ExtraaLearn Lead Conversion Model

Classification & Hypothesis Testing Project

Shafat Ahsan
Date: 09/05/24

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

- ExtraaLearn an EdTech startup is facing challenges with lead conversion
- They want specific data on which leads are being converted to customers
- Python has been used to analyze the different types of leads related to age, online and offline media
- Decision Tree model and Random Forrest model have been used to predict lead conversion

Data Overview

Observations from Univariate Analysis

First Interaction: The majority of users first interacted with the platform via the website (about 55%), with the remaining using the mobile app (45%).

Profile Completed: Most users have a “High” or “Medium” level of profile completion, with “High” slightly more common. Very few users have a “Low” profile completion level.

Last Activity: The last activity of most users is related to “Email Activity” or “Phone Activity,” with “Website Activity” being the least common.

Print Media Type 1: The majority of users did not engage with Print Media Type 1, with a small portion (around 10%) who did.

Print Media Type 2: Even fewer users engaged with Print Media Type 2, with only about 5% of users interacting with it.

Digital Media: Around 11% of users engaged with Digital Media, while the vast majority (89%) did not.

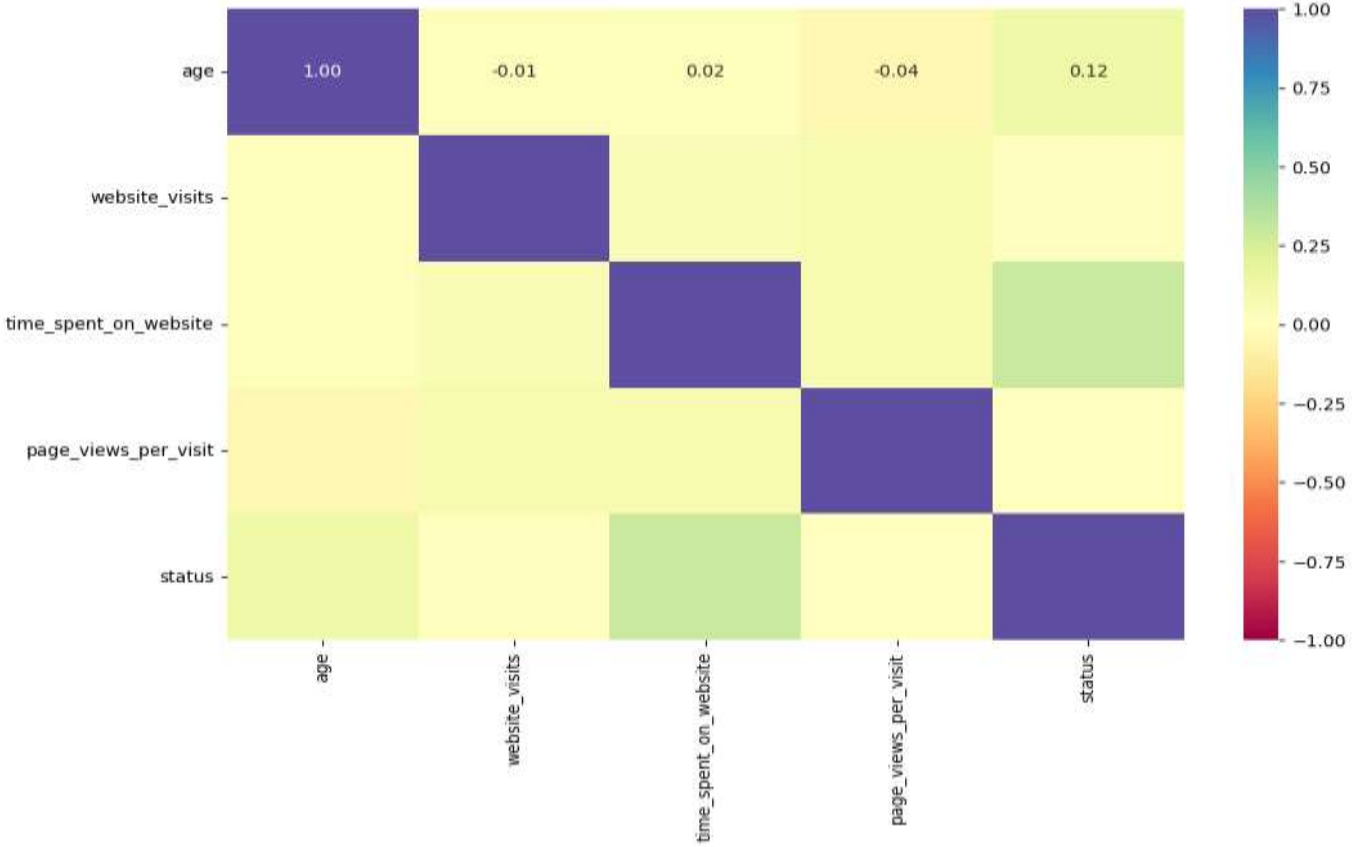
Educational Channels: Most users (85%) did not engage with Educational Channels, while a minority did (15%).

Referral: A very small percentage of users (around 2%) were referred by others, while the overwhelming majority were not.

Status: The “status” variable, which indicates whether a user is a potential customer, shows that around 30% of the users are potential customers, while 70% are not.

