EXTRAALEARN

POTENTIAL LEADS

Natasha Gitlin
January 7, 2024

BUSINESS PROBLEM OVERVIEW

- **Context**: The EdTech industry is rapidly growing, with a projected worth of $286.62 billion by 2023.
- **Challenge**: ExtraaLearn, an early-stage startup, faces the dilemma of efficiently identifying leads with the highest potential for conversion among the numerous leads generated regularly.
- **Objective**: Allocate resources effectively by pinpointing leads likely to convert to paid customers.

SOLUTION

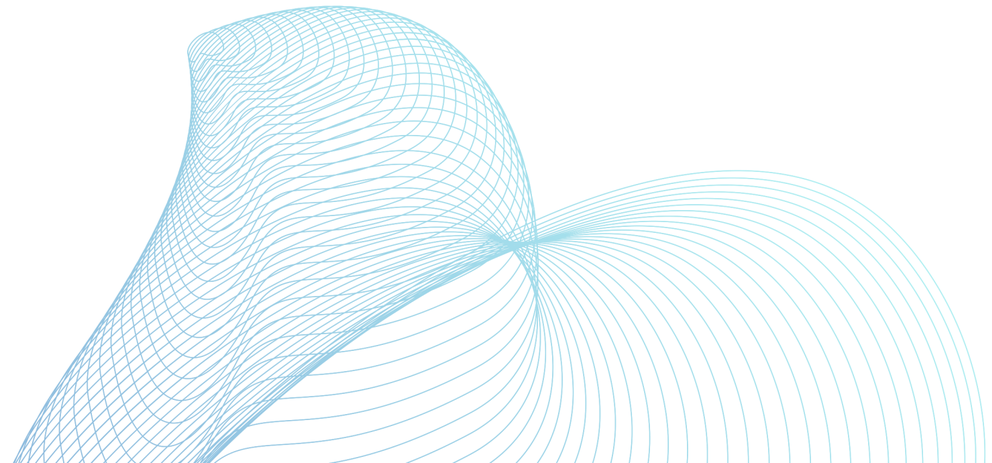
1. **Machine Learning Model:**
    - Objective: Develop an ML model to predict which leads are more likely to convert.
    - Benefits: Enhance resource allocation by focusing efforts on high-potential leads.
2. **Factor Analysis:**
    - Objective: Identify and understand the key factors influencing lead conversion.
    - Benefits: Inform targeted marketing strategies and improve overall conversion rates.
3. **Lead Profiling:**
    - Objective: Create profiles for leads likely to convert based on analyzed factors.
    - Benefits: Streamline marketing efforts and tailor engagement strategies for specific lead segments.

DATA OVERVIEW

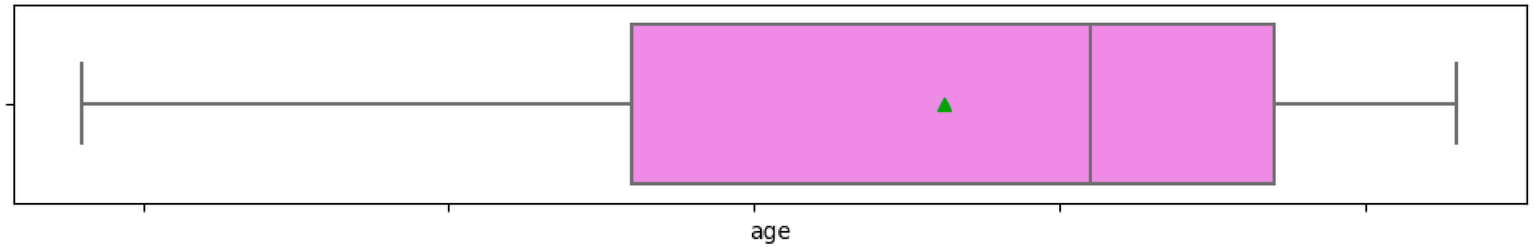
The different attributes of leads and their interaction details with ExtraaLearn. The detailed data dictionary is given below.

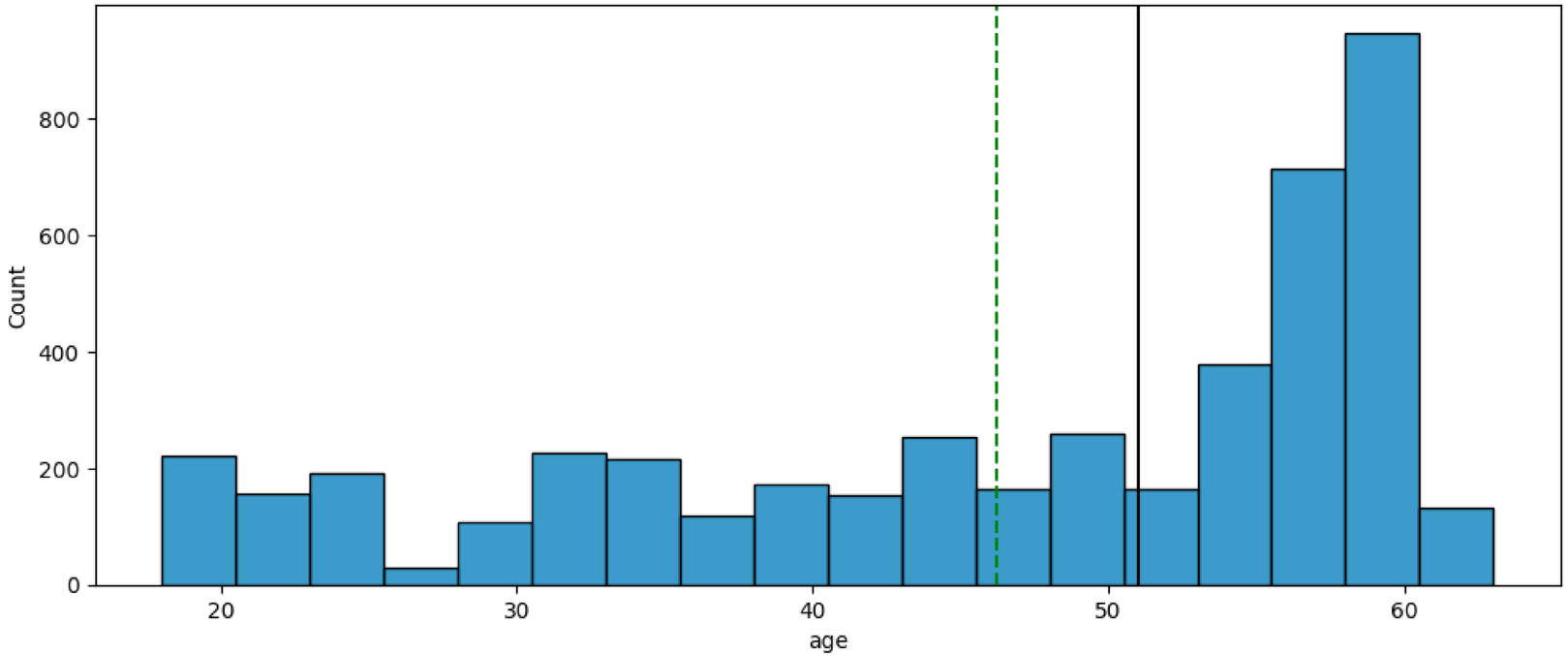
- Features include **ID**, **age**, **Current Occupation**, and **First Interaction** (**website** or **mobile** app).
- **Profile completion** percentage categorized as **Low** (0-50%), **Medium** (50-75%), **High** (75-100%).
- Metrics such as **website visits**, **time spent on the website**, and **page views per visit**.
- Activities recorded: **Email, Phone, Website interactions**, and **last activity** details.
- Flags indicating exposure to **ExtraaLearn ads** in print and digital media.
- **Educational channels** and **referral flags** indicating lead source awareness.
- "**Status**" flag denotes lead conversion, with **class 1** as a **paid** customer and **class 0** as an **unpaid** customer.

