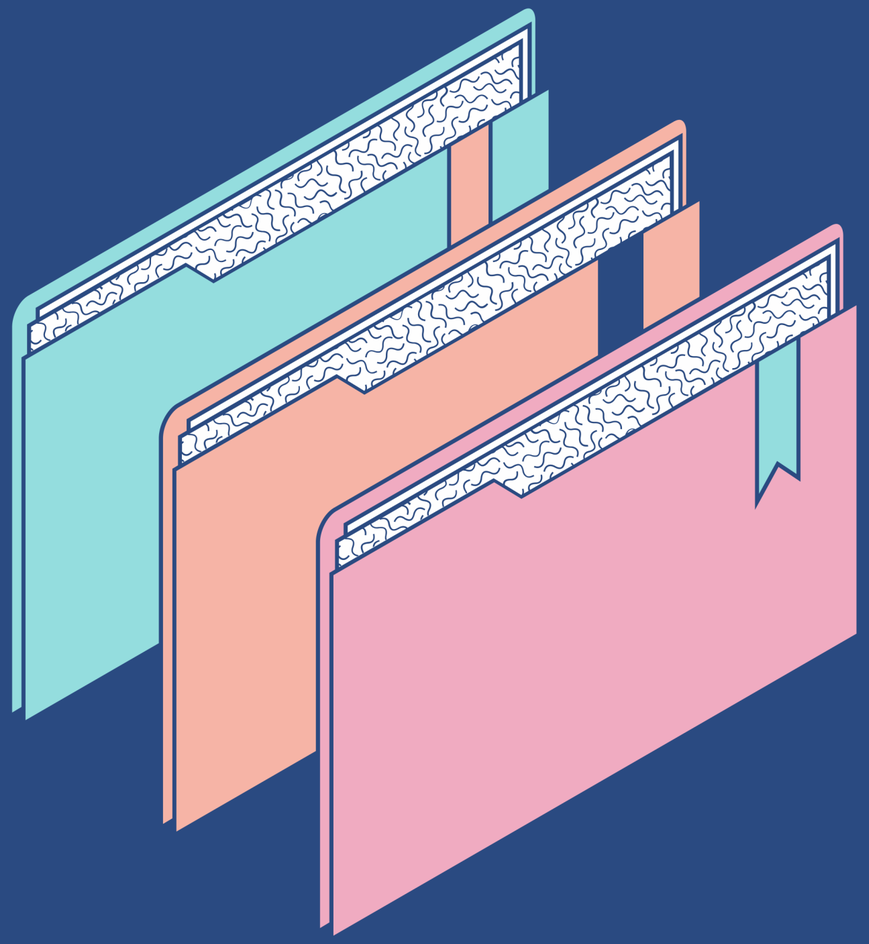
Group 8

TEACHER:
NGUYỄN HUY ANH



Member:

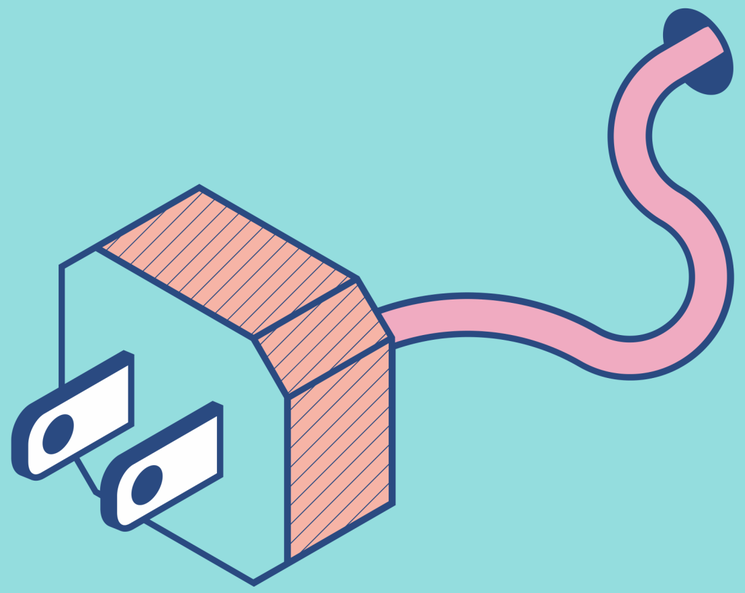
NGUYỄN QUANG HUY

NGUYỄN TRỌNG VINH

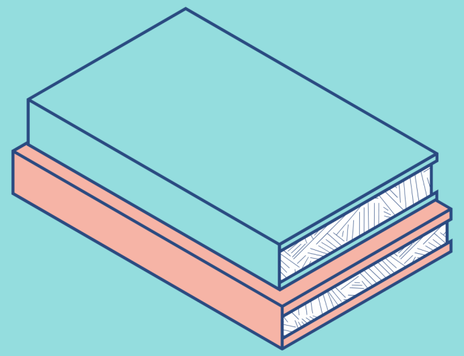
HOÀNG QUANG MINH

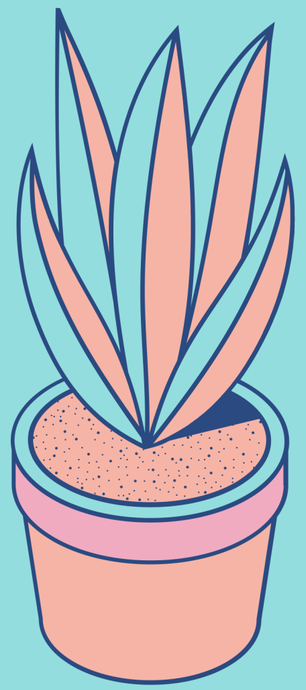
Introduction

ExtraaLearn is a startup in the EdTech industry that focuses on providing innovative educational solutions. The goal is to revolutionize the way people learn by leveraging technology and personalized learning experiences. By offering a range of online courses and educational resources to help individuals enhance their skills and knowledge in various fields



With a large number of leads being generated on a regular basis, one of the issues faced by ExtraaLearn is to identify which of the leads are more likely to convert so that they can allocate the resources accordingly





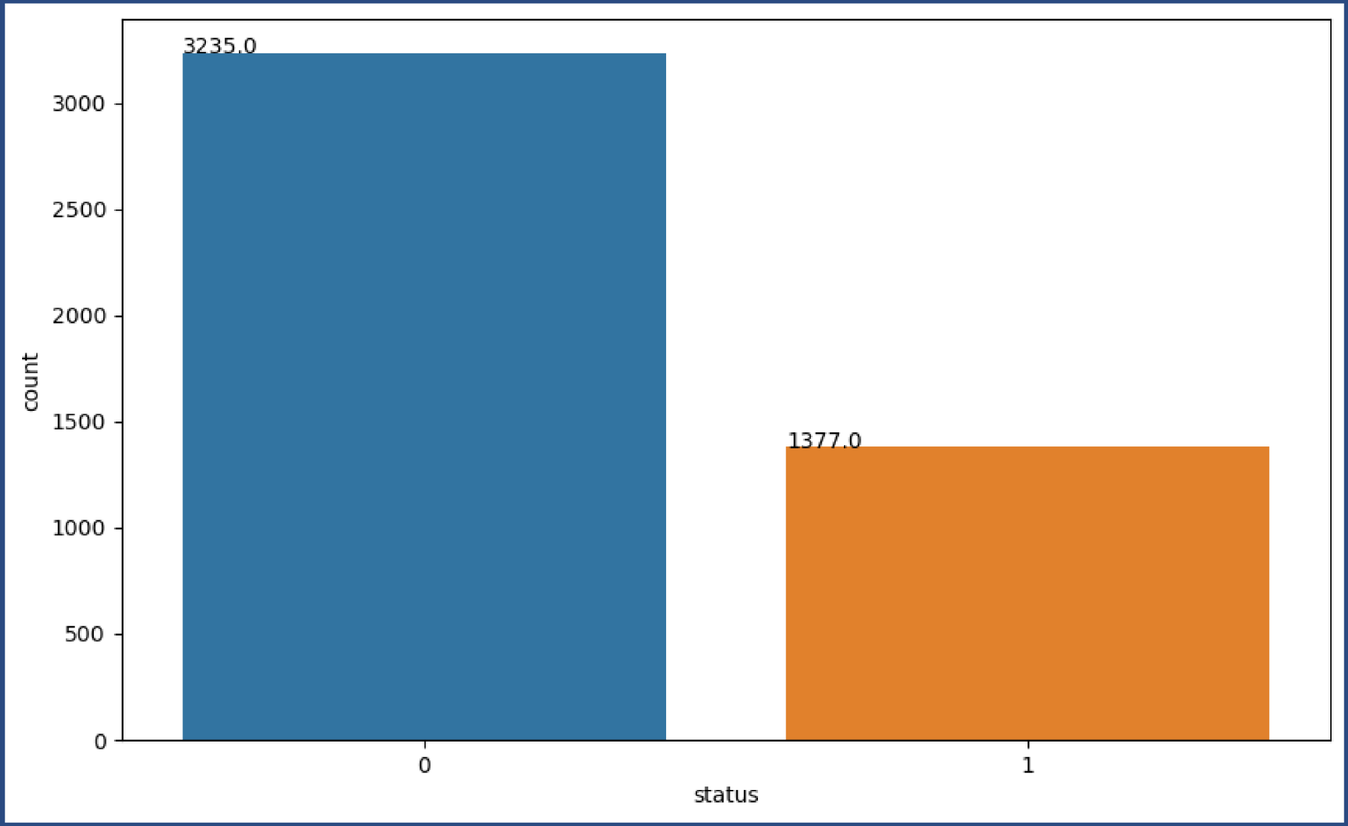
We will perform exploratory data analysis and built classification models using Decision Tree and Random Forest algorithms to identify which of the leads are more likely to convert.