Data Overview

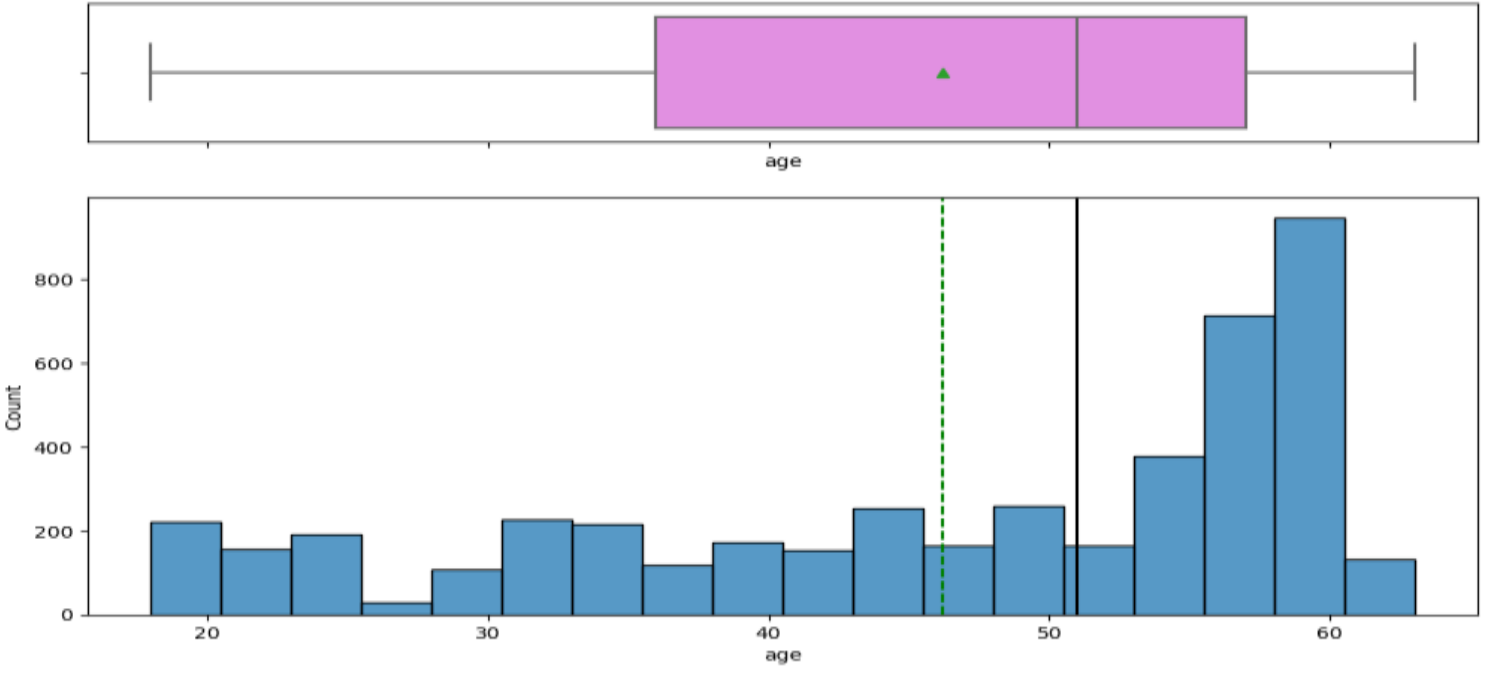
Observations from Bivariate Analysis



Data Overview

Observations on Age

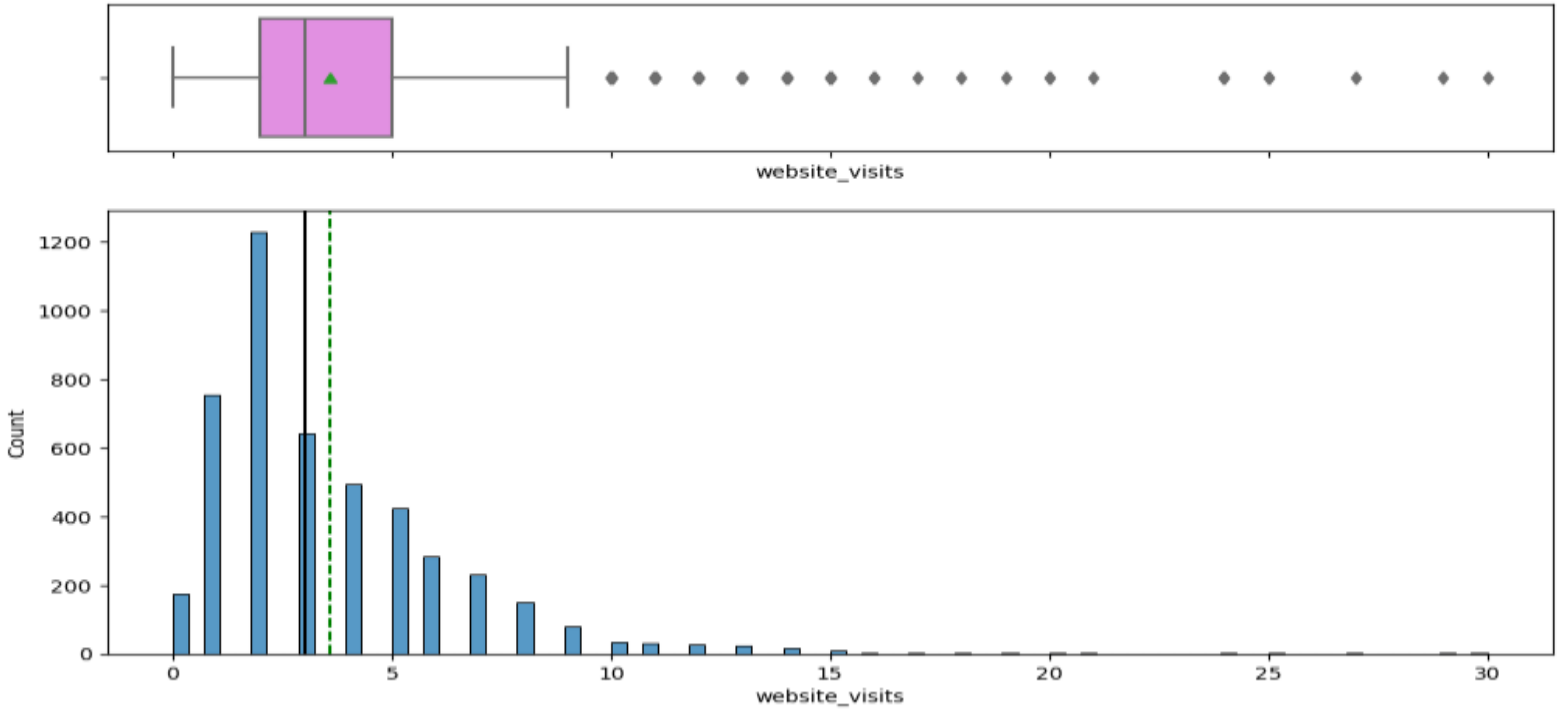
- The distribution of age shows a somewhat uniform distribution, with a slight concentration around the age of 50.
- The boxplot indicates a median age of around 51, with the mean also close to this value.
- There are no significant outliers in the age distribution.



Data Overview

Observations on Website Visits

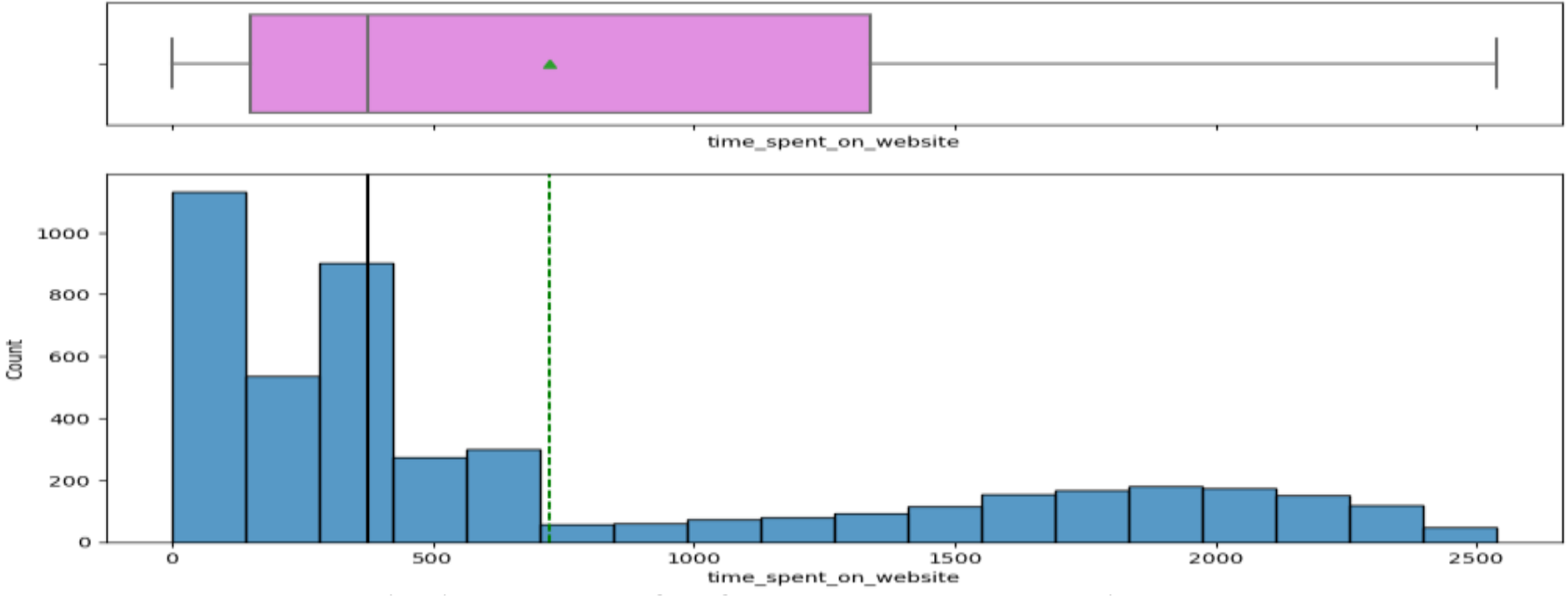
- The distribution of website visits is right-skewed, with most users visiting the website between 2 to 5 times.
- The median number of visits is 3, with the mean slightly higher due to the skewness in the data.
- A few outliers exist where the number of website visits is significantly higher than the majority, reaching up to 30 visits.