EDA RESULTS



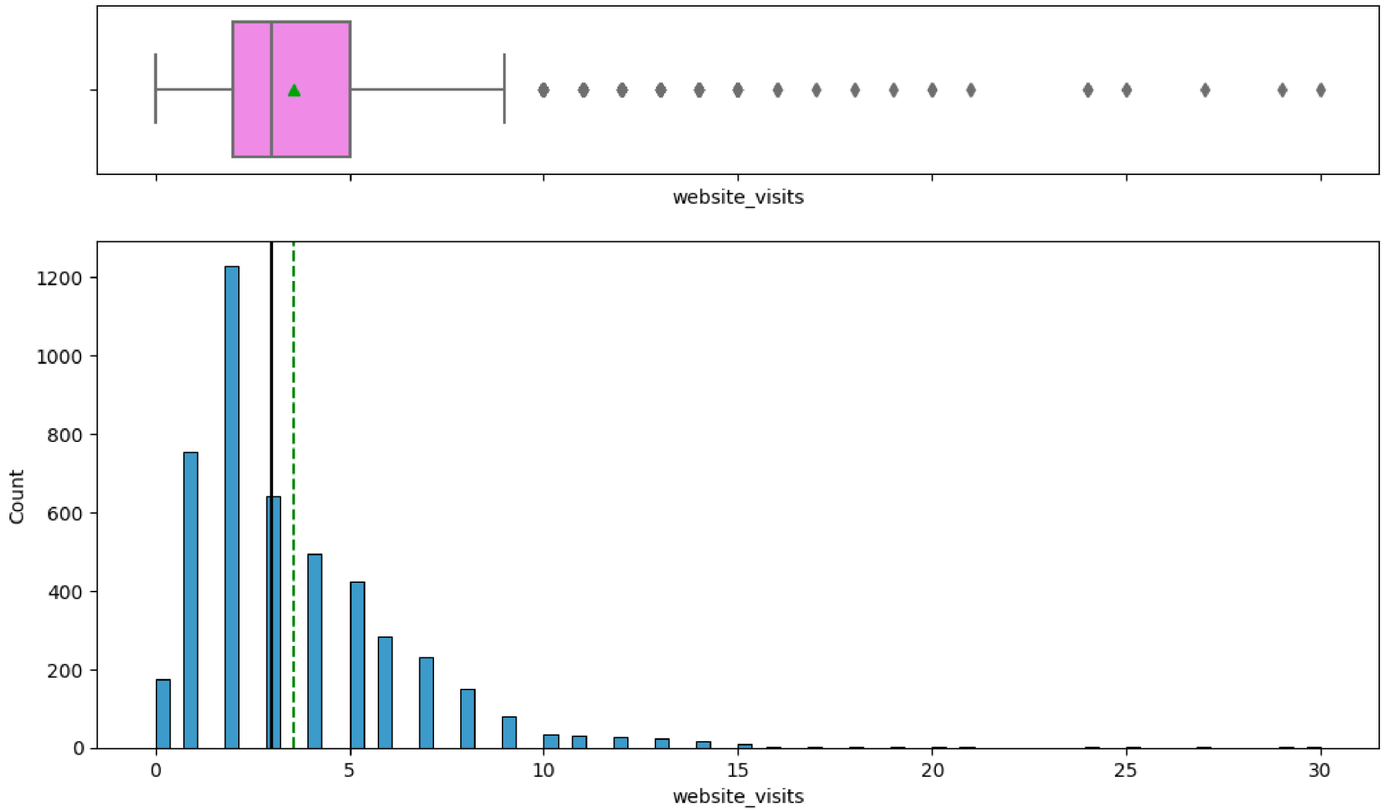
AGE





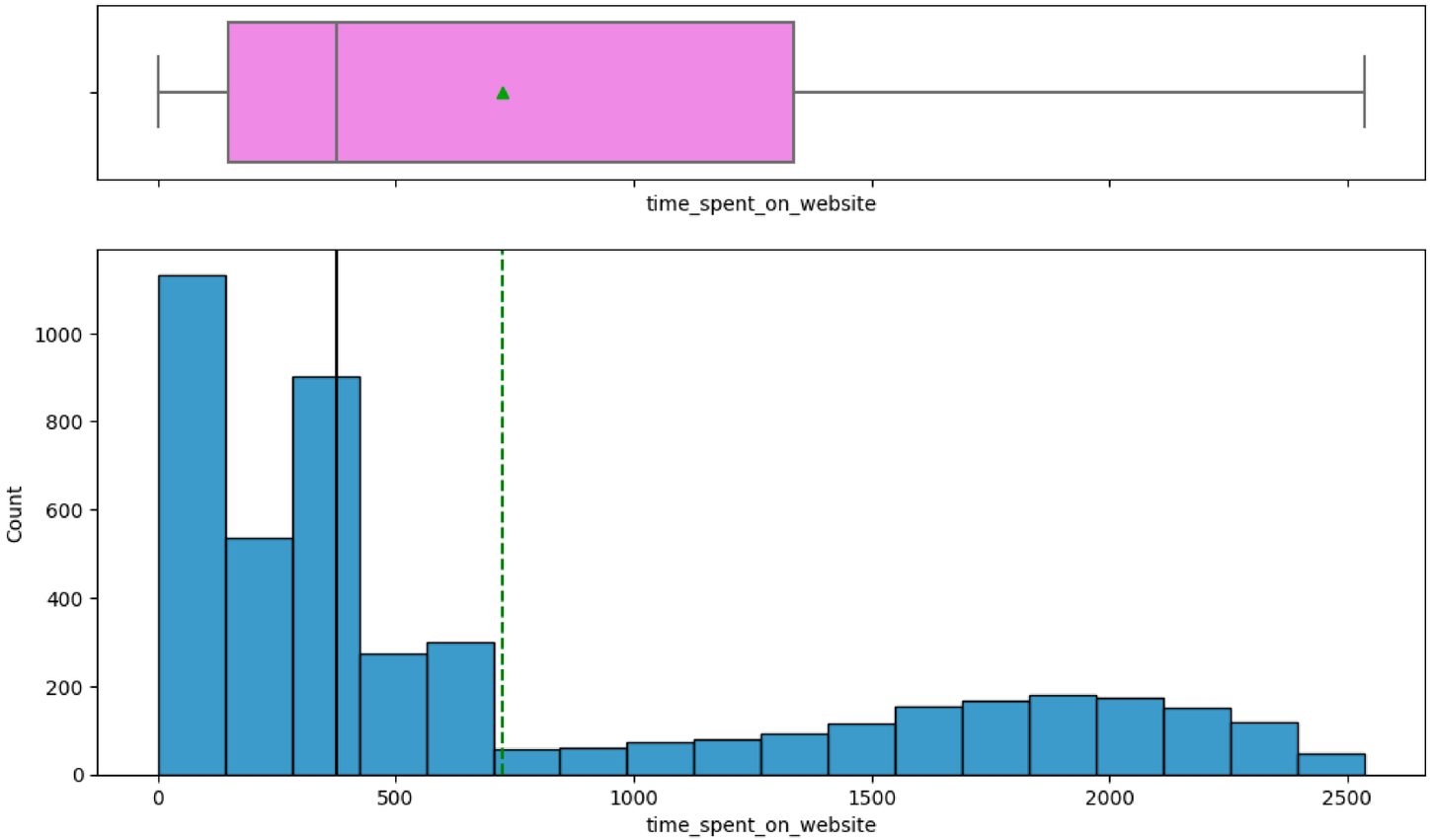
- Age distribution is skewed to the left
- The age distribution has a slight concentration in the middle age range
- The age column has many outliers for this variable
- Age doesn't seem to have much determining power on whether a specific lead will convert.

WEBSITE VISITS



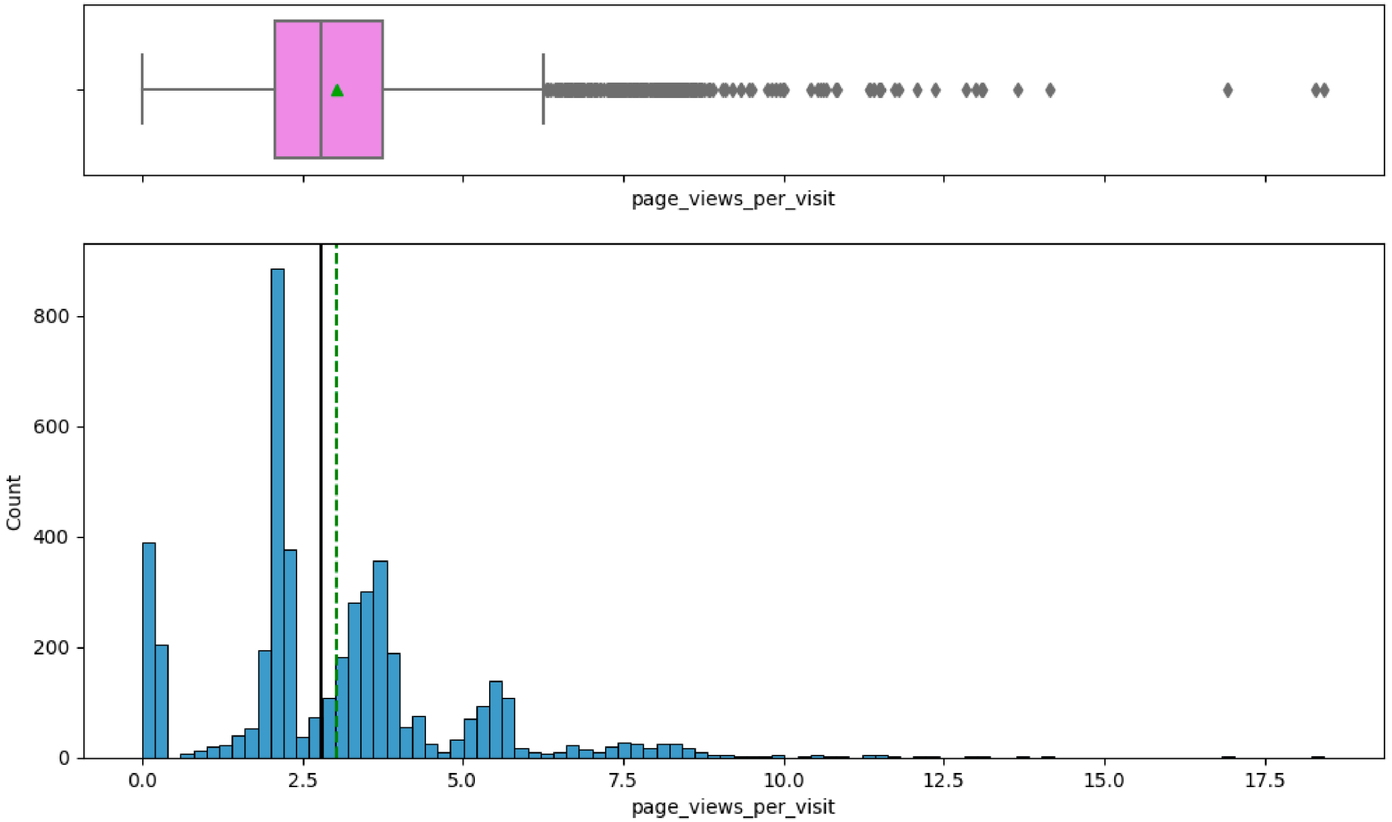
- The distribution for the duration of pitch is right-skewed
- The average number of website visits is less than 5 times
- You could consider outliers for website visits more than ~6 times but most data points showing fewer than 10 visits

