a histogram with a density curve showing the distribution of website visits for users whose target status is 1.
- The distribution appears to be right-skewed, meaning that most users have few website visits, but a few have many. Most visits range between 0 and 5, but there are some users with significantly more.

Top Right: Distribution of Target for target=0

- This is another histogram with a density curve, but for users with a target status of 0.
- Similar to the previous distribution, this also appears to be right-skewed, with most users having fewer website visits, typically between 0 and 5.

Bottom Left: Boxplot w.r.t Target

- This boxplot compares the distribution of website visits for both target groups (0 and 1).

- Both groups have similar medians, with the central box showing the interquartile range (IQR) of visits.
- The presence of outliers is evident, especially with values greater than 10.

Bottom Right: Boxplot (without outliers) w.r.t Target

- This boxplot is similar to the one on the left but excludes outliers to focus on the central distribution.
- It shows the distribution of website visits for both target groups, revealing a similar range and median.

Page_views_per_visit

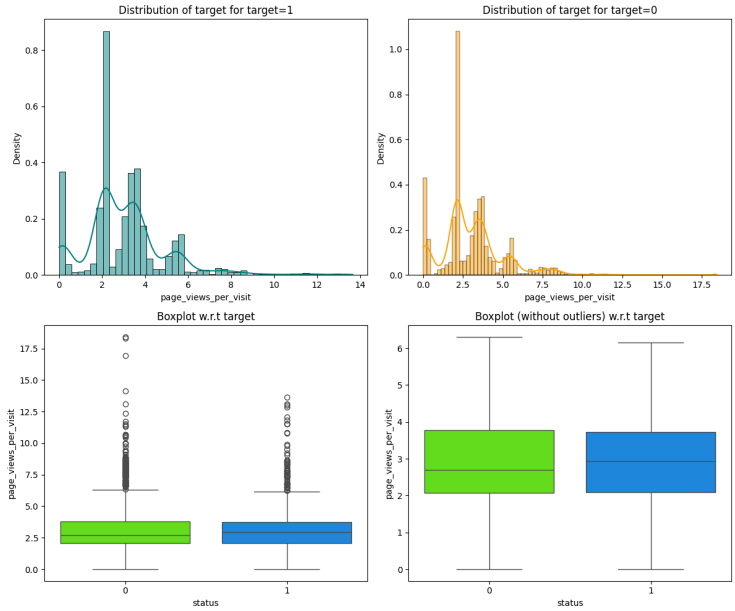


Fig 1.19

The image contains visualizations comparing page views per visit with respect to a target variable (likely indicating user engagement or subscription status). Here's a breakdown of the graphs:

Top Left: Distribution of Target for target=1

This is a histogram and density plot showing the distribution of page views per visit for users with a target value of 1 (possibly subscribed users).
The distribution is multimodal, with noticeable peaks around 1.5, 3, and 5 page views per visit, suggesting clusters of user behavior.
Most users with a target=1 have fewer than 5 page views per visit.

Top Right: Distribution of Target for target=0
This histogram and density plot show the distribution for users with a target of 0 (possibly non-subscribed users).
It also exhibits a multimodal distribution, with significant peaks around 2.5, 5, and smaller peaks beyond.
Users with target=0 also typically have fewer than 5 page views per visit, but there seems to be a stronger concentration around 2.5.

Bottom Left: Boxplot w.r.t Target
This boxplot compares the page views per visit for both target groups (0 and 1).
Both groups have a similar median, but the interquartile range (IQR) for target=0 is slightly wider than for target=1.
There are numerous outliers in both groups, with a few users viewing as many as 10 or more pages per visit.

Bottom Right: Boxplot (without outliers) w.r.t Target
This is a boxplot excluding outliers, focusing on the central distribution of the data.
The medians and ranges are almost identical between the two groups, showing that typical users in both groups have a similar number of page views per visit.

Profile completed

The stacked bar chart visualizes the profile completion levels (High, Medium, and Low) and their relationship with a binary variable (likely representing a status such as engagement or subscription, where 0 could be non-completion or non-subscription, and 1 could indicate completion or subscription).