Data Overview

Observations on time spent on website

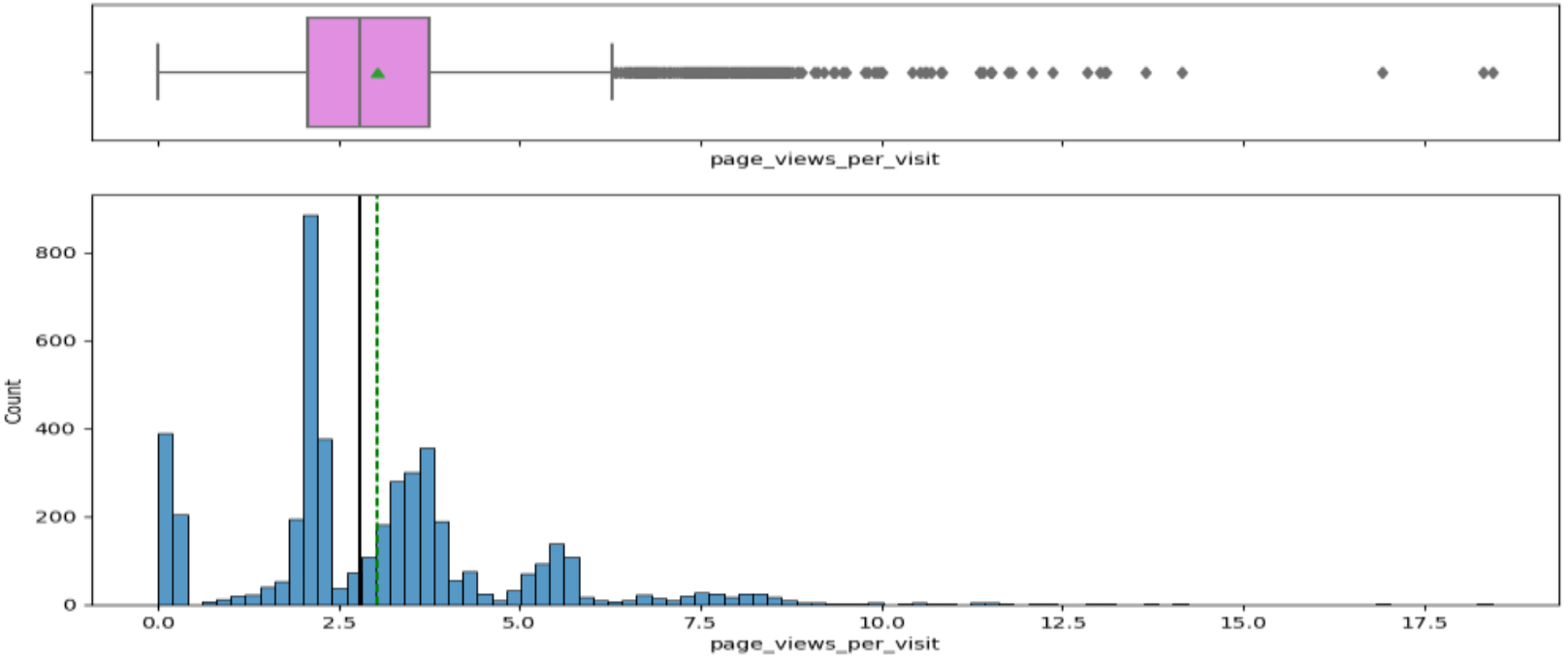
- The distribution of time spent on the website is heavily right-skewed, with the majority of users spending relatively little time on the website.
- The median time spent is much lower than the mean, indicating the presence of a few users who spend a significantly higher amount of time on the website.
- There are several outliers, with a few users spending up to 2537 seconds on the website.



Data Overview

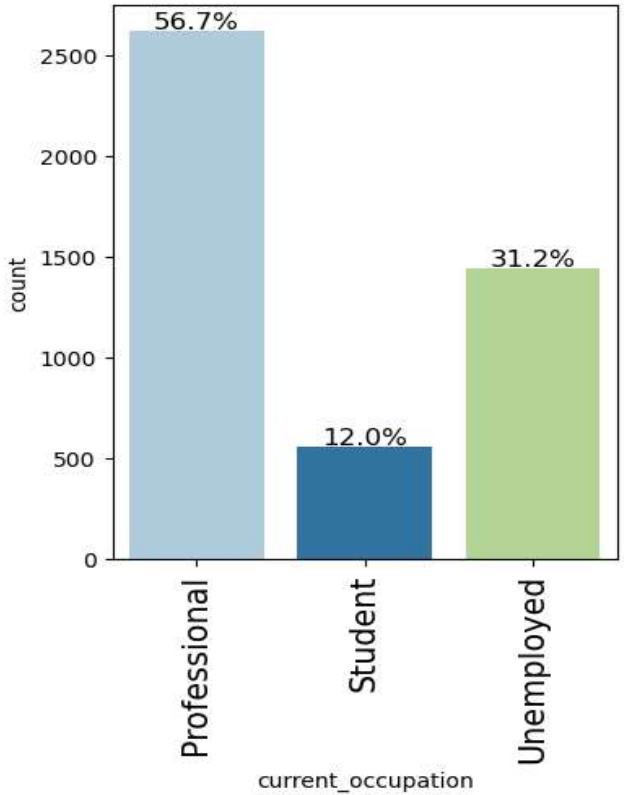
Observations on number of page views per visit

- The distribution of page views per visit is also right-skewed, with most users viewing around 2 to 4 pages per visit.
- The median and mean are relatively close, but the presence of outliers causes the mean to be slightly higher.
- A few users view a very high number of pages per visit, which is reflected in the right tail of the distribution.