TIME SPENT ON WEBSITE

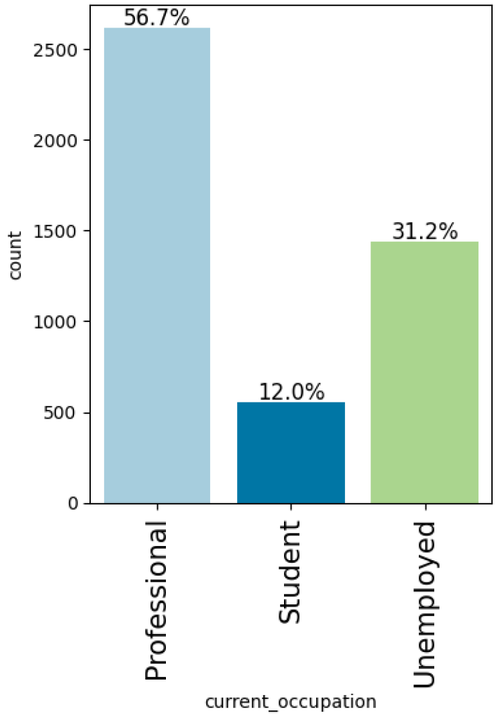


- The distribution for time spent on the website is right-skewed
- Where the average time spent on there is around 8 minutes and the median is ~7 minutes
- The time_spent_on_website is an important factor to consider as the more time spent on there, the more interested they are
- It doesn't look like there are any outliers for this column and should be used to build our model

PAGE VIEWS PER VISIT

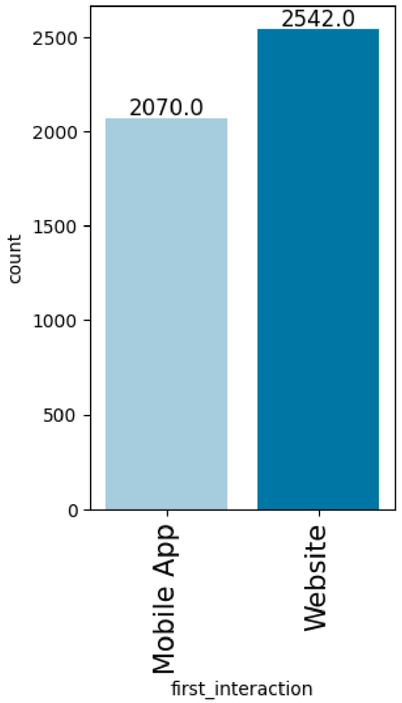


- This distribution for page views per visit is right-skewed
- There are a lot of outliers for page views of 7.5 or more
- However, it seems to be a normal distribution around 2.5 pages
- While this column could be zoomed in and filtered based on IQR, it might not be a strong determining factor on predicting 'status'



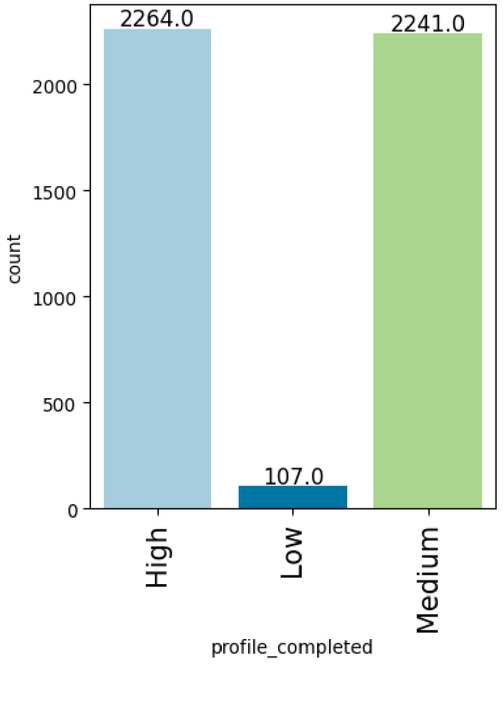
CURRENT OCCUPATION

Significant difference between occupations where Professional is at highest frequency at close to 60%



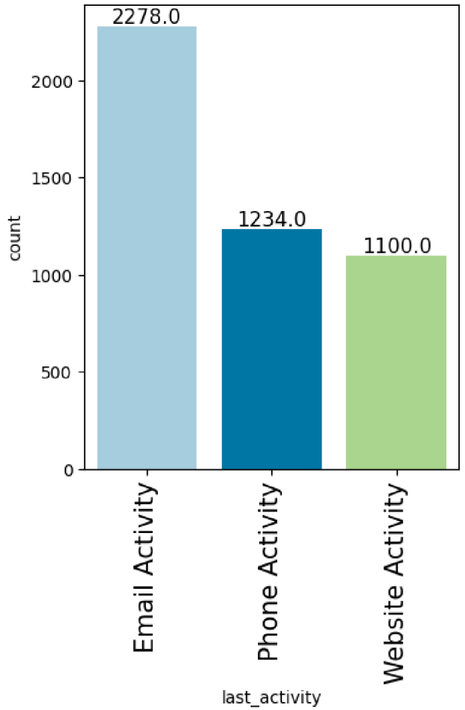
FIRST INTERACTION

Website is the most common first interaaction point for leads with a total of 2542 instances. Both apps are well-utilized with Mobile at 2070 instances



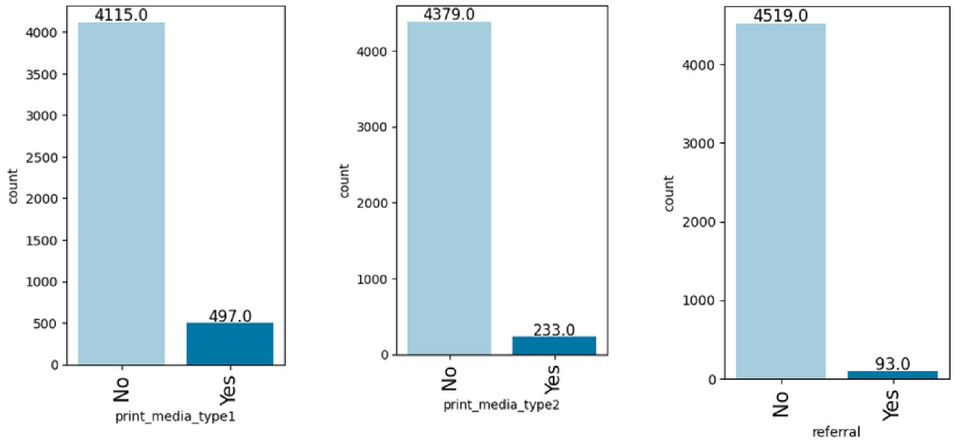
PROFILE COMPLETED

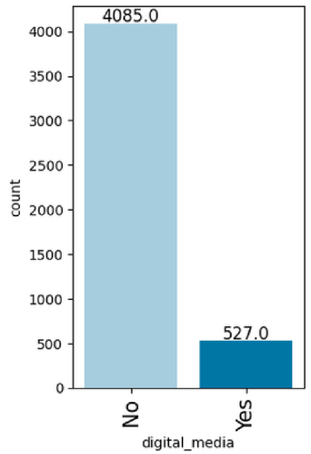
Shows that there is a high frequency of completing their profile. There is a high level of engagement

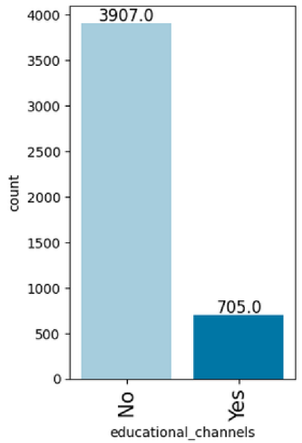


LAST ACTIVITY

Highest frequency of activity is Email which is significantly higher at 2278. The phone and website activity are closer together.