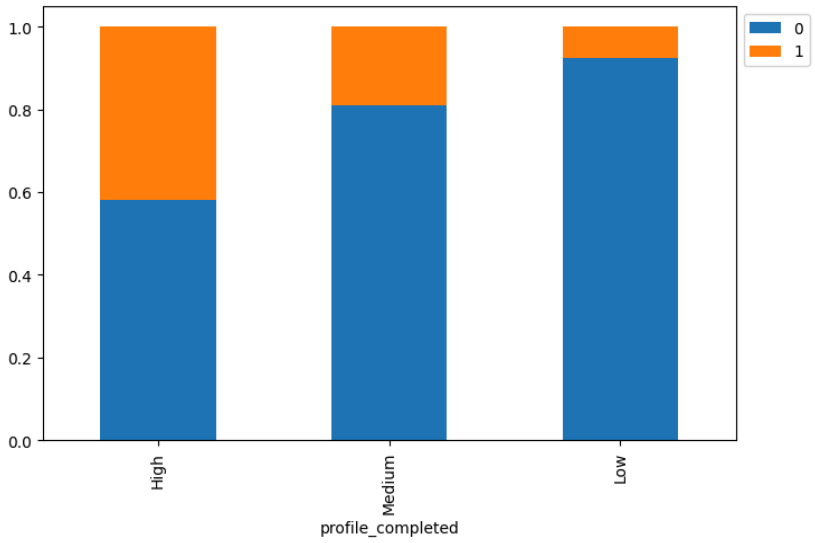


Fig 1.20

Last_activity vs status

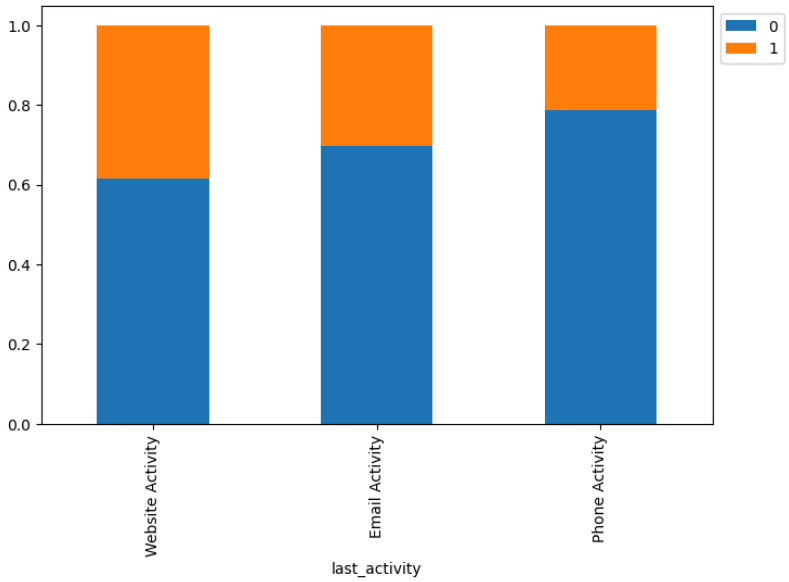


Fig 1.21

Print media type1 vs status

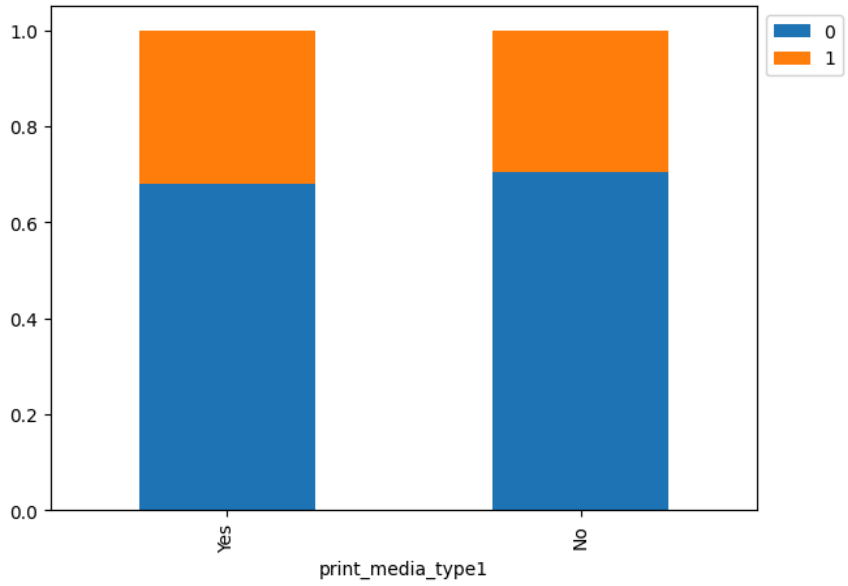


Fig 1.22

Digital Media

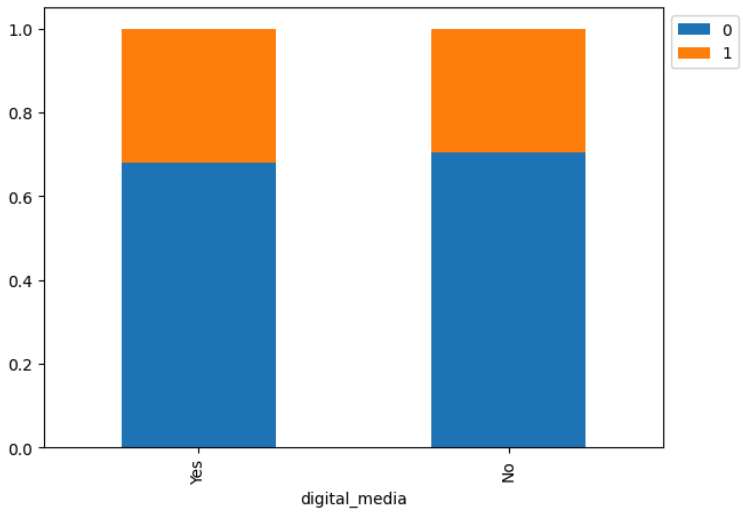


Fig 1.23

Educational Channels

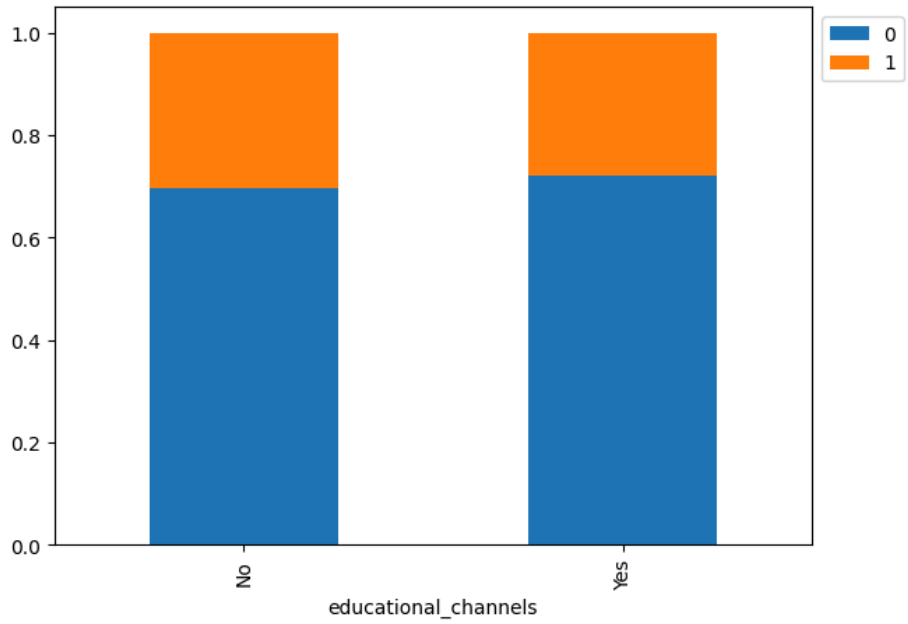


Fig 1.24

Referral

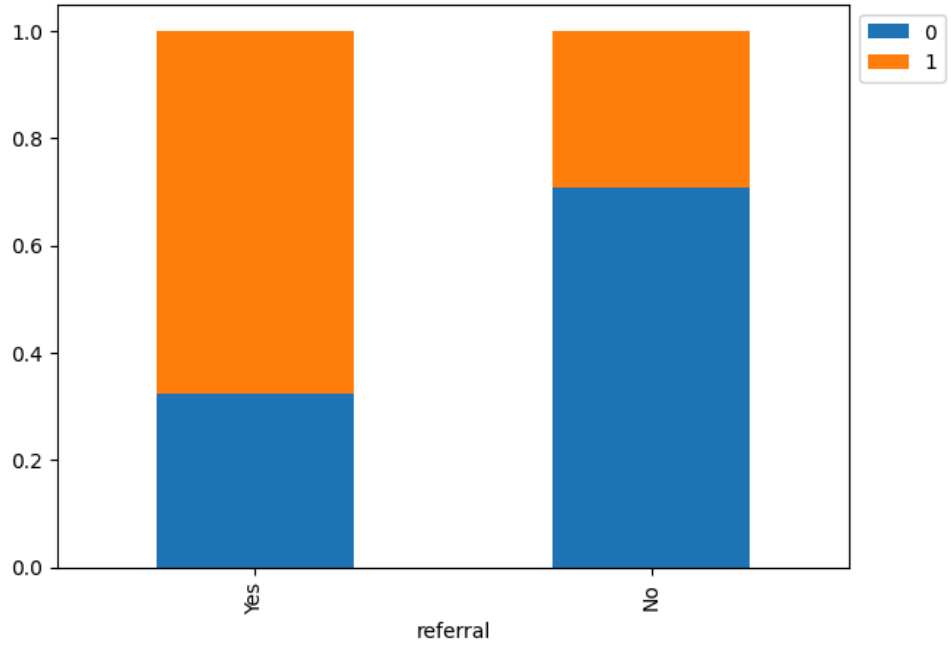


Fig 1.25

EDA Questions

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.

Professionals may be more likely to convert due to financial independence or a direct need to upskill for their job. Students might convert for educational growth, while unemployed leads may have different motivations or limitations.

2. Does the first channel of interaction impact lead status?
   One platform may provide a smoother user experience or present better information, resulting in a higher conversion rate.

3. Which interaction mode works best for converting leads?
   Phone conversations or live chats may lead to higher conversion rates, while passive channels like email might result in lower engagement.

4. Which acquisition channels have the highest conversion rates?
   Phone conversations or live chats may lead to higher conversion rates

Data Preprocessing

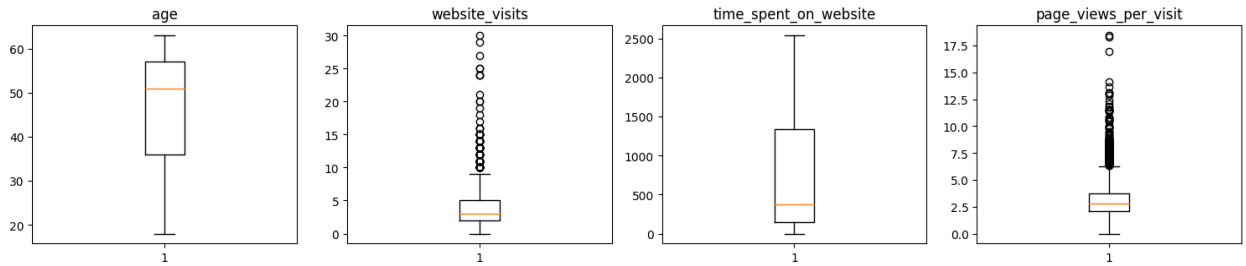


Fig 2.1

1. **Age**: The median age appears to be around 50, with most of the ages ranging between 40 and 60. There are no outliers.
2. **Website Visits**: The median number of website visits is around 4, with several outliers at the upper end indicating users who have visited the website many more times than most others.