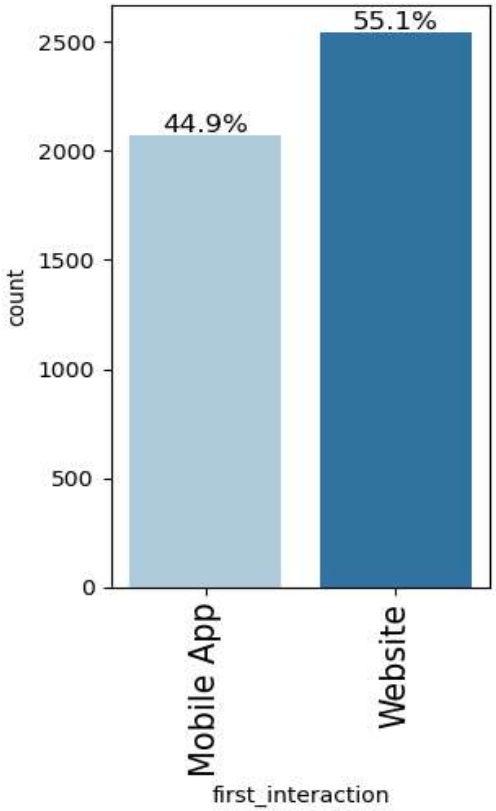
Data Overview

Observations on current occupation



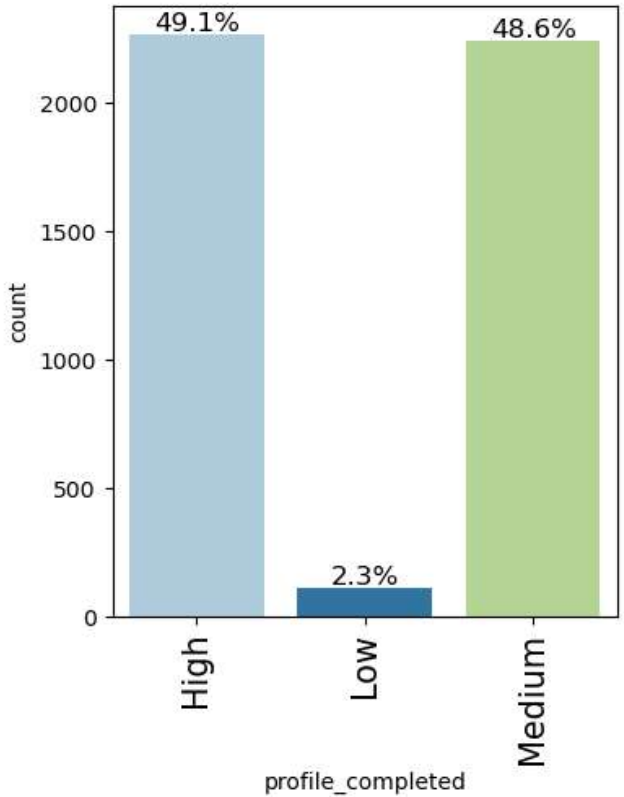
Data Overview

Observations on number of first interaction



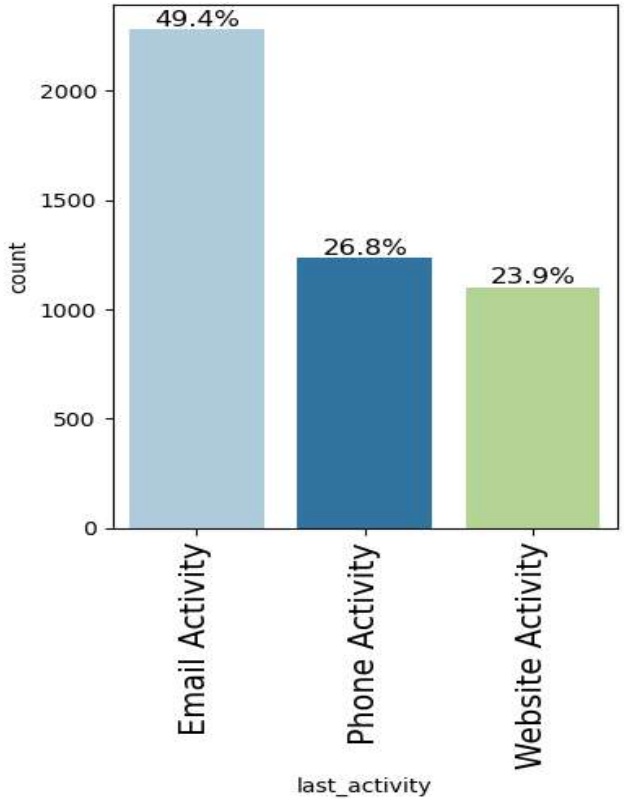
Data Overview

Observations on profile completed



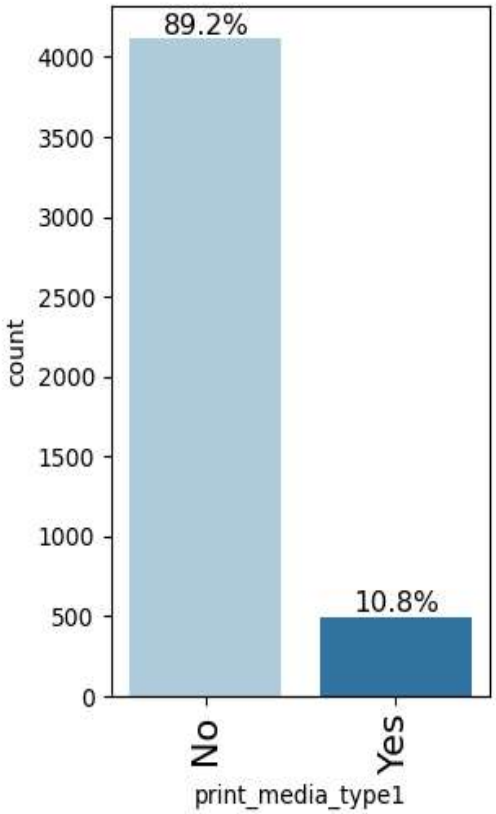
Data Overview

Observations on last activity



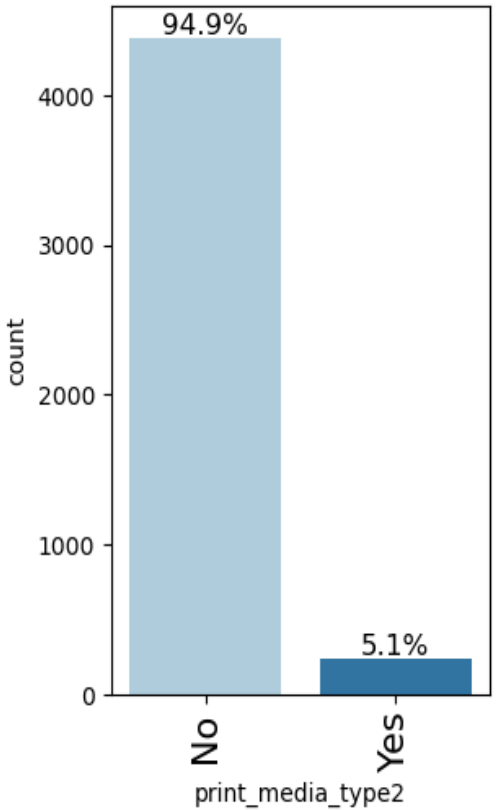
Data Overview

Observations on Print Media type 1



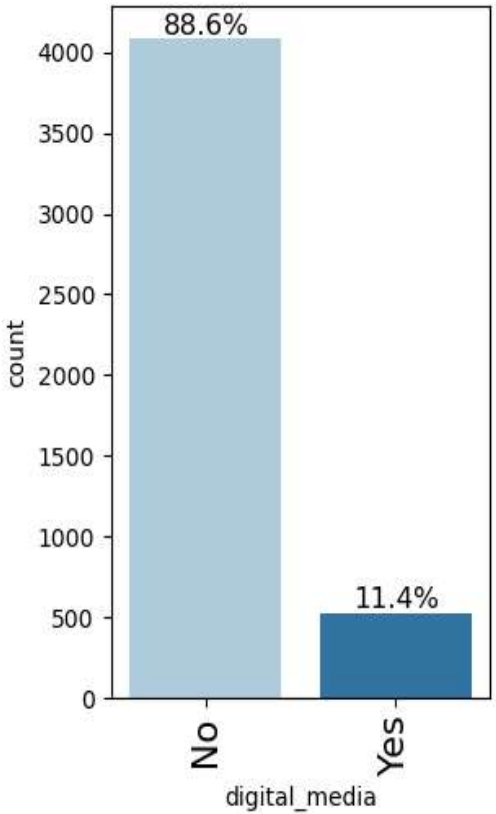
Data Overview

Observations on Print Media type 2



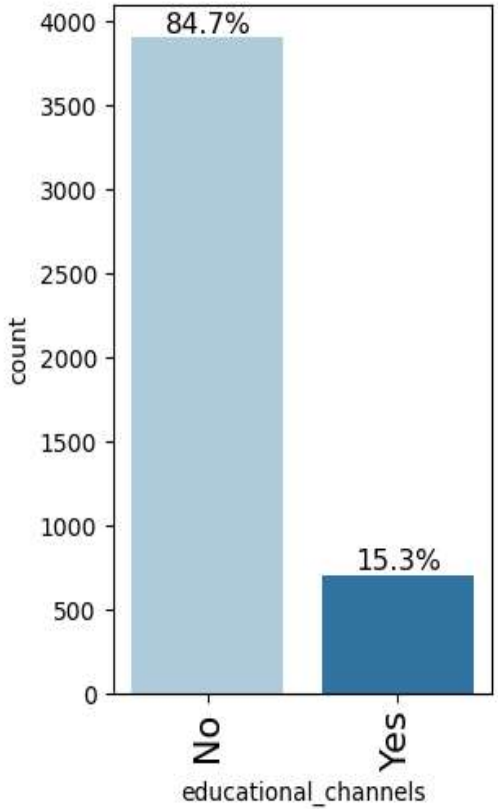
Data Overview

Observations on Digital Media



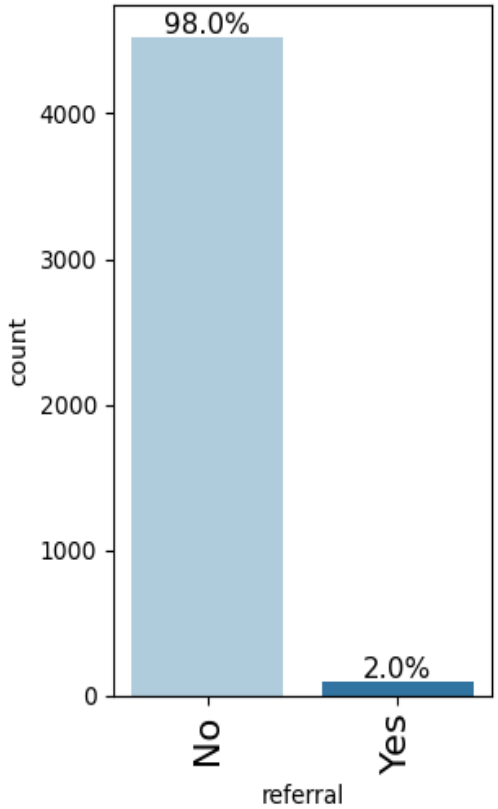
Data Overview

Observations on Educational Channels



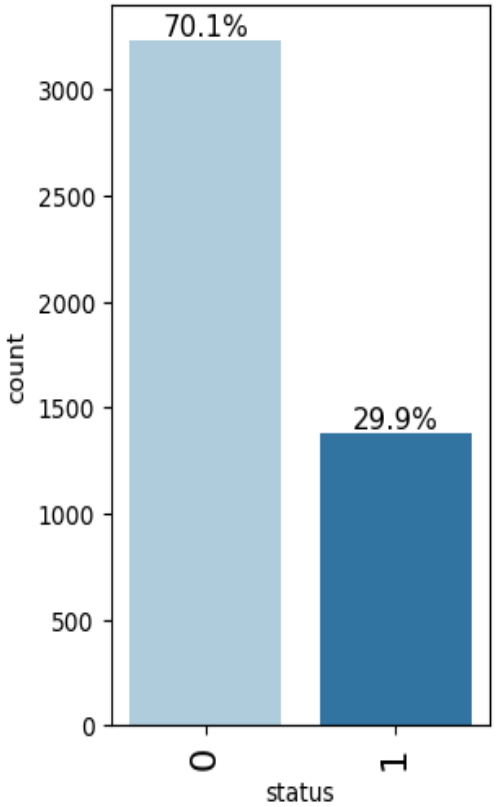
Data Overview

Observations on Referral

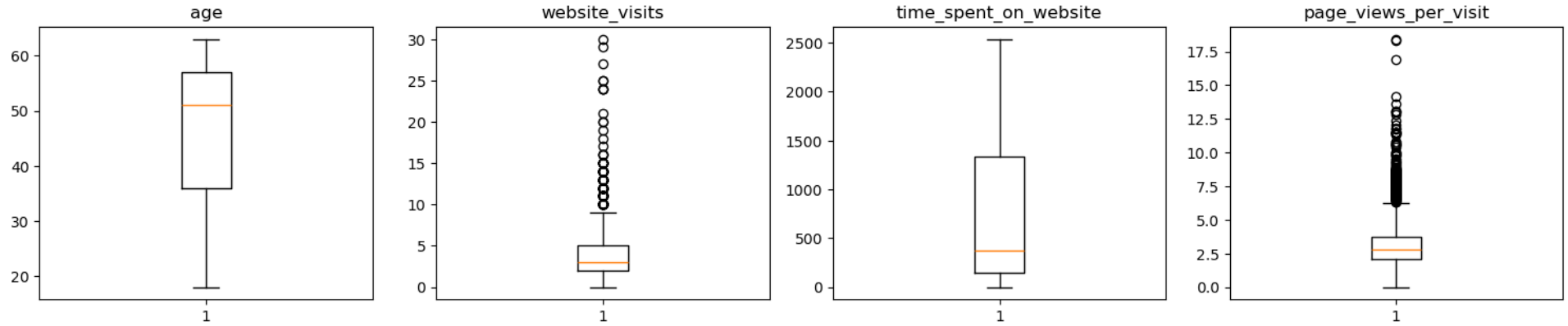


Data Overview

Observations on Status



Outlier Check



Observations

- Age: The boxplot for age shows very few, if any, outliers. The distribution is relatively consistent, with no extreme values.