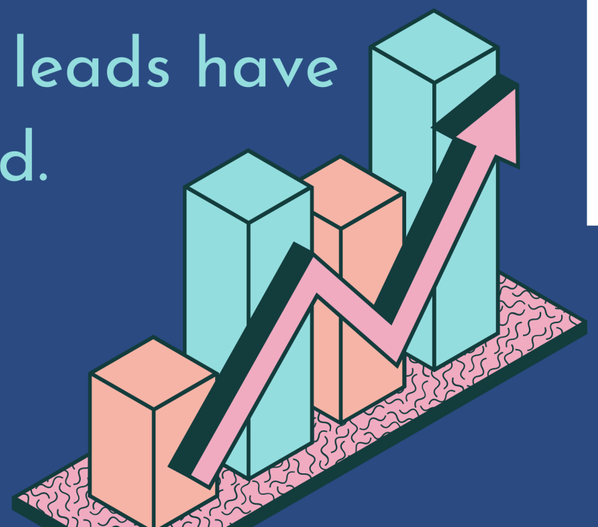
The data includes various customer attributes such as ID, age, current occupation, initial access method, profile completion level, number of visits, total time spent, average pages viewed per visit, last interaction, email activities, phone activities, website activities, and other advertising factors.

	age	current_occupation	first_interaction	profile_completed	website_visits	time_spent_on_website	page_views_per_visit	last_activity	print_me
0	57	Unemployed	Website	High	7	1639	1.861	Website Activity	
1	56	Professional	Mobile App	Medium	2	83	0.320	Website Activity	
2	52	Professional	Website	Medium	3	330	0.074	Website Activity	
3	53	Unemployed	Website	High	4	464	2.057	Website Activity	
4	23	Student	Website	High	4	600	16.914	Email Activity	
...	...	...	...	...	...	...	...	...	
4607	35	Unemployed	Mobile App	Medium	15	360	2.170	Phone Activity	
4608	55	Professional	Mobile App	Medium	8	2327	5.393	Email Activity	
4609	58	Professional	Website	High	2	212	2.692	Email Activity	
4610	57	Professional	Mobile App	Medium	1	154	3.879	Website Activity	
4611	55	Professional	Website	Medium	4	2290	2.075	Phone Activity	