Media Ad Interactions

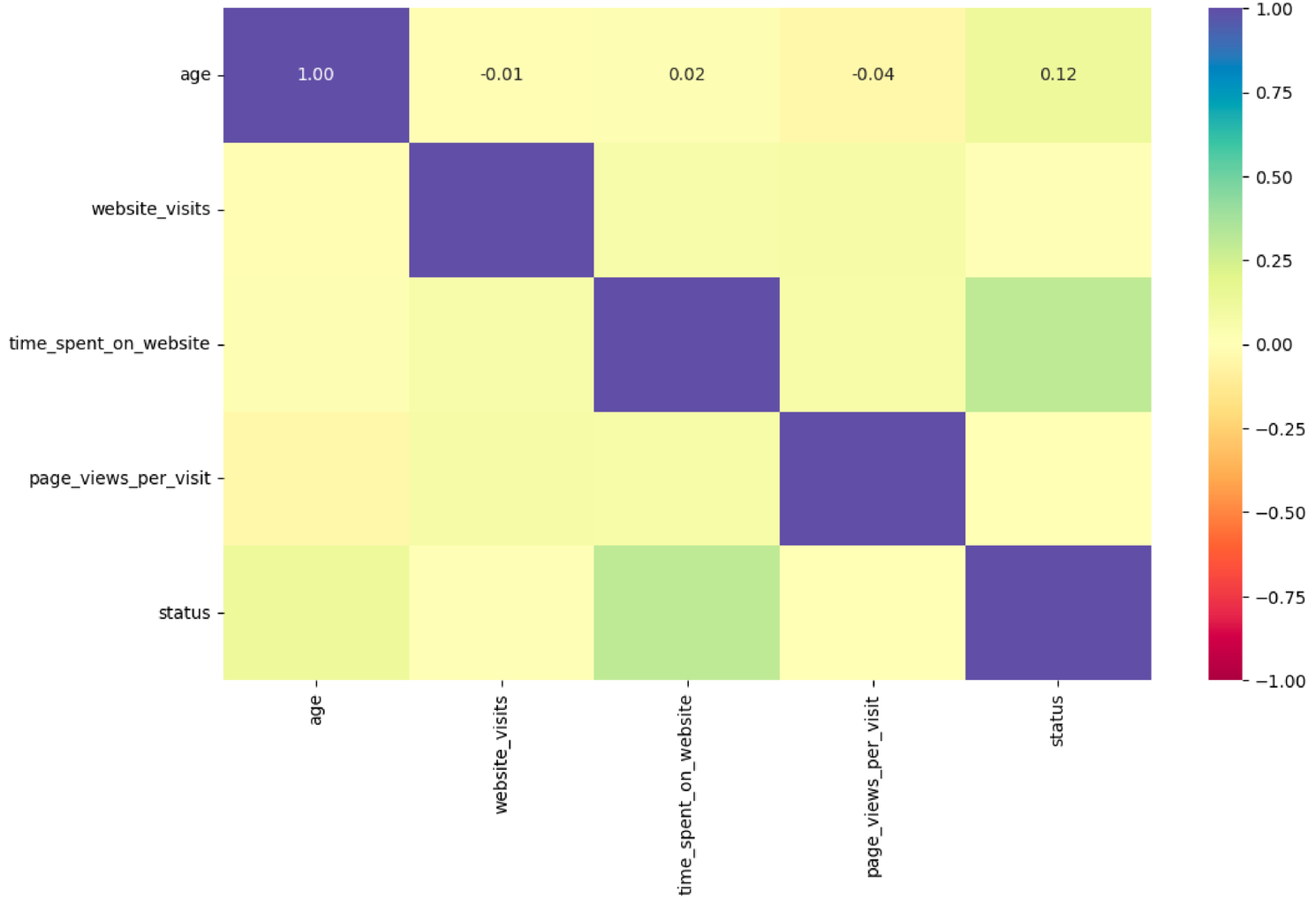
There was low frequency of leads using additional outlets such as ad media, educational channels and referrals. It shows that these methods are outdated and not as helpful in determining potential leads

OBSERVATIONS FROM UNIVARIATE ANALYSIS

- Overall, there is a pretty high signficance of website activity that shows what ExtraaLearn has to offer.
- Email marketing is successful out of the three interaction types, website, email or website.
- It shows there is a high significance of conversion rate for percentage of profile completed and could be a strong contender for building a predictive model
- The columns we should be considering are current_occupation, profile_completed, time_spent_on_website, website_visits, first_interaction
- The columns that show 'how' a lead heard of ExtraaLearn would in theory be helpful however, there is not enough strong data to use as most of it is No

BIVARIATE ANALYSIS

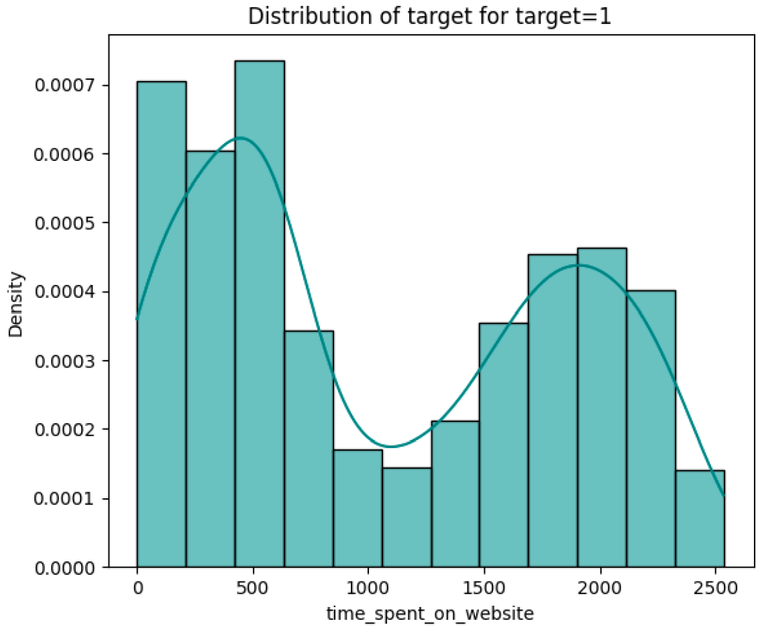
Some variables show little to no correlation with others as you can see.

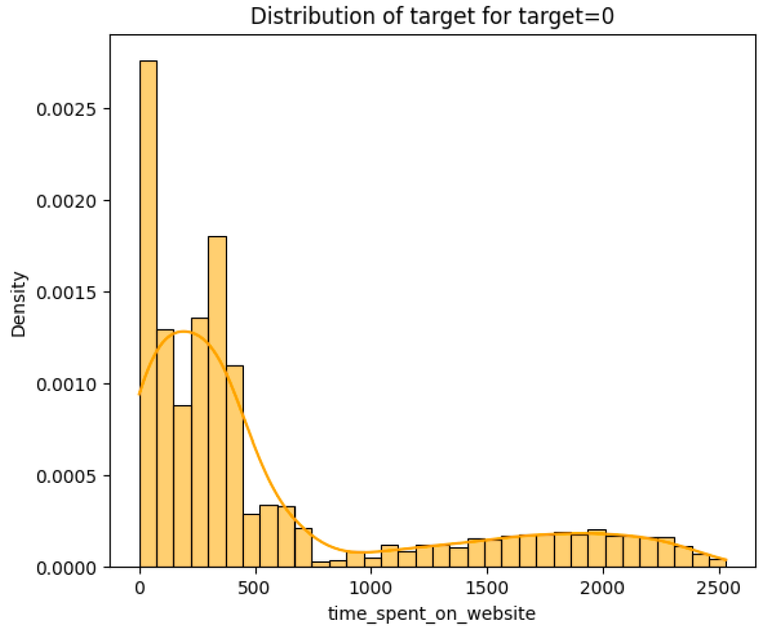


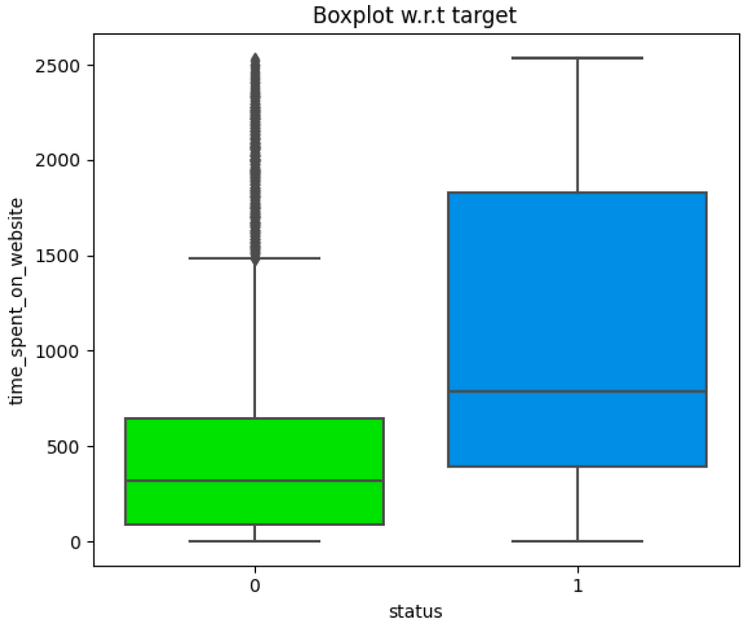
TIME SPENT ON WEBSITE AND STATUS

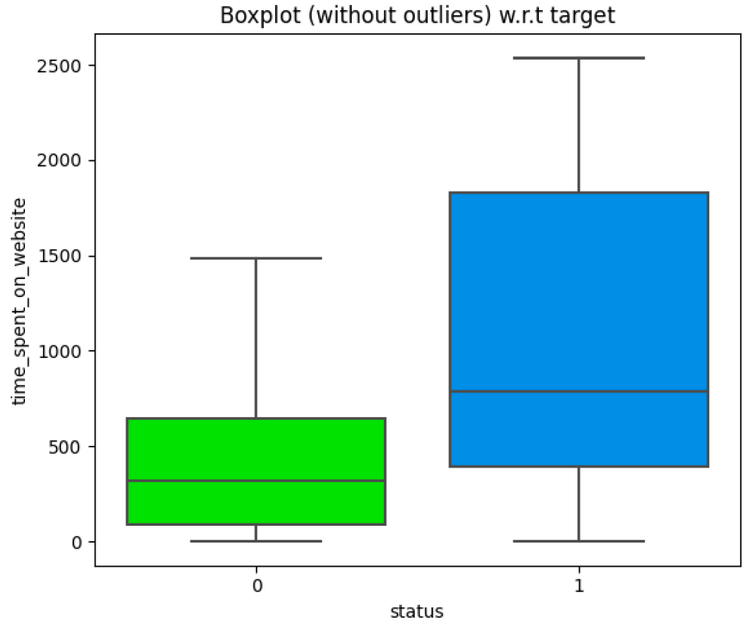
Engagement and Conversion:

- Individuals who spend more time on the website may be more engaged with the content or services offered.
- Longer time spent on the website could be a positive indicator of potential conversion to a paid status when it comes to paid (0) or non-paid (1) customers



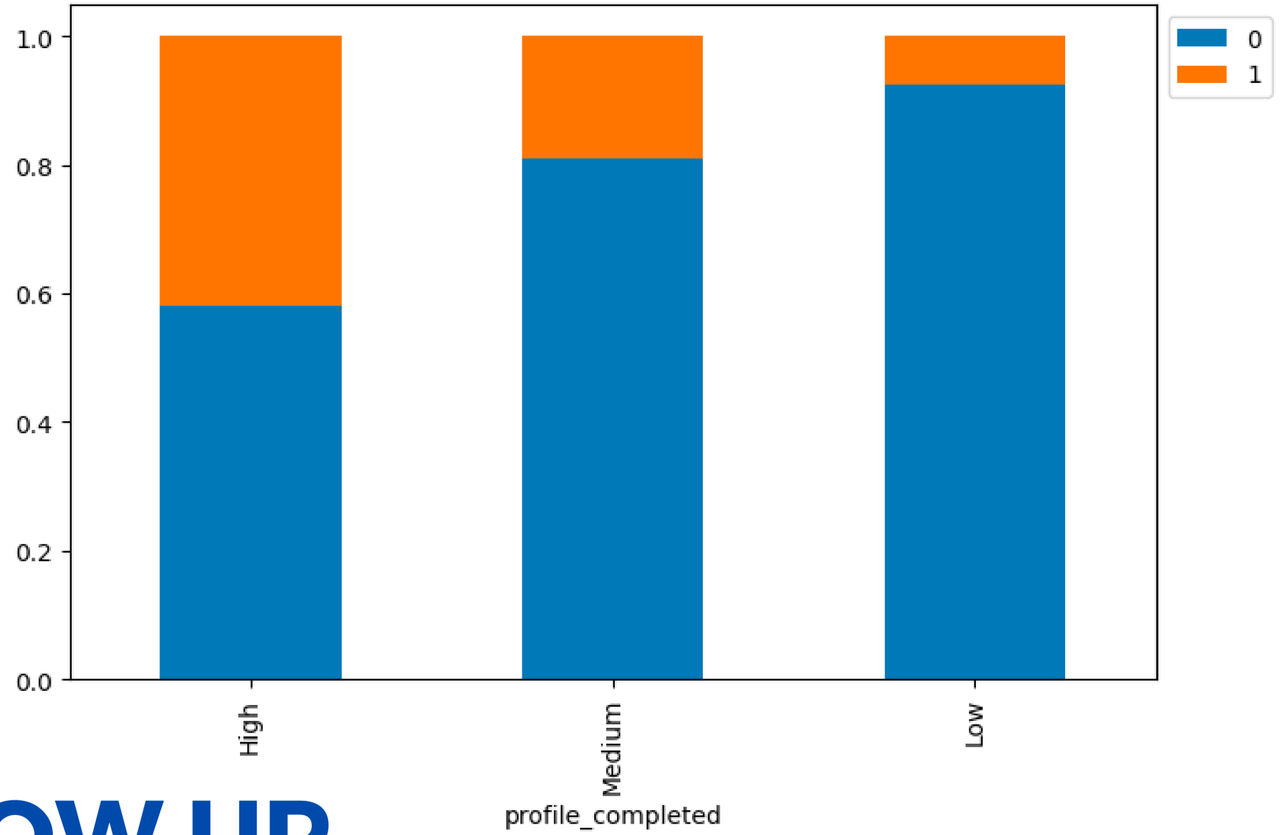


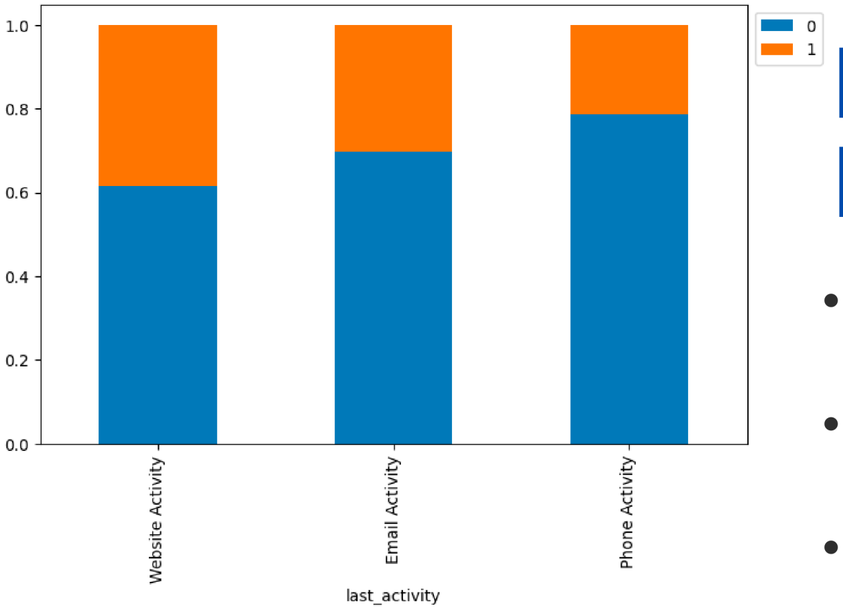




PROFILE COMPLETION

- A "**High**" profile completion level may indicate greater user engagement, but it seems to be more associated with unpaid status.
- "**Medium**" profile completion levels might be associated with a more balanced distribution between unpaid and paid statuses.

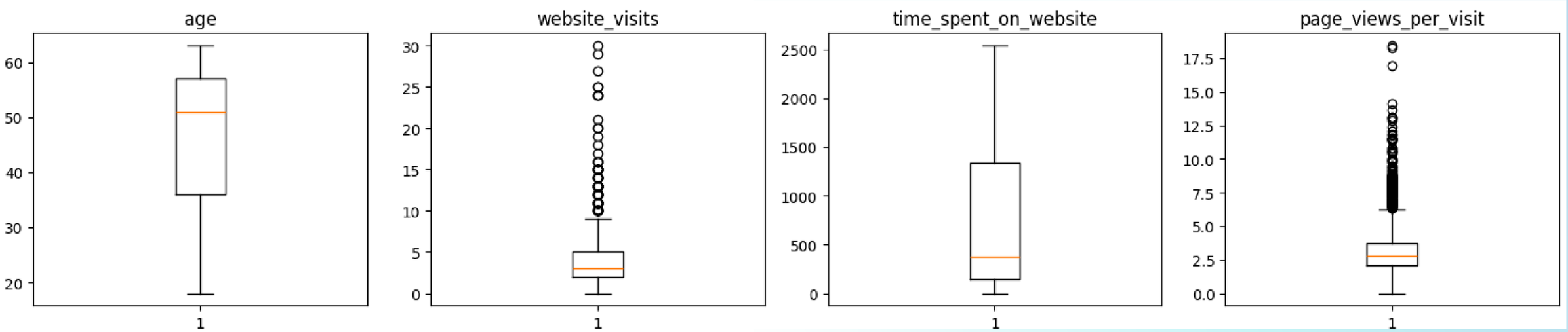




FOLLOW UP INTERACTIONS

- **Email Activity** might be a common interaction, but it seems to be associated more with unpaid customers.
- **Website Activity** has a more balanced distribution of status, indicating potential effectiveness in engaging both unpaid and paid customers.
- **Phone Activity** is more associated with status 0, suggesting that this mode of interaction may be less effective in converting users to paid status.

OUTLIERS



From the outlier analysis:

- Age: **0** outliers
- Website Visits: **154** outliers
- Time Spent on Website: **0** outliers
- Page Views Per Visit: **40** outliers
- This indicates that 'website_visits' and 'page_views_per_visit' columns have a number of outlier values that we may need to handle or investigate

MODEL BUILDING

Company would want Recall to be maximized, greater the Recall score higher are the chances of minimizing False Negatives.

Using Decision Tree and Random Forest Classifiers

1. Predicting a lead **will not** be converted to a paid customer in reality, the lead would have converted to a paid customer.
2. Predicting a lead **will** be converted to a paid customer in reality, the lead would not have converted to a paid customer.