- Website Visits: There are several outliers in the number of website visits. While most users visit the website between 2 and 5 times, some visit up to 30 times, which are considered outliers.
- Time Spent on Website: The boxplot indicates significant outliers in the time spent on the website. Most users spend a moderate amount of time, but a few spend a very high amount of time, which deviates significantly from the rest.
- Page Views per Visit: There are notable outliers in the number of page views per visit. While most users view around 2 to 4 pages per visit, some users view up to 18 pages, which is an outlier in this context.

Building Decision Tree Model

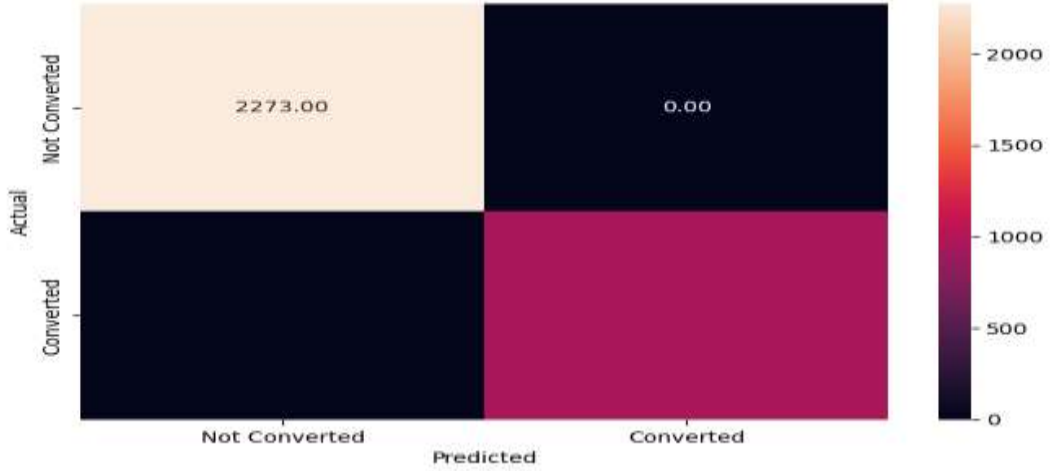
Model Performance on Training Set

Observations:

- Precision, Recall, F1-Score: For both classes (0 and 1), the precision, recall, and f1- score are all 1.00. This means that the model perfectly identified all instances of both classes without any errors.
- Accuracy: The overall accuracy is 100%, indicating that the model correctly predicted the class for every instance in the training set.
- Macro and Weighted Averages: Both macro and weighted averages are also 1.00, further confirming the model's perfect performance on the training data.