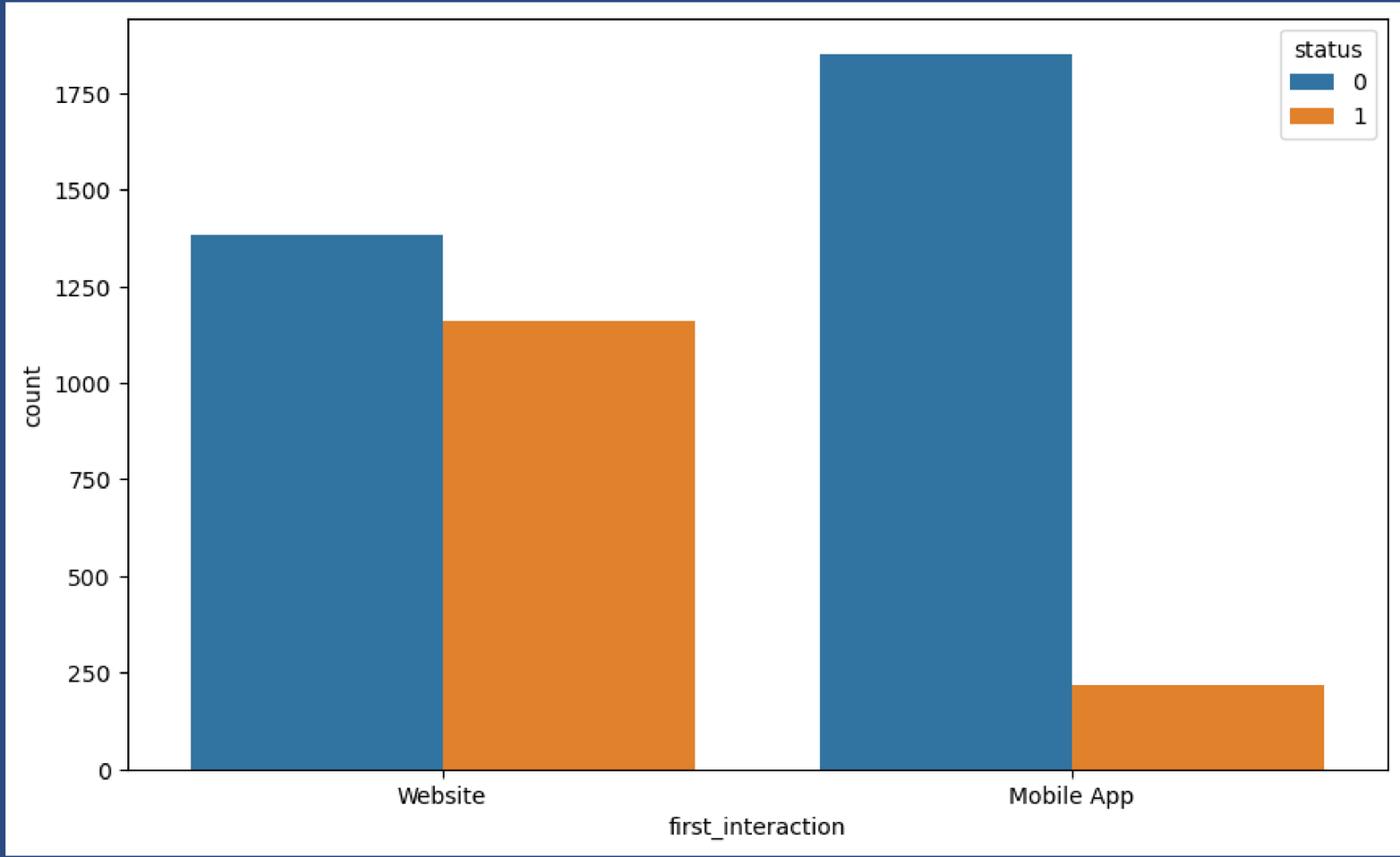
visits	time_spent_on_website	page_views_per_visit	last_activity	print_media_type1	print_media_type2	digital_media	educational_channels	referral	status
7	1639	1.861	Website Activity	Yes	No	Yes	No	No	1
2	83	0.320	Website Activity	No	No	No	Yes	No	0
3	330	0.074	Website Activity	No	No	Yes	No	No	0
4	464	2.057	Website Activity	No	No	No	No	No	1
4	600	16.914	Email Activity	No	No	No	No	No	0
...	...	...	...	...	...	...	...	...	...
15	360	2.170	Phone Activity	No	No	No	Yes	No	0
8	2327	5.393	Email Activity	No	No	No	No	No	0
2	212	2.692	Email Activity	No	No	No	No	No	1
1	154	3.879	Website Activity	Yes	No	No	No	No	0