Decision Tree

Testing Round #1:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	2273
1	1.00	1.00	1.00	955
accuracy			1.00	3228
macro avg	1.00	1.00	1.00	3228
weighted avg	1.00	1.00	1.00	3228

- The Decision Tree Classifier has achieved perfect precision, recall and F1-score of 1.00 on the training data, which shows it has memorized the training dataset.
- Confusion matrix confirms that there are no misclassifcations in the training set

Testing Round #2:

	precision	recall	f1-score	support
0	0.87	0.86	0.86	962
1	0.69	0.70	0.70	422
accuracy			0.81	1384
macro avg	0.78	0.78	0.78	1384
weighted avg	0.81	0.81	0.81	1384

- The accuracy on the test data is approximately 84.61%
- While the model has perfect performance on the training data, it's performance on unseen data is lower, which is indicative of overfitting

DECISION TREE - HYPERPARAMETER TUNING

Train Data - Tuned

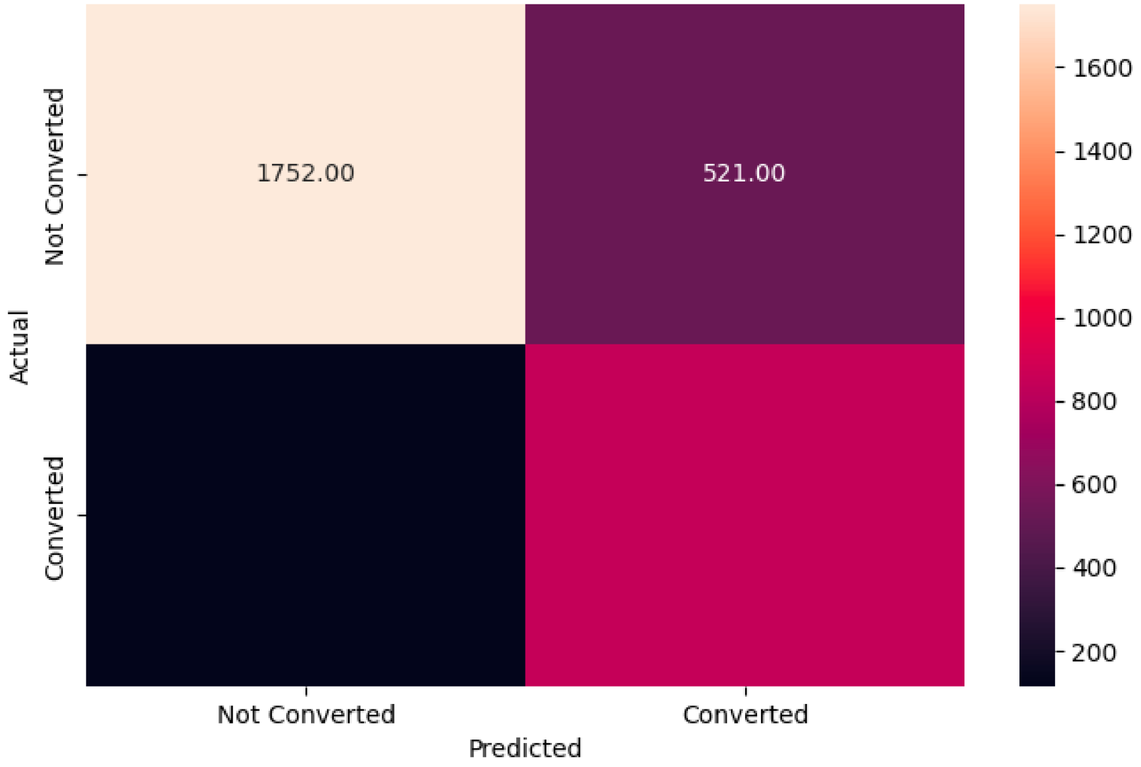
Class 0 (Unpaid Customers):

- **Precision: 94%, Recall: 77%, F1-Score: 85%**
- The model effectively identifies unpaid customers with high precision but shows a trade-off with lower recall.
- This shows a reduction in overfitting compared to the previous model, as the gap between training and test accuracy has decreased

Class 1 (Paid Customers):

- **Precision: 62%, Recall: 88%, F1-Score: 73%**
- The model captures paid customers well with high recall but at the expense of lower precision.

	precision	recall	f1-score	support
0	0.94	0.77	0.85	2273
1	0.62	0.88	0.73	955
accuracy			0.80	3228
macro avg	0.78	0.83	0.79	3228
weighted avg	0.84	0.80	0.81	3228



DECISION TREE - HYPERPARAMETER TUNING

Test Data - Tuned

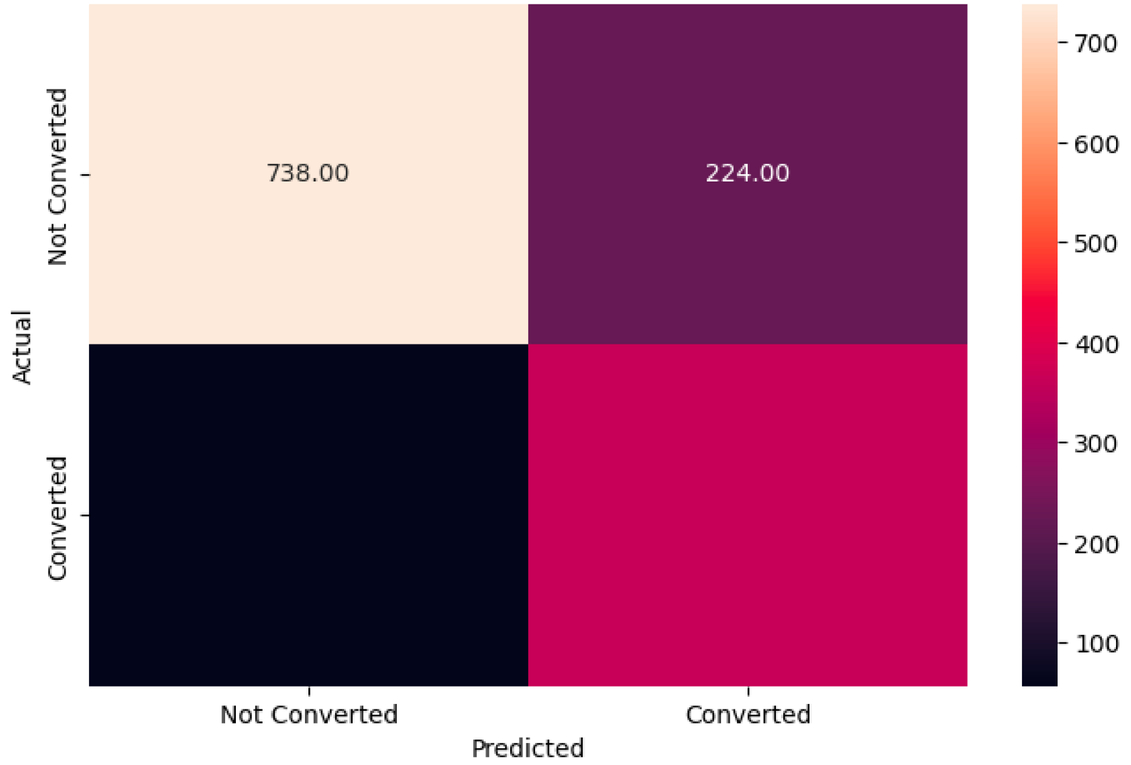
	precision	recall	f1-score	support
0	0.93	0.77	0.84	962
1	0.62	0.86	0.72	422
accuracy			0.80	1384
macro avg	0.77	0.82	0.78	1384
weighted avg	0.83	0.80	0.80	1384

Class 0 (Unpaid Customers):

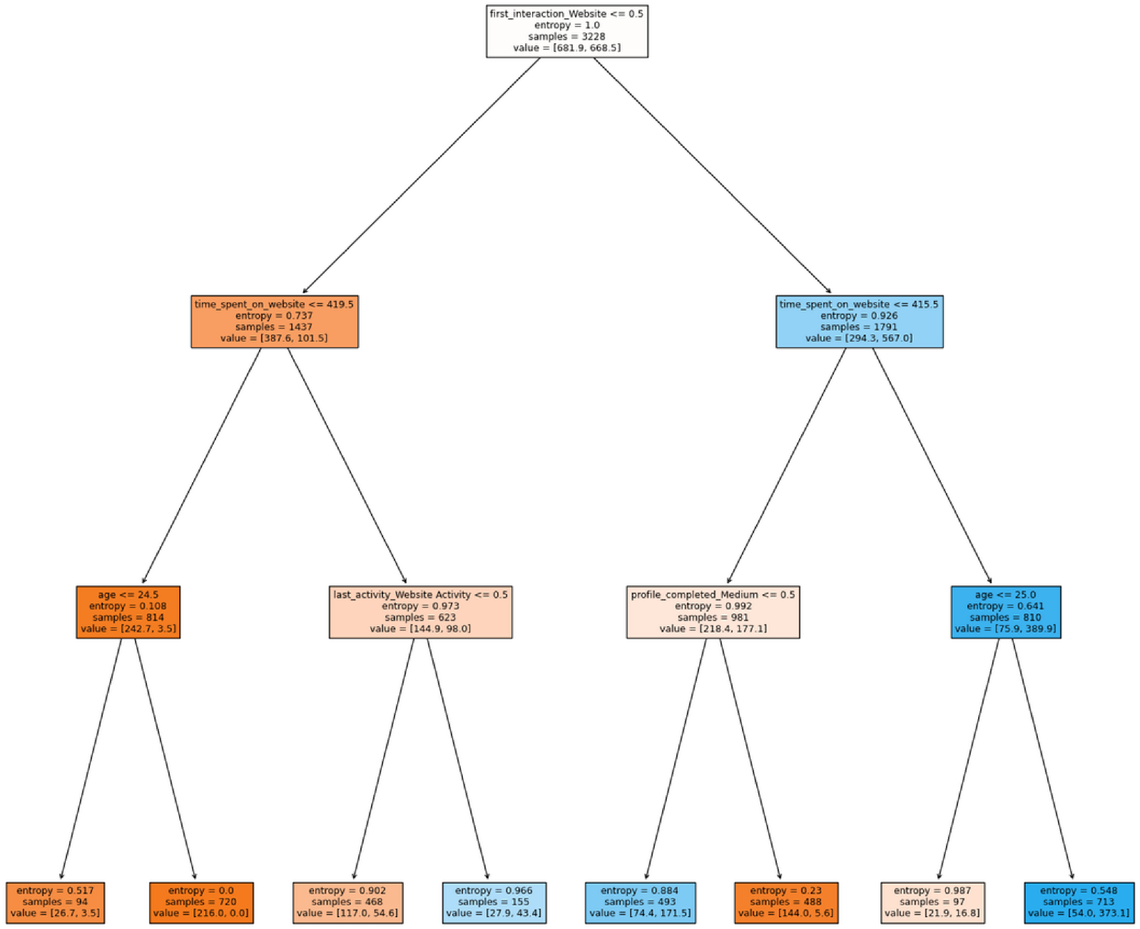
- **Precision: 93%, Recall: 77%, F1-Score: 84%**
- The model effectively identifies unpaid customers with high precision but shows a trade-off with lower recall.