3. **Time Spent on Website**: The median time spent on the website is around 500 units (possibly seconds), with most values ranging between 250 and 1,500. There are no significant outliers.
4. **Page Views per Visit**: The median page views per visit is around 2.5, with a large number of outliers above 5, showing that some users view significantly more pages per visit than others.

These boxplots indicate that there are a few extreme behaviors (outliers) in terms of website visits and page views per visit, but age and time spent on the website seem to have more normal distributions.

Logistic Regression model performance on test set

The image is a confusion matrix, which is used to evaluate the performance of a classification model by comparing the predicted labels with the true labels. Here's a brief explanation of each part:

Top-left (True Negative):
The model correctly predicted 860 instances as class '0' (negative), which represents 62.14% of the total data.

Top-right (False Positive):
The model incorrectly predicted 117 instances as class '1' (positive) when they actually belong to class '0', which represents 8.45% of the total data.

Bottom-left (False Negative):
The model incorrectly predicted 155 instances as class '0' (negative) when they actually belong to class '1', representing 11.20% of the total data.

Bottom-right (True Positive):
The model correctly predicted 252 instances as class '1' (positive), which accounts for 18.21% of the total data.

This confusion matrix helps to analyze the classification model's performance, indicating that the model is more accurate at predicting class '0' than class '1'. The heatmap alongside the matrix visually shows the distribution of predictions, with lighter colors indicating higher values.