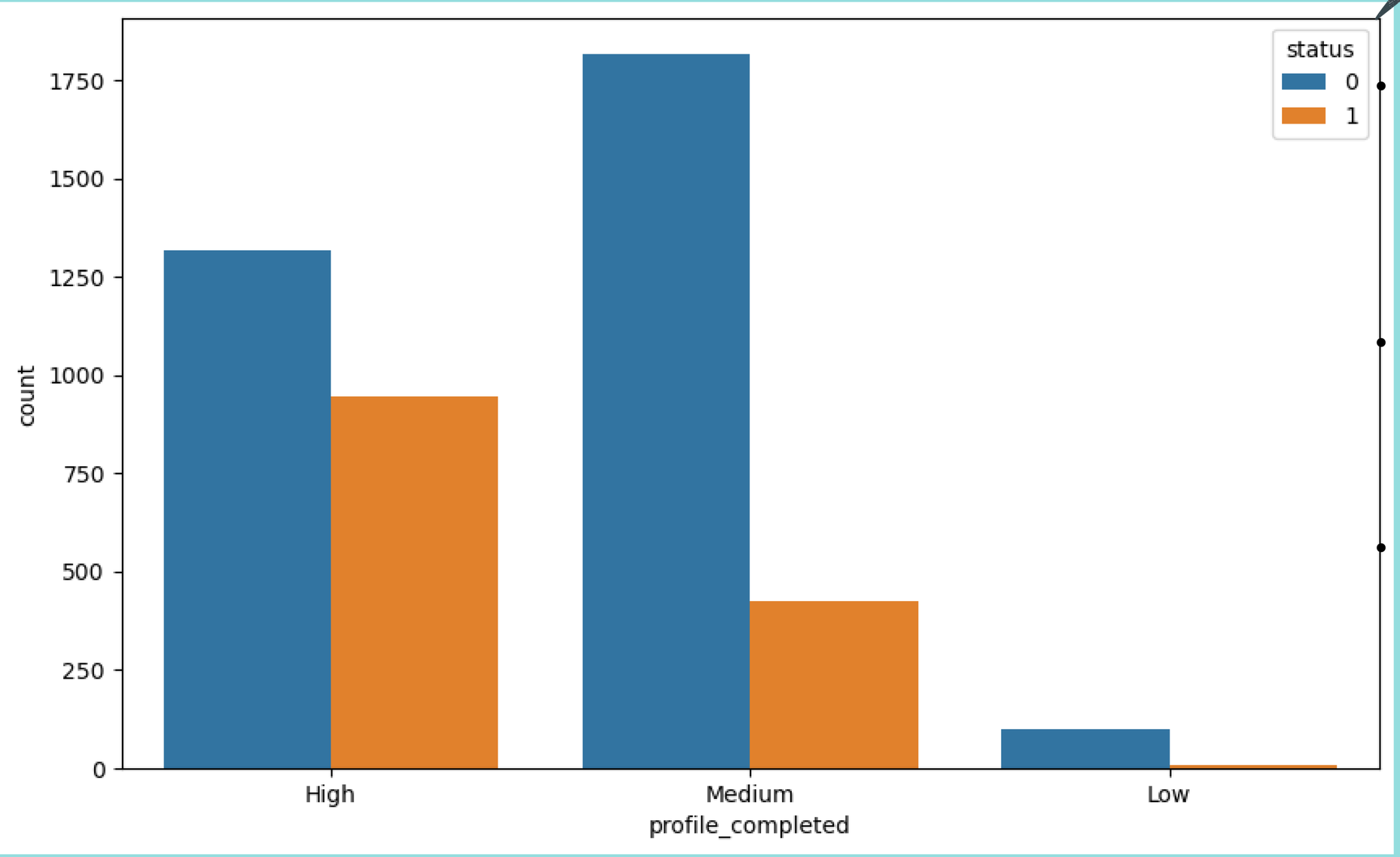
- The above plot shows that number of leads converted are significantly less than number of leads not converted which can be expected.
- The plot indicates that ~30% (1377/4612) of leads have been converted.