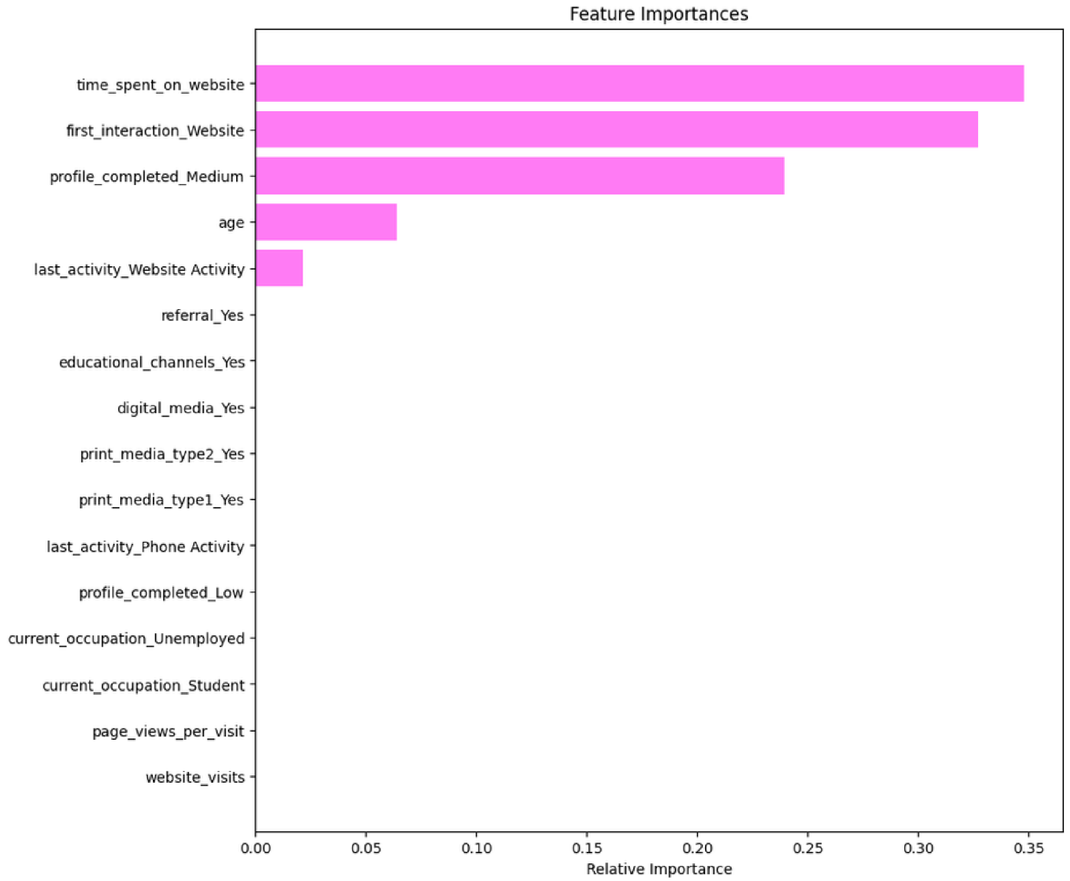
Class 1 (Paid Customers):

- **Precision: 62%, Recall: 86%, F1-Score: 72%**
- The model captures paid customers well with high recall but at the expense of lower precision.
- F1-score suggests an imbalance between precision and recall



VISUALIZE DECISION TREE





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

Random Forest Classifier

Testing Round #1:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	2273
1	1.00	1.00	1.00	955
accuracy			1.00	3228
macro avg	1.00	1.00	1.00	3228
weighted avg	1.00	1.00	1.00	3228

- Again, like the decision tree, the random forest classifier has achieved perfect precision, recall and F1-score of 1.00
- Confusion matrix confirms that there are no misclassifcations in the training set which means we need to reassess its generalization ability

Testing Round #2:

	precision	recall	f1-score	support
0	0.87	0.92	0.89	962
1	0.79	0.68	0.73	422
accuracy			0.85	1384
macro avg	0.83	0.80	0.81	1384
weighted avg	0.84	0.85	0.84	1384

- Recall sensitivity identifies at 92% for class 0 (non-converted customers) and only captures 68% of the instances for class 1.
- Suggests that our model might be better at identifying non-converted customers than converted
- The F1-score indicates a balance between precision and recall
- Unpaid customers has a higher F1-score suggesting more balance than class 1
- Overall accuracy is 85% which shows good performance, with high precision and recall for class 0.
- There is room for imporovement for correctly identifying instances of class 1

RANDOM FOREST CLASSIFIER - HYPERPARAMETER TUNING

Train Data - Tuned

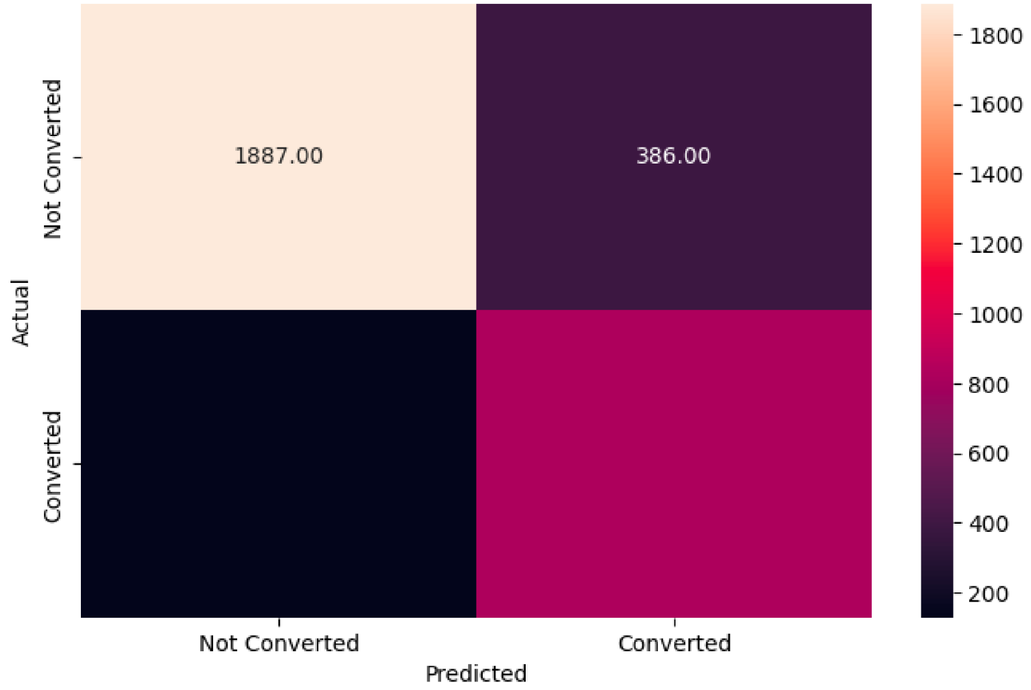
Class 0 (Unpaid Customers):

- **Precision** (Positive Predictive Value): **94%**
- **Recall** (Sensitivity or True Positive Rate): **83%**
- F1-Score: **88%**
- The number of actual instances of class 0 is **2273**.

Class 1 (Paid Customers):

- **Precision** (Positive Predictive Value): **68%**
- **Recall** (Sensitivity or True Positive Rate): **86%**
- F1-Score: **76%**
- The number of actual instances of **class 1 is 955**.