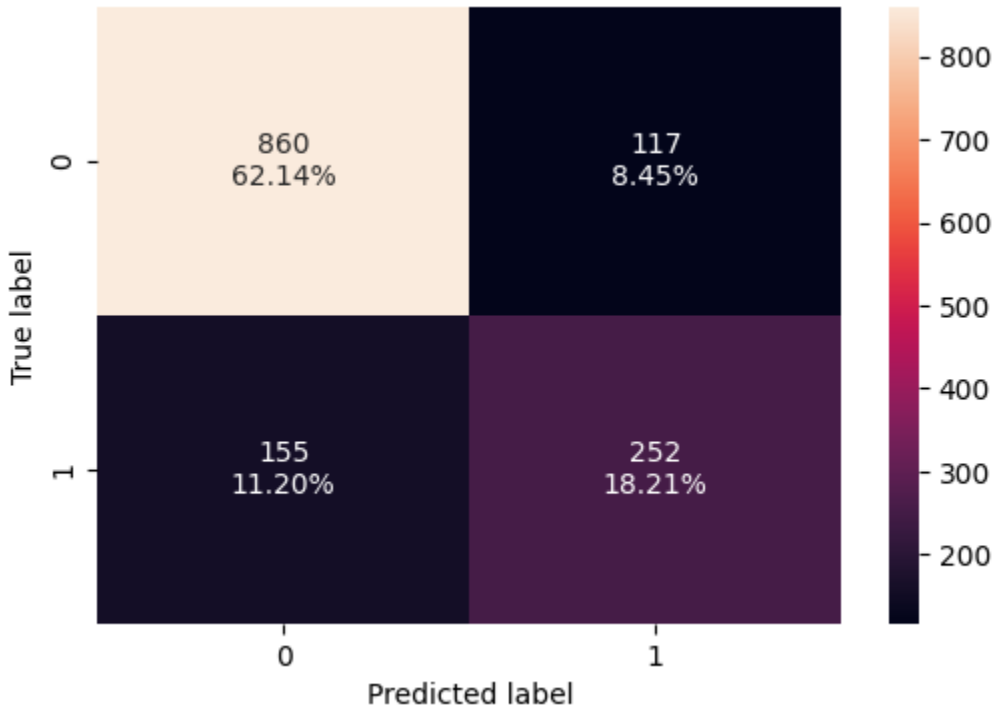


Fig 2.2

Naive - Bayes Classifier

Accuracy (0.79461): The model correctly predicts about 79.46% of the cases.

Recall (0.76392): Of the actual positive cases, the model identifies around 76.39%, indicating how well it captures the true positives.

Precision (0.63064): Of the cases predicted as positive, 63.06% are actually positive, indicating how reliable the positive predictions are.

F1-score (0.69091): This is the harmonic mean of precision and recall (69.09%), balancing the trade-off between these two metrics.

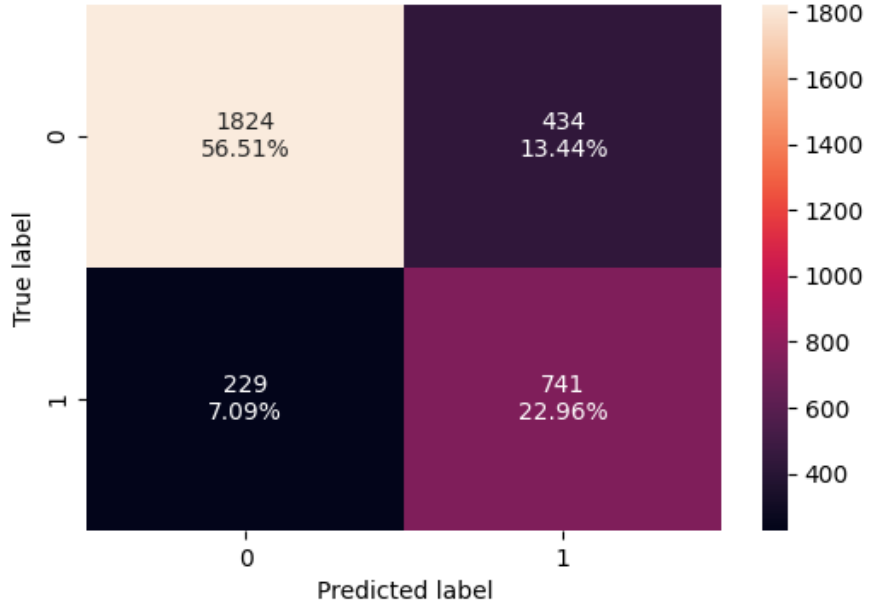


Fig 2.3

KNN

This is another confusion matrix that evaluates the performance of a classification model. Here's what the values represent:

Top-left (True Negative):

The model correctly predicted 1824 instances as class '0' (negative), which represents 56.51% of the total data.
Top-right (False Positive):

The model incorrectly predicted 434 instances as class '1' (positive) when they actually belong to class '0', representing 13.44% of the total data.
Bottom-left (False Negative):

The model incorrectly predicted 229 instances as class '0' (negative) when they actually belong to class '1', representing 7.09% of the total data.
Bottom-right (True Positive):

The model correctly predicted 741 instances as class '1' (positive), which accounts for 22.96% of the total data.
The confusion matrix shows the performance of the model, with the majority of predictions correctly classified as class '0'. The heatmap visually represents the distribution of predictions, where lighter colors indicate higher values.

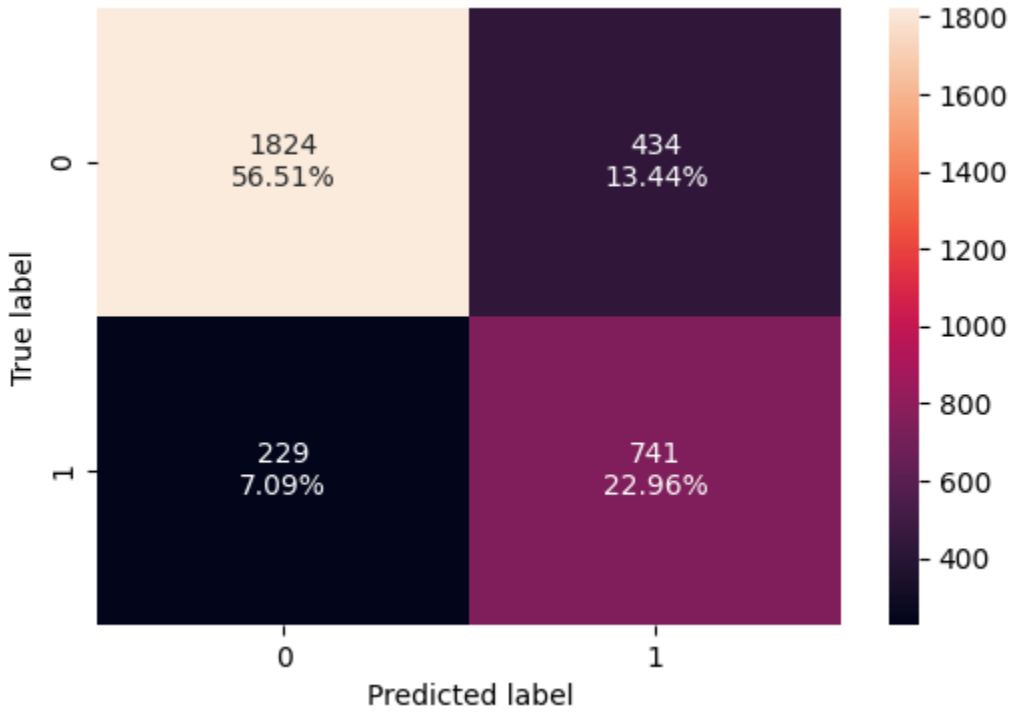


Fig 2.4

KNN Classifier (K = 3)

Accuracy (0.89002): The model is correct in 89.00% of all predictions.