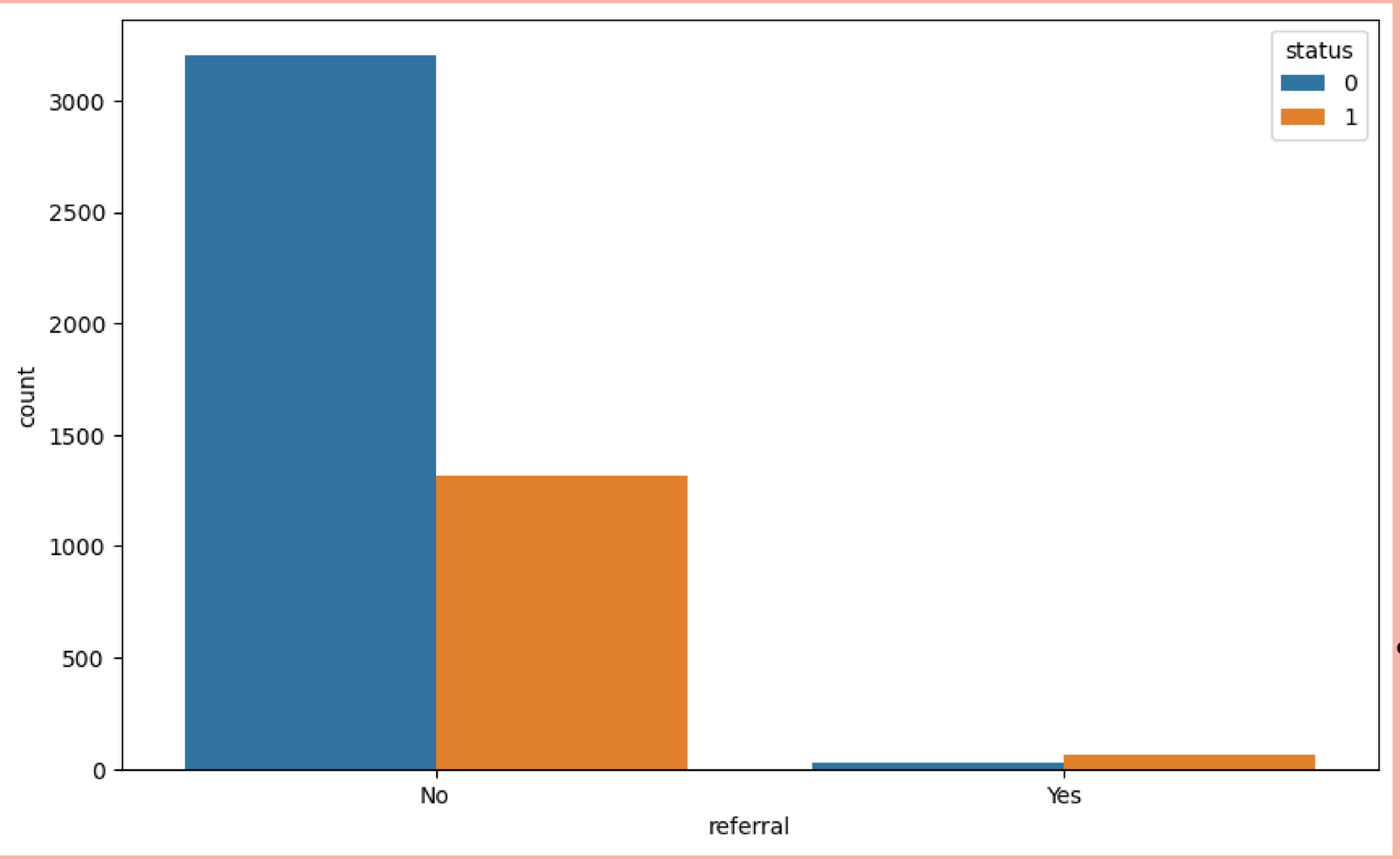


- The website seems to be doing a good job as compared to mobile app as there is a huge difference in the number of conversions of the leads who first interacted with the company through website and those who interacted through mobile application.
-
- Majority of the leads who interacted through websites were converted to paid customers, while only a small number of leads, who interacted through mobile app, converted.