Potential Concern:

This level of performance on the training data might suggest that the model is overfitting, especially given that decision trees can easily overfit if not properly tuned. Overfitting means the model may not perform as well on unseen (test) data.



Building Decision Tree Model

Model Performance on Test Data to see if model is overfitting

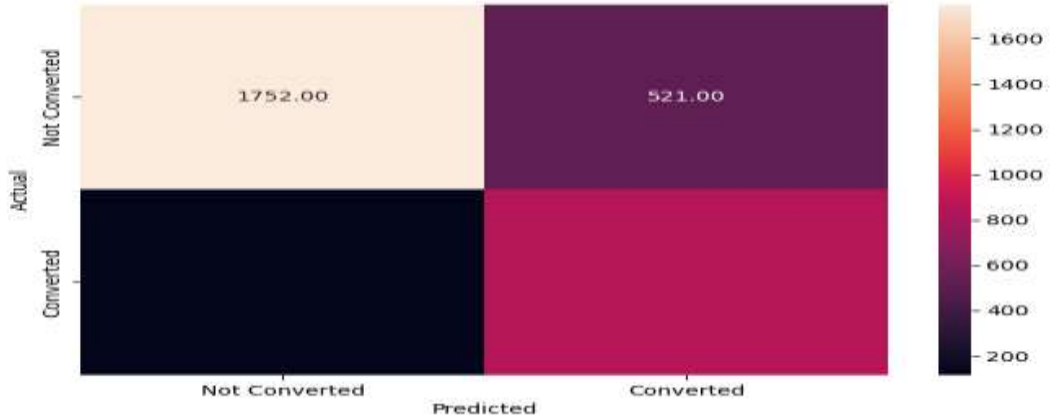
Observations:

- Precision, Recall, F1-Score: For class 0, the precision, recall, and f1-score areapproximately 0.87, 0.86, and 0.86, respectively. This indicates that the model isreasonably good at identifying non-potential customers. • For class 1, the precision,recall, and f1-score are around 0.69, 0.70, and 0.70, respectively. The performancefor predicting potential customers is lower compared to class 0.
- Accuracy: The overall accuracy is 81%, which is decent but shows that the modeldoes not perform perfectly on the test data, indicating that the model overfitted onthe training data.
- Macro and Weighted Averages: Both macro and weighted averages are around0.78 to 0.81, showing that the model's performance is balanced between the two classes but is not as strong as on the training data.

Conclusion:

The model is overfitting, as evidenced by the perfect training performance and lower test performance. This suggests that the model may need to be tuned, or a more complex model like Random Forest should be considered to improve performance on unseen data.

Decision Tree- Hyperparameter Tuning



Model Performance on Tuned Decision Tree Model (training data set)

Observations:

Class 0 (Not Potential Customer):

Precision: 0.94 - The model remains highly precise in identifying non-potential customers.

Recall: 0.77 - The model correctly identifies 77% of non-potential customers, but misses 23%.

F1-Score: 0.85 - This indicates a good balance between precision and recall for class 0.

Class 1 (Potential Customer):

Precision: 0.62 - The precision for identifying potential customers is moderate, with some false positives.

Recall: 0.88 - The recall for potential customers is strong, correctly identifying 88% of them.

F1-Score: 0.73 - This shows a reasonably balanced performance for class 1.

Overall Performance

Accuracy: 0.80 - The model correctly classifies 80% of the training data.

Macro Average: Precision: 0.78

Recall: 0.83

F1-Score: 0.79

Weighted Average:

Precision: 0.84

Recall: 0.80

F1-Score: 0.81

Decision Tree- Hyperparameter Tuning

Model Performance on Tuned Decision Tree Model (test data set)

Observations:

Precision, Recall, F1-Score:

•For class 0: The precision, recall, and f1-score are 0.93, 0.77, and 0.84, respectively. This indicates that the model is highly precise in identifying non-potential customers, but it misses some of them, as seen by the lower recall.

•For class 1: The precision, recall, and f1-score are 0.62, 0.86, and 0.72, respectively. The model shows a strong recall for potential customers, meaning it correctly identifies 86% of them, but the precision is moderate, with some false positives.

Accuracy:

•The overall accuracy is 80%, which is consistent with the performance on the training data, suggesting that the model generalizes well without significant overfitting.

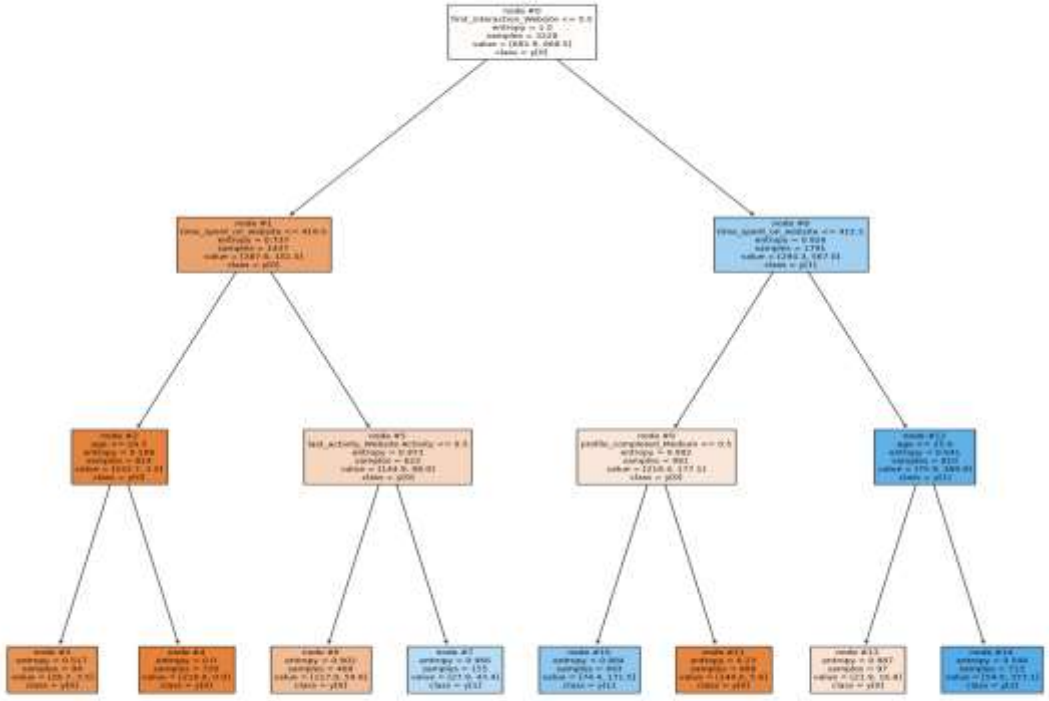
Macro and Weighted Averages:

• Macro Average: The precision, recall, and f1-score are 0.77, 0.82, and 0.78, respectively, showing a balanced performance across both classes.