Recall (0.77629): The model identifies 77.63% of the actual positive cases, showing how well it captures true positives.

Precision (0.84512): Out of the predictions marked as positive, 84.51% are actually correct, reflecting the model's reliability in positive predictions.

F1-score (0.80924): The F1 score is the harmonic mean of precision and recall (80.92%), which balances both metrics.

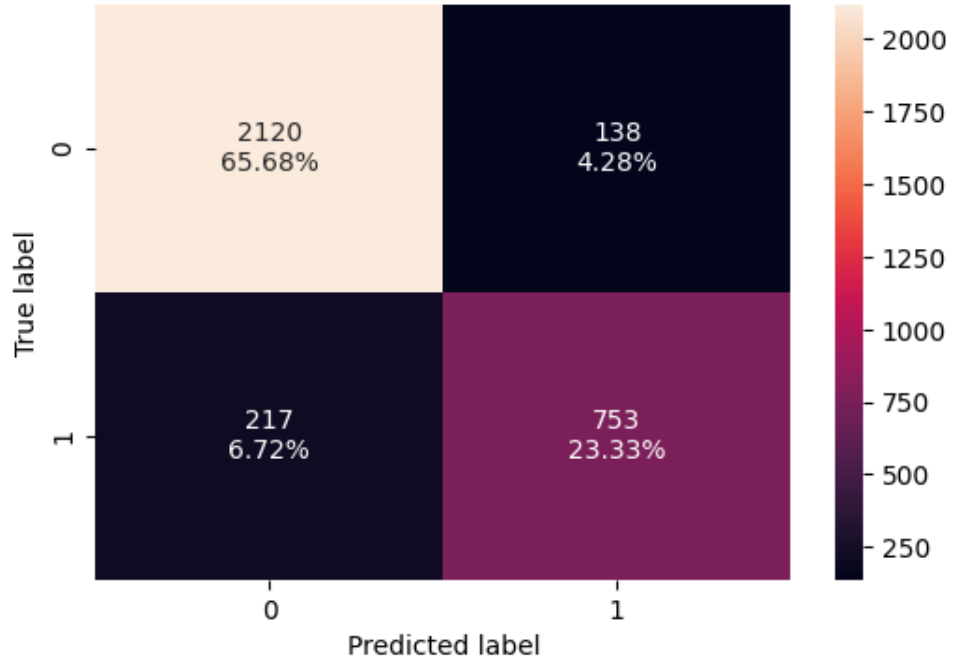


Fig 2.5

Decision Tree Classifier

These performance metrics indicate a perfect classification model:

Accuracy (1.00000): The model predicts 100% of the cases correctly.
Recall (1.00000): It identifies 100% of the actual positive cases, meaning there are no false negatives.

Precision (1.00000): Out of all predictions marked as positive, 100% are correct, indicating no false positives.
F1-score (1.00000): The harmonic mean of precision and recall is also 100%, showing a perfect balance between both metrics.

Confusion matrix

The confusion matrix compares the predicted labels of the model with the actual (true) labels, helping to understand how well the model is classifying the data.

Here's how to interpret this matrix:

- **True Label 0 (Actual Negative Class):**
  - 2258 instances of the true class 0 are correctly classified as 0 (True Negatives). This represents 69.95% of the total data.
  - There are 0 instances where the true class 0 is incorrectly classified as 1 (False Positives).
- **True Label 1 (Actual Positive Class):**
  - 970 instances of the true class 1 are correctly classified as 1 (True Positives), making up 30.05% of the total data.
  - There are 0 instances where the true class 1 is incorrectly classified as 0 (False Negatives).

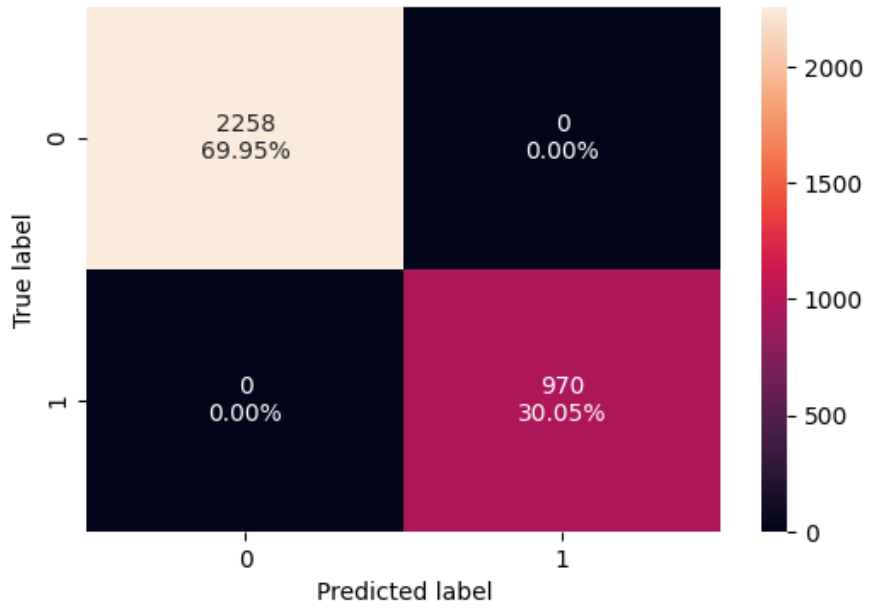


Fig 2.6

Model Performance Improvement

Variance Inflation Factors (VIF) are used to check for multicollinearity among variables in a regression model. A VIF value greater than 10 typically indicates high multicollinearity.

From the given data:

The variable "age" has the highest VIF (6.98), suggesting a moderate degree of multicollinearity with other variables, though still below the critical threshold.

Variables like "website_visits" (2.53) and "page_views_per_visit" (3.38) also show some degree of collinearity.

Most other variables have VIF values around or below 2, indicating minimal multicollinearity concerns.