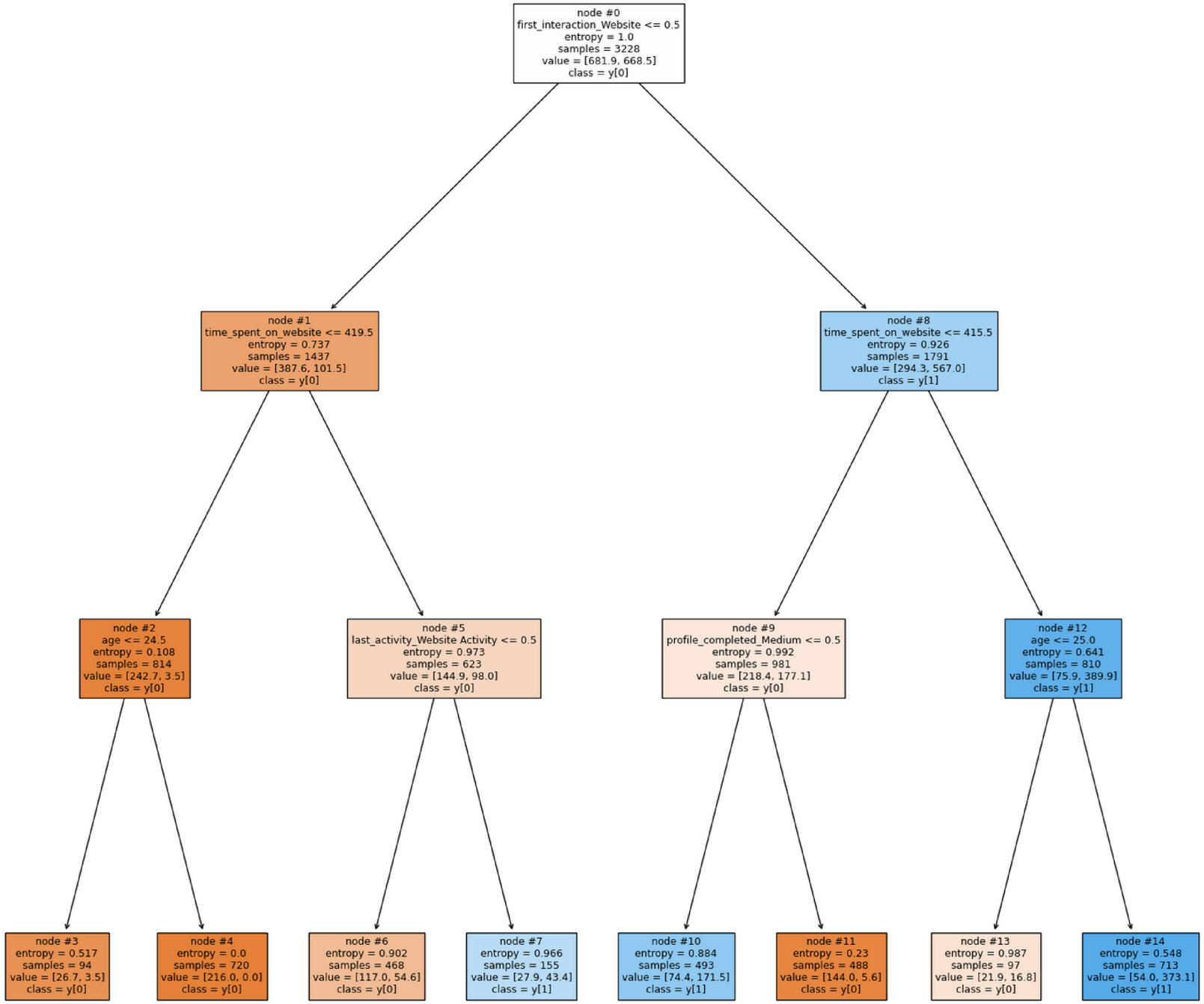


- The leads whose profile completion level is high converted more in comparison to other levels of profile completion.
- The medium and low levels of profile completion saw comparatively very less conversions.
- The high level of profile completion might indicate a lead's intent to pursue the course which results in high conversion.

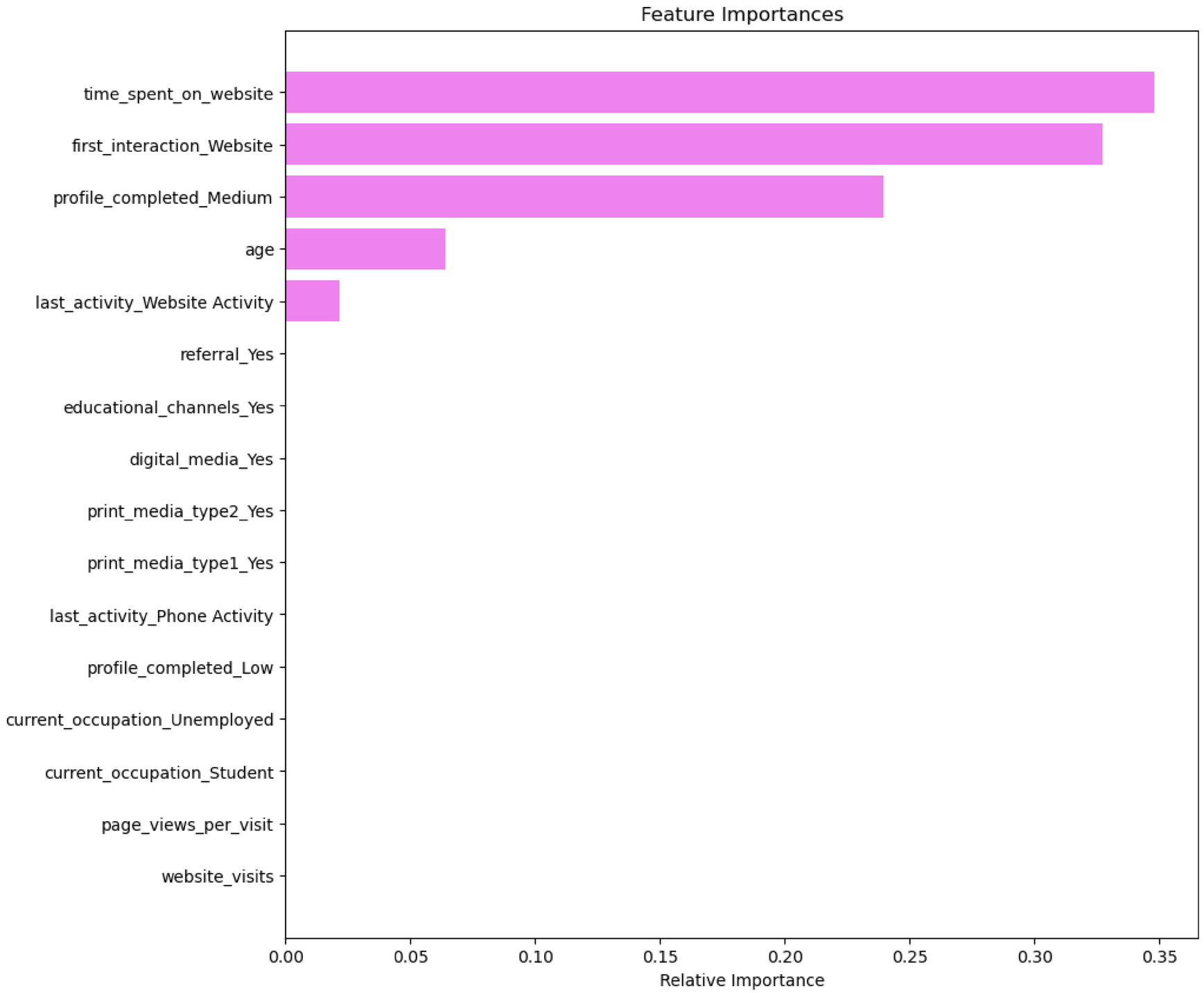


- **Company should try to get more leads through referrals by promoting rewards for existing customer base when they refer someone.**

There are a very less number of referrals but the conversion is high.