• Weighted Average: The precision, recall, and f1-score are 0.83, 0.80, and 0.80, respectively, indicating that the model's performance is well-distributed according to the class distribution in the test set.

These observations indicate that the tuned decision tree model performs consistently on both the training and test data, with a particular strength in recalling potential customers (class 1) while maintaining good overall accuracy.

Visualizing the Decision Tree



Note: Blue leaves represent the converted leads, i.e., **y[1]**, while the orange leavesrepresent the not converted leads, i.e., **y[0]**. Also, the more the number of observationsin a leaf, the darker its color gets

Visualizing the Decision Tree

Observations from the Decision Tree Diagram:

1. Root Node (First Interaction – Website vs. Mobile App): The root node splits based on whether the first interaction was via the website or not. This suggests that the method of first interaction is a significant factor in predicting whether someone will become a potential customer.

2. Time Spent on Website:
- The next significant split happens based on the time spent on the website. Users who spend less time on the website (less than around 420 seconds) are more likely to be non-potential customers (class 0).
- Conversely, those who spend more time on the website are more like to be potential customers (class 1).

3. Age and Last Activity:
- Age is another important factor, particularly in further splitting groups of users who spend less time on the website. Younger users (age <= 24.5) are less likely to be potential customers.
- Last activity (Website Activity) is important for users who spent less time on the website but are still likely to convert, particularly if their last activity wasn't on the website itself.

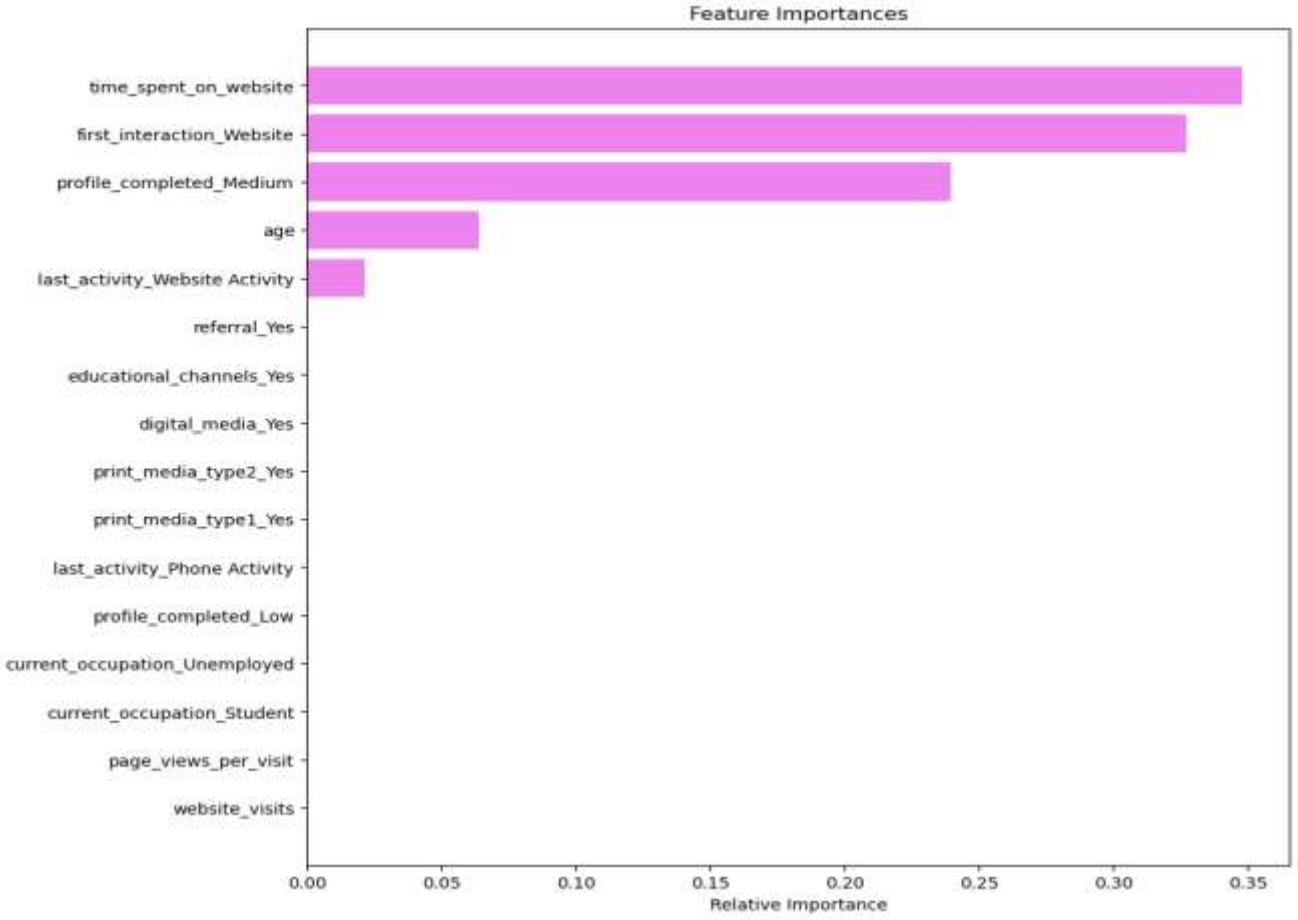
4. Profile Completion:
- The level of profile completion (High/Medium) is also a decisive factor, particularly for users who spend a moderate amount of time on the website.

5. Leaf Nodes:
- The leaf nodes represent the final classification decisions, with colors indicating the predicted class (orange for class 0 and blue for class 1). The entropy values at these nodes indicate how pure each final group is (with lower entropy indicating higher purity).

6. Class Distribution:
- The decision tree shows a clear pattern where users who interact primarily via the website, spend more time on it, and have a medium or high profile completion level are more likely to be potential customers.

Conclusion:
The decision tree highlights key factors that influence whether a user is likely to convert into a potential customer. These factors include the method of first interaction, time spent on the website, age, last activity, and profile completion. The visualization shows how the model uses these features to make predictions, with certain splits clearly favoring one class over the other.

Feature Importance of Tunes Decision Tree model



Feature Importance of Tuned Decision Tree model

Observations:

1. Time Spent on Website:
•This feature has the highest importance in the model. The amount of time a user spends on the website is a strong predictor of whether they will become a potential customer. Users who spend more time on the website are more likely to convert.

2. First Interaction (Website vs. Mobile App):
•The method of first interaction is the second most important feature. Users who initially interacted via the website have a different likelihood of converting compared to those who used the mobile app.3.Profile Completed (Medium):•The level of profile completion is also significant, particularly for those with a “Medium” completion level. This suggests that users who take the time to fill out their profile moderately are more engaged and hence more likely to become potential customers.

4. Age:
• Age is another important factor, indicating that certain age groups are more likely to convert than others.
5. Last Activity (Website Activity):
• The last activity, particularly if it was on the website, plays a role in determining the likelihood of conversion. This suggests that recent engagement with the website is a positive indicator.

6. Lower Importance Features:
• Other features, such as whether the user was referred, engagement with educational channels, digital media, and print media, as well as their occupation, have lower importance. These features contribute to the model but are less decisive compared to the top predictors.