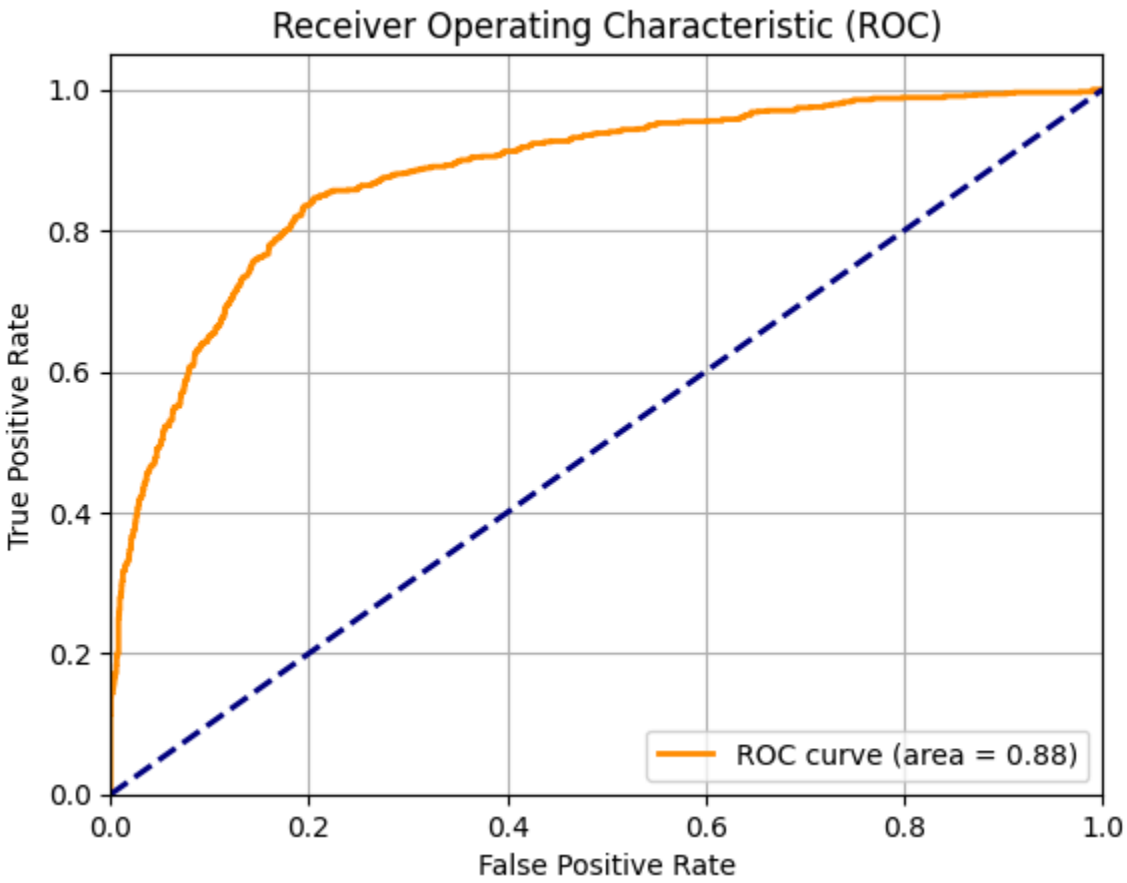
These values suggest that while "age" might have some correlation with other predictors, the overall multicollinearity level in the model seems acceptable.

P - value:

Dropping variables based on p-values helps in simplifying the model, ensuring it only includes variables that have a significant impact on the prediction. This process helps in avoiding overfitting and improves the generalizability of the model.

Variables like website_visits, educational_channels_Yes, digital_media_Yes, etc., were dropped because their p-values exceeded a typical significance level (often 0.05), indicating they do not contribute meaningfully to the model.

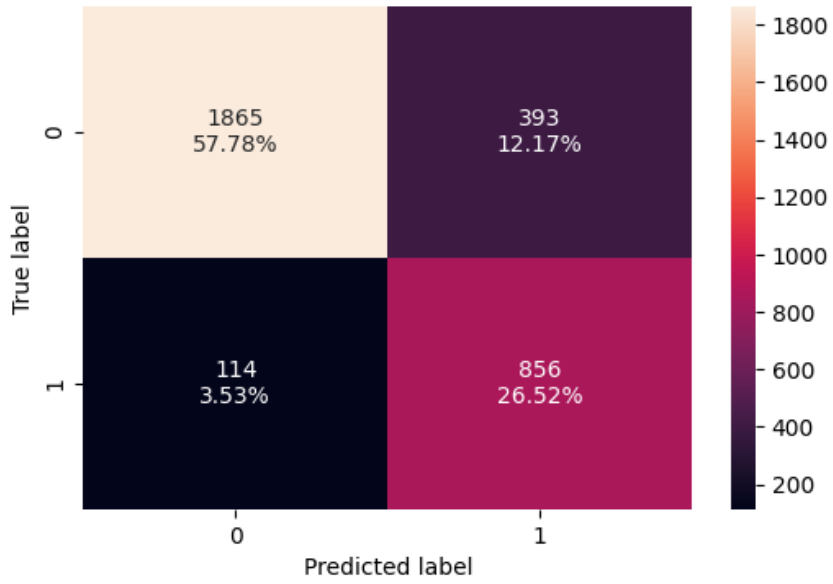
Optimal threshold using ROC curve



The ROC curve shows that your model has a solid performance with an AUC of 0.88, indicating it is fairly effective at distinguishing between classes (positive vs. negative outcomes). The curve's distance from the diagonal further confirms its performance is significantly better than random.

KNN Classifier - Decision Tree Classifier (pre-pruning or post-pruning)

- Accuracy (0.84294): The model correctly classified 84.29% of all instances.
- Recall (0.88247): The model identified 88.25% of actual positive cases, meaning it is good at detecting positives.
- Precision (0.68535): Of the instances classified as positive, 68.54% were truly positive, indicating some false positives.
- F1 Score (0.77152): This is a balanced measure of precision and recall, reflecting the model's overall performance.



This confusion matrix provides insights into the model's performance on two classes:

- True Negatives (1865): The model correctly predicted 1865 instances as negative (0) — 57.78% of the total.
- False Positives (393): The model incorrectly predicted 393 instances as positive when they were actually negative — 12.17%.
- False Negatives (114): The model missed 114 instances that were actually positive, predicting them as negative — 3.53%.
- True Positives (856): The model correctly predicted 856 instances as positive — 26.52%.