The overall accuracy of the Random Forest model in correctly predicting both classes is **84%.**

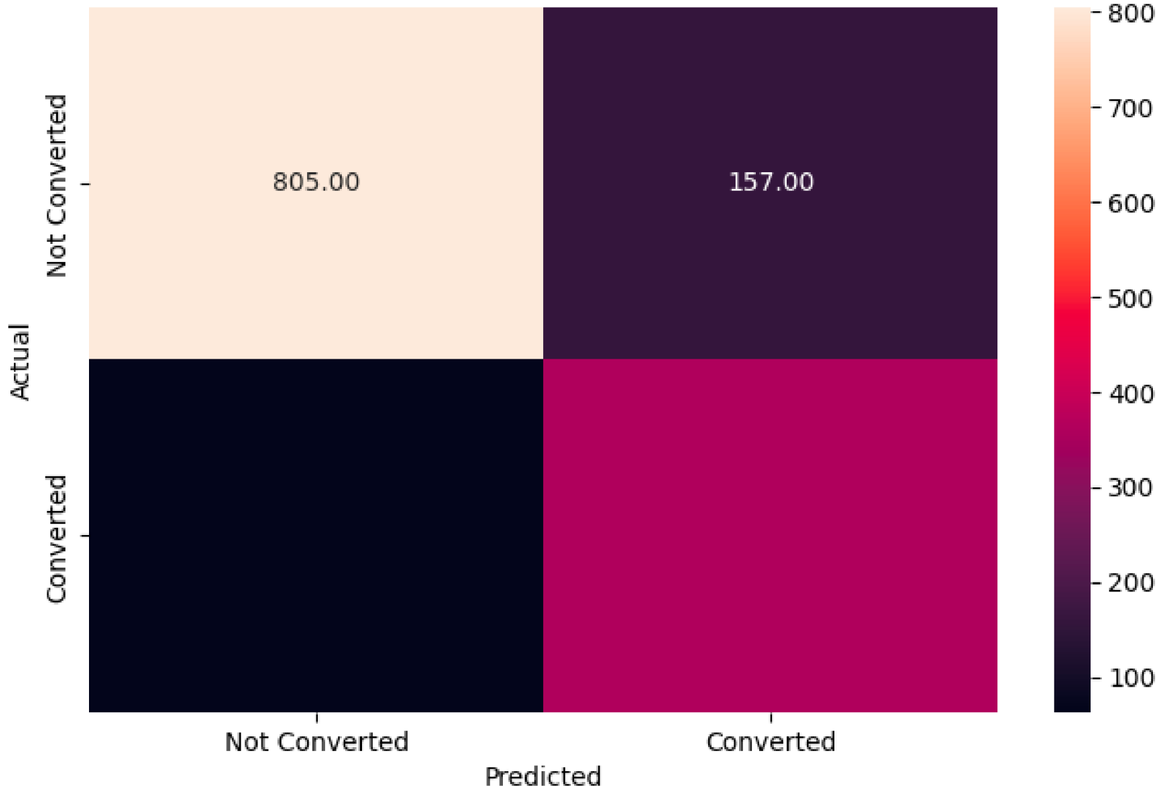


	precision	recall	f1-score	support
0	0.94	0.83	0.88	2273
1	0.68	0.86	0.76	955
accuracy			0.84	3228
macro avg	0.81	0.85	0.82	3228
weighted avg	0.86	0.84	0.85	3228

RANDOM FOREST CLASSIFIER - HYPERPARAMETER TUNING

Test Data - Tuned

	precision	recall	f1-score	support
0	0.93	0.84	0.88	962
1	0.70	0.85	0.77	422
accuracy			0.84	1384
macro avg	0.81	0.84	0.82	1384
weighted avg	0.86	0.84	0.84	1384



Class 0 Performance:

- The model performs well in predicting class 0 with high precision (93%) and good recall (84%). This suggests that the model is effective in correctly identifying instances of class 0.

Class 1 Performance:

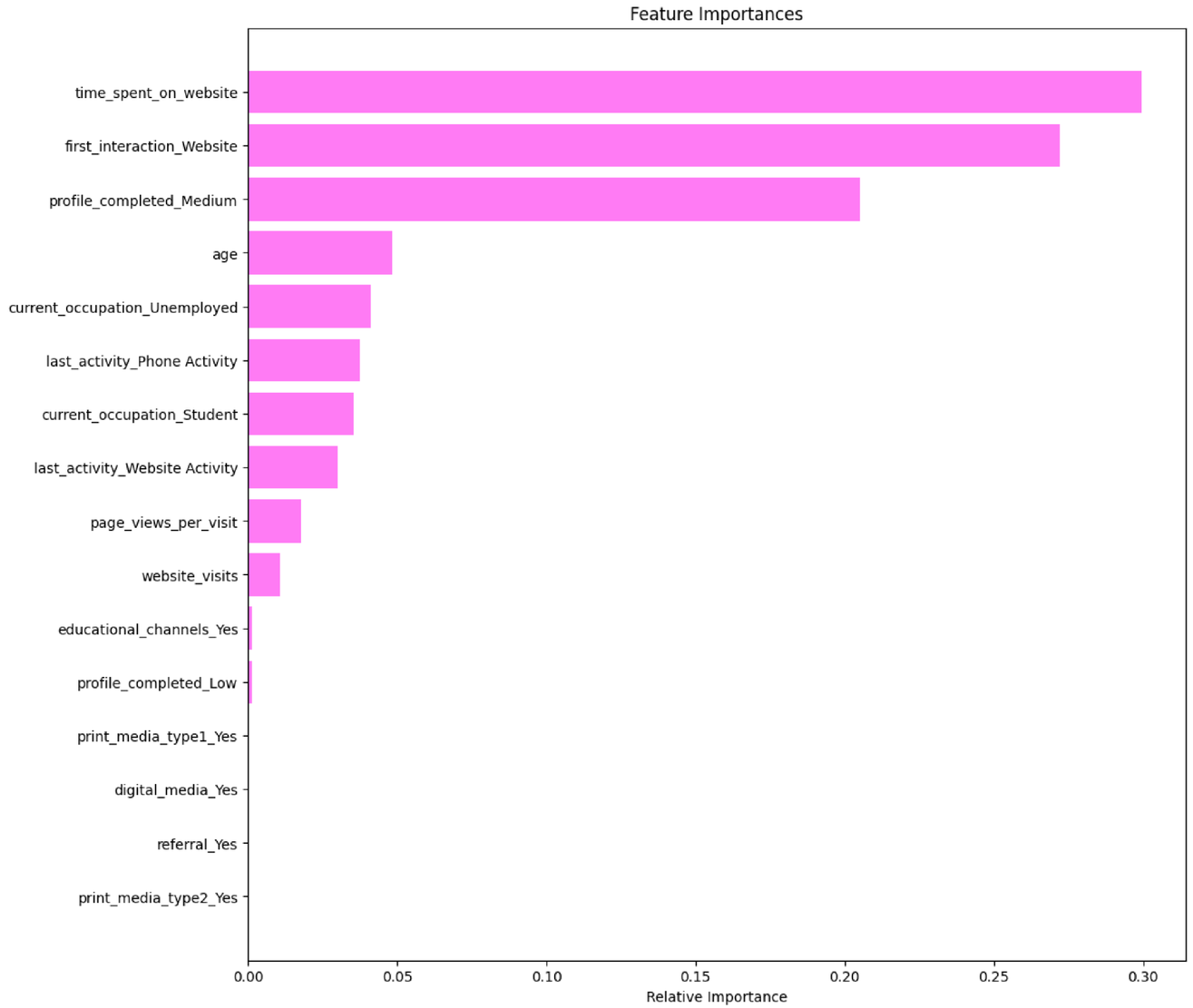
- The model performs reasonably well in predicting class 1 with a precision of 70% and a higher recall of 85%. There is room for improvement in precision for class 1, indicating a higher rate of false positives.

There is an imbalance in the number of instances between class 0 (unpaid customers) and class 1 (paid customers). Class 0 has a higher support (962) compared to class 1 (422).

Can consider balancing precision and recall for both classes

FEATURE IMPORTANCE

- Similar to the decision tree model, **time spent on 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.



CONCLUSIONS

DECISION TREE

- The tuned Decision Tree model exhibits a trade-off between precision and recall for both converted and non-converted customers, achieving an **accuracy of 80%.**
- Compared to the untuned Decision Tree, the tuned model shows **improved recall** for paid customers, indicating better identification of converted customers.
- The model faces challenges in precision for unpaid customers, **suggesting potential false positives** among predicted unpaid customers.

RANDOM FOREST CLASSIFIER

- The Random Forest model demonstrates **balanced** performance with an **accuracy of 84%**, indicating its ability to effectively classify both converted and non-converted customers.
- Precision for paid customers (Class 1) is higher, **reducing the likelihood of false positives**. This is particularly important for minimizing the misclassification of potential paying customers.
- The model maintains a **high recall** for unpaid customers (Class 0), suggesting its effectiveness in identifying most non-converted customers.
- The success of the Random Forest model suggests potential benefits from exploring other ensemble methods to further enhance predictive performance.

BUSINESS RECOMMENDATIONS

- A focus on website improvements may be particularly beneficial for improving the user experience and engagement, thereby influencing conversion rates positively.
- Bounce-rates and time spent on web-site seem effective and could focus on these metrics of data for further model building
- Can further identify specific aspects of website interactions or user behaviors that significantly influence conversion rates
- Could consider additional analysis to understand the specific user journeys that lead to misclassification
- Could identify specific areas of the website that may benefit from improvements, focusing on aspects that contribute to higher conversion rates such as follow up interaction, CTA buttons on web homepage that incentivize leads to complete their profile
- Establishing a feedback mechanism to collect user feedback on website changes which could provide qualitative insights into user satisfaction
- Continuous monitoring performance of the website and updating the model as needed as well as refining the areas regarding follow-up interactions through email