Building Random Forest Model

Model Performance on Training Set

Observations:

- Precision, Recall, F1-Score: • For both classes (0 and 1), the precision, recall, and f1- score are all 1.00. This indicates that the random forest model has perfectly classified all instances in the training set without any errors.
- Accuracy: • The overall accuracy is 100%, meaning the model correctly predicted the class for every instance in the training set.
- Macro and Weighted Averages: • Both macro and weighted averages are also 1.00, confirming the model's perfect performance on the t

Building Random Forest Model

Model Performance on Test Data

Observations:

Precision, Recall, F1-Score:

For class 0:

•Precision: 0.87 - The model is quite precise in identifying non-potential customers, with a low rate of false positives.

•Recall: 0.91 - The model correctly identifies 91% of non-potential customers.

•F1-Score: 0.89 - This indicates a strong balance between precision and recall for class 0.

For class 1:

•Precision: 0.78 - The model has moderate precision for identifying potential customers, with a higher rate of false positives.

•Recall: 0.68 - The model correctly identifies 68% of potential customers, indicating that some potential customers are missed.

•F1-Score: 0.73 - This reflects a decent but not perfect balance between precision and recall for class 1.

Accuracy:

• The overall accuracy is 84%, which indicates that the model performs well on the test data but is not perfect.

Random Forrest- Hyperparameter Tuning

Model Performance on Tuned Random Forrest Model (test data set)

Observations:

Class 0 (Not Potential Customer):

•Precision: 0.88

•Recall: 0.93•F1-Score: 0.90

•The model is highly effective at identifying non-potential customers with good precision and recall.

Class 1 (Potential Customer):

• Precision: 0.81

• Recall: 0.70

• F1-Score: 0.75

• The model has good precision for identifying potential customers but is slightly weaker in recall, missing some potential customers.

Overall Accuracy:

• The accuracy of the test data is 86%, which indicates that the model generalizes fairly well to unseen data, but with some trade-offs.

Macro and Weighted Averages:

• Precision: 0.84

• Recall: 0.81

• F1-Score: 0.83

Weighted Avg:

• Precision: 0.86

• Recall: 0.86

• F1-Score: 0.86

Decision Tree- Hyperparameter Tuning

Model Performance on Tuned Random Forrest Model (training data set)

Observations:

Precision, Recall, F1-Score:

•For both classes (0 and 1), the precision, recall, and f1-score are all 1.00, indicating perfect classification of the training data.

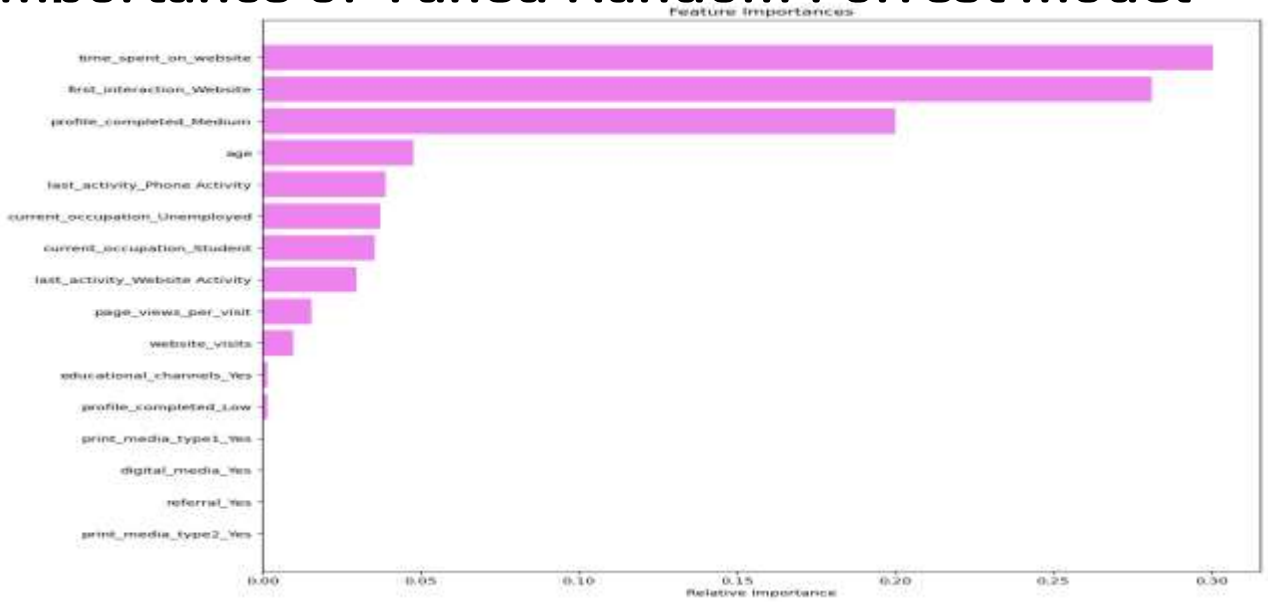
Accuracy:

•The overall accuracy is 100%, showing that the model has perfectly memorized the training data.

Conclusion:

•This indicates that the model is overfitting on the training data, as it has learned the training set too well.

Feature Importance of Tuned Random Forrest model



Observations:

- Similar to the decision tree model, **time spent on the website, first_interaction_website, profile_completed, and age are the top four features that** help distinguish between not converted and converted leads.
- Unlike the decision tree, **the random forest gives some importance to other variables like occupation, page_views_per_visit, as well.** This implies that the random forest is giving importance to more factors in comparison to the decision tree.

Conclusion and Recommendations

Conclusions:

1. Key Predictors of Conversion:

• Time Spent on Website: This is the most significant predictor of whether a lead will convert, indicating that users who spend more time on the website are more likely to become customers.

• First Interaction via Website: Leads who initially interact via the website have a higher likelihood of conversion compared to those who first use the mobile app.

• Profile Completion: A “Medium” level of profile completion is strongly associated with conversion, suggesting that engaged users who take time to fill out their profiles are more likely to convert.

• Age: Age also plays an important role, with certain age groups showing a higher likelihood of conversion.

2. Model Performance:

•Random Forest Model: The random forest model performs well, particularly with an accuracy of 86% on the test data. It also highlights additional factors like occupation and page views per visit, which contribute to the prediction of conversion likelihood.

•Overfitting: The random forest model shows signs of overfitting on the training data, achieving perfect accuracy. However, it generalizes reasonably well to test data, indicating a balance between complexity

Conclusion and Recommendations

Business Recommendations:

1. Enhance User Engagement:

•Increase Time Spent on Website: Since time spent on the website is a key predictor of conversion, efforts should be made to engage users longer. This could be through content improvements, interactive features, or personalized experiences.

• Encourage Profile Completion: Promote profile completion by offering incentives or highlighting the benefits of a more complete profile. This can increase user engagement and improve conversion rates.

2. Targeted Marketing Strategies:

•Focus on Web Interactions: Given that first interactions via the website are strongly linked to conversions, prioritize web-based marketing and ensure the website provides a smooth, informative, and engaging experience.

•Age-Specific Campaigns: Tailor marketing campaigns to target age groups that are more likely to convert, using age-appropriate messaging and offers.

3. Leverage Additional Insights:

•Occupation-Based Targeting: Use the insights on occupation to better target unemployed individuals or students with specific campaigns or offers that resonate with their needs and circumstances.

•Monitor and Optimize Page Views: Since the number of page views per visit is also a contributing factor, analyze user navigation patterns and optimize the website to encourage exploration and discovery.

4. Continuous Monitoring and Model Updating:

• Regularly update the model with new data to ensure it remains accurate and relevant. Continuously monitor its performance and adjust the strategy based on the latest insights.

These recommendations should help the business improve lead conversion rates by focusing on the most influential factors and optimizing user experience accordingly.

Thank You!