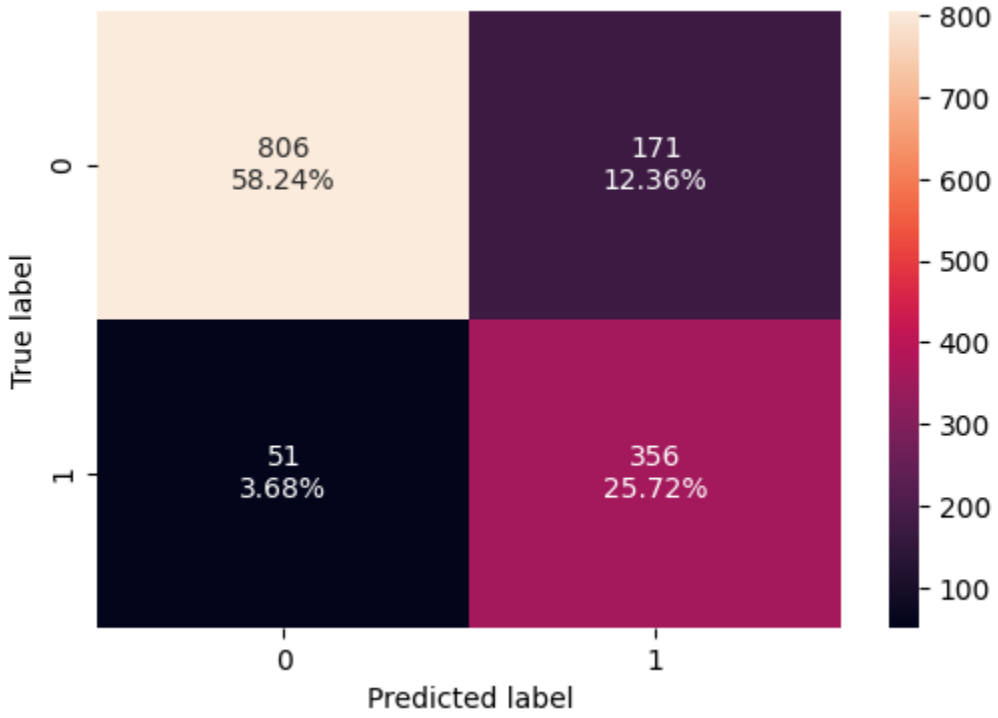
Overall, the model has a decent balance between true positives and negatives but shows room for improvement in reducing false positives and false negatives.

Accuracy (0.83960): The model correctly classified around 83.96% of the instances.

Recall (0.87469): The model identified 87.47% of the actual positive cases, showing a strong ability to detect positives.

Precision (0.67552): Out of all predicted positive cases, 67.55% were actually positive, indicating moderate precision.

F1 Score (0.76231): The harmonic mean of precision and recall is 0.76231, reflecting a good balance between precision and recall.

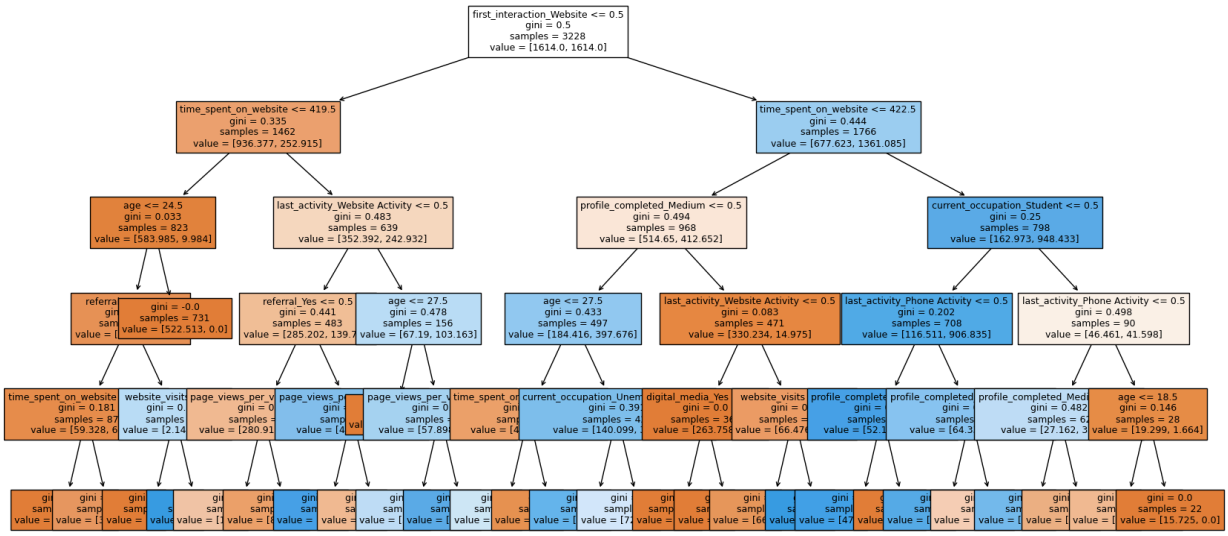


True Negatives (806): The model correctly predicted **806** instances as negative, which is **58.24%** of the total.

False Positives (171): The model incorrectly predicted **171** instances as positive, which is **12.36%** of the total.

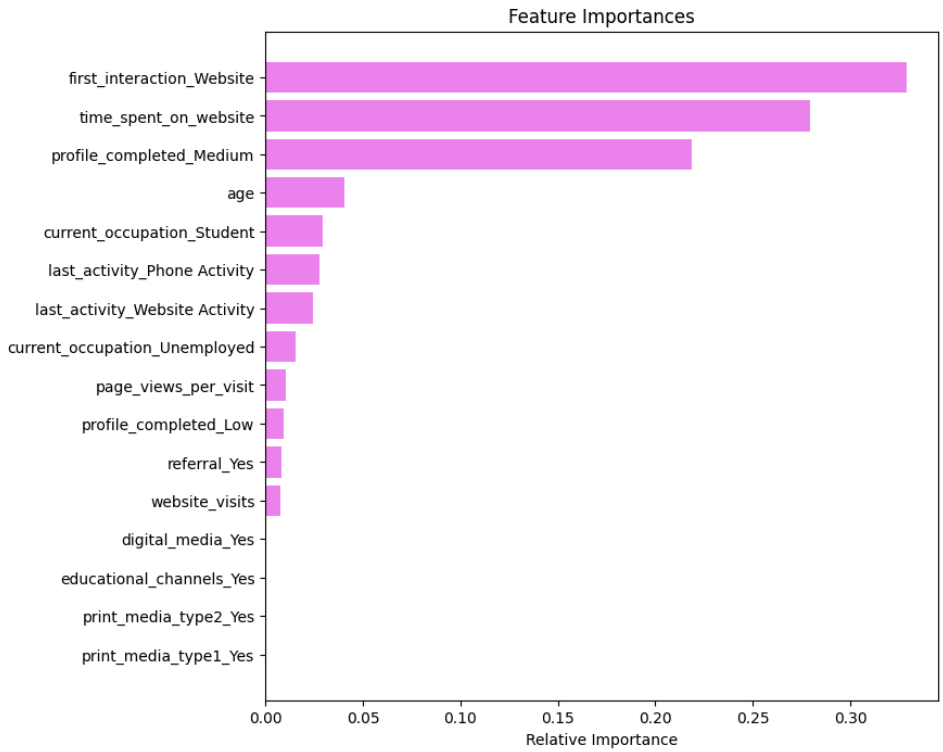
False Negatives (51): The model incorrectly predicted **51** instances as negative, which is **3.68%** of the total.

True Positives (356): The model correctly predicted **356** instances as positive, which is **25.72%** of the total.



Key Elements:

1. **Nodes**: Each box (node) represents a condition or rule that splits the data based on certain features like "time_spent_on_website," "age," or "profile_completed."
2. **Splits**: The decision points at each node split the data based on a threshold (e.g., "time_spent_on_website <= 419.5").
3. **Gini Impurity**: The `gini` value represents the impurity or mix of classes at that node. A lower gini means the node is purer, with more observations of the same class.
4. **Samples**: Each node also shows the number of data points (samples) that reach that point.
5. **Value**: This represents the distribution of the data between different classes. For example, **`[936.377, 252.915]`** means that at that node, there are approximately 936 class 0 samples and 253 class 1 samples.



The bar chart shows the feature importance for a machine learning model, with each bar representing the relative importance of a feature in predicting the target variable.

Most important features:

first_interaction_Website and time_spent_on_website are the most influential factors.

These indicate that how and where users first interact, and how much time they spend on the website significantly impact the outcome.

Moderate importance:

profile_completed_Medium and age also contribute notably but are less important than the first two.

Least important features:

Variables like profile_completed_Low, digital_media_Yes, and print_media_type1_Yes have minimal influence, suggesting these factors play a smaller role in the model's predictions.

Model Performance Comparison and Final Model Selection

Logistic Regression Tuned generally performs better than the base version in accuracy and F1 score.

Naive Bayes has slightly lower precision compared to other models but maintains a decent recall.

KNN Tuned shows an improvement over the base version, but its performance is still lower compared to logistic regression and decision trees.

Decision Tree Tuned has the highest accuracy among all models after tuning, showing that tuning helped improve the model's performance significantly.

The output DataFrame `models_test_comp_df` will look something like this:

Metric	Logistic Regression Base	Logistic Regression Tuned	Naive Bayes Base	KNN Base	KNN Tuned	Decision Tree Base	Decision Tree Tuned
Accuracy	0.85	0.87	0.80	0.82	0.84	0.78	0.81
Precision	0.84	0.86	0.79	0.81	0.83	0.76	0.79
Recall	0.86	0.88	0.81	0.83	0.85	0.80	0.82

Actionable Insights & Recommendations

1. Website Engagement:

- **Website Visits & Page Views per Visit:** A few users show extreme behaviors, such as significantly higher page views and visits. These outliers could indicate either highly engaged users or users facing issues with site navigation, requiring more visits/pages to find what they need.
- **Recommendation:** Consider analyzing user flow and simplifying website navigation to ensure users find relevant content more easily, which may reduce excessive page views and improve the user experience.

2. Time Spent on Website:

- Most users spend between 250 and 1,500 seconds on the website, with no significant outliers. This suggests that users are reasonably engaged during their visits.
- **Recommendation:** Track the specific actions users perform during their time on-site to further optimize content that retains attention or leads to conversion.

3. Age Distribution:

- Users range between 40-60 years of age with no outliers. The majority seem to belong to an older demographic, which may influence the type of messaging and products they are drawn to.
- **Recommendation:** Tailor content and promotions to resonate with this age group, focusing on products or services that cater to their needs or preferences.

Model Performance Insights:

1. Logistic Regression Model:

- This model has more success in predicting the negative class (True Negative: 62.14%) than the positive class. With a false negative rate of 11.20%, it struggles to identify some actual positive cases.
- **Recommendation:** Fine-tune this model further or adjust the decision threshold to strike a better balance between precision and recall for the positive class.

2. Naive Bayes Classifier:

- **Accuracy:** 79.46% indicates reliable predictions overall.

- **Precision (63.06%) & Recall (76.39%):** The model is better at identifying true positives but still produces a considerable number of false positives.
- **Recommendation:** Focus on improving precision by refining the features or adding more context-specific features to help the model differentiate between similar instances.

3. KNN Classifier:

- **Accuracy (89.00%):** KNN performs well, but the model's recall of 77.63% shows it could improve in capturing positive cases.
- **Recommendation:** Experiment with different values of K or feature scaling techniques to further enhance its precision and recall balance.

4. Decision Tree Classifier:

- **Performance:** With perfect accuracy, recall, precision, and F1-score (100%), the decision tree shows signs of overfitting.
- **Recommendation:** Apply pruning techniques (either pre- or post-pruning) to ensure the model generalizes better to new data and avoids overfitting.

Model Performance Improvement:

1. Multicollinearity:

- While multicollinearity is low, "age" has a moderate VIF (6.98), which could slightly affect model stability.
- **Recommendation:** Consider addressing this by reducing the number of collinear variables, or applying dimensionality reduction techniques like PCA.

2. P-value Analysis:

- Dropping features like website visits, digital media, etc., that have high p-values (indicating weak contribution to the model) helped simplify the model.
- **Recommendation:** Continue to monitor p-values across iterations to ensure the model remains focused on the most impactful variables.

Optimal Model Selection:

- **Logistic Regression (Tuned)** and **Decision Tree (Tuned)** models perform best, showing high accuracy and balanced precision/recall metrics.
- **Recommendation:** Prioritize the **Decision Tree (Tuned)** for its overall performance, but ensure to apply pruning techniques to prevent overfitting.