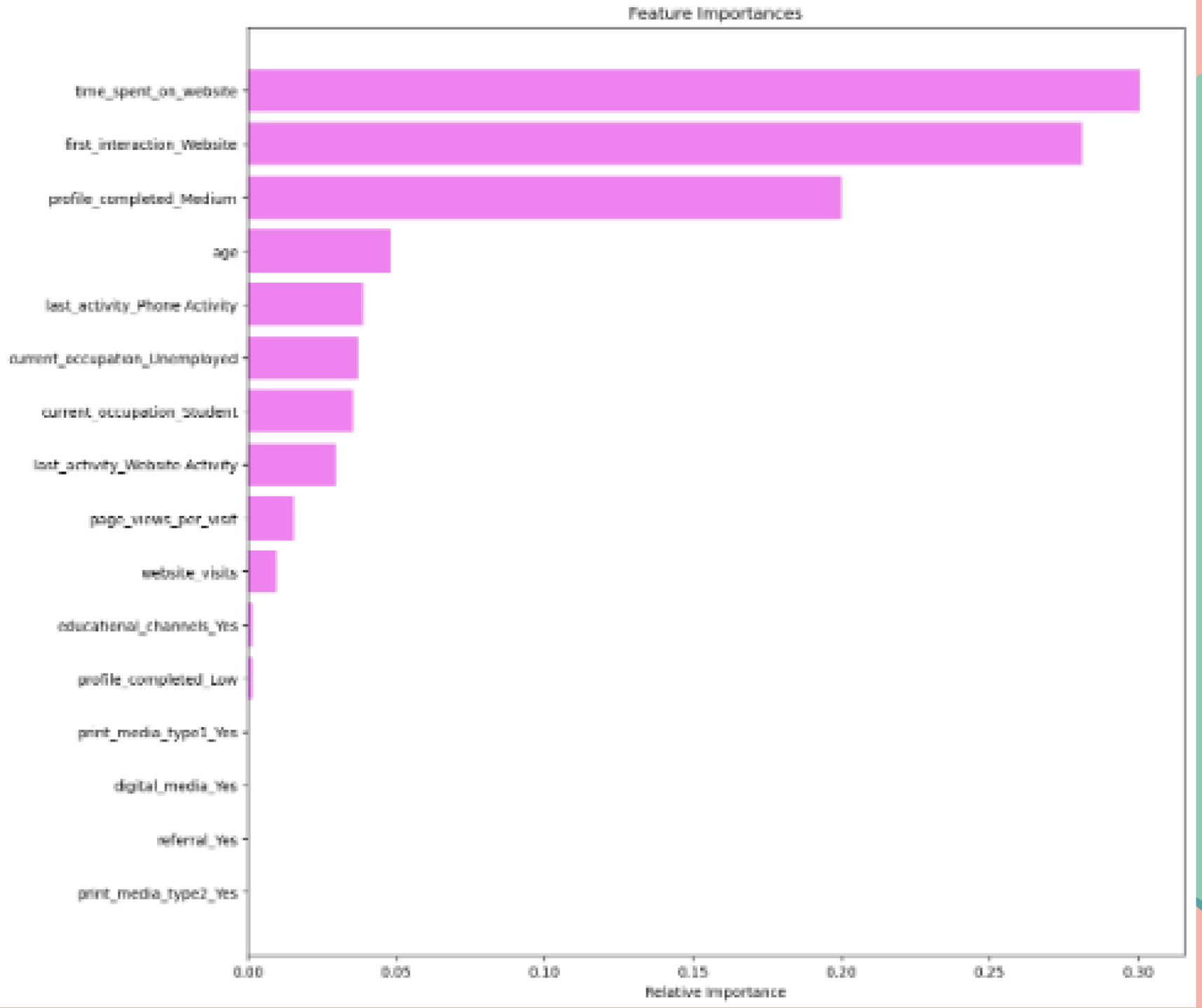


- Leads older than 25 have a higher likelihood of becoming paid customers.
- Less time spent on the website and younger age groups showed a higher likelihood of not being converted to paid customers.
- Leads who spent more time on the website have a higher conversion rate



Decision Tree feature importance

- 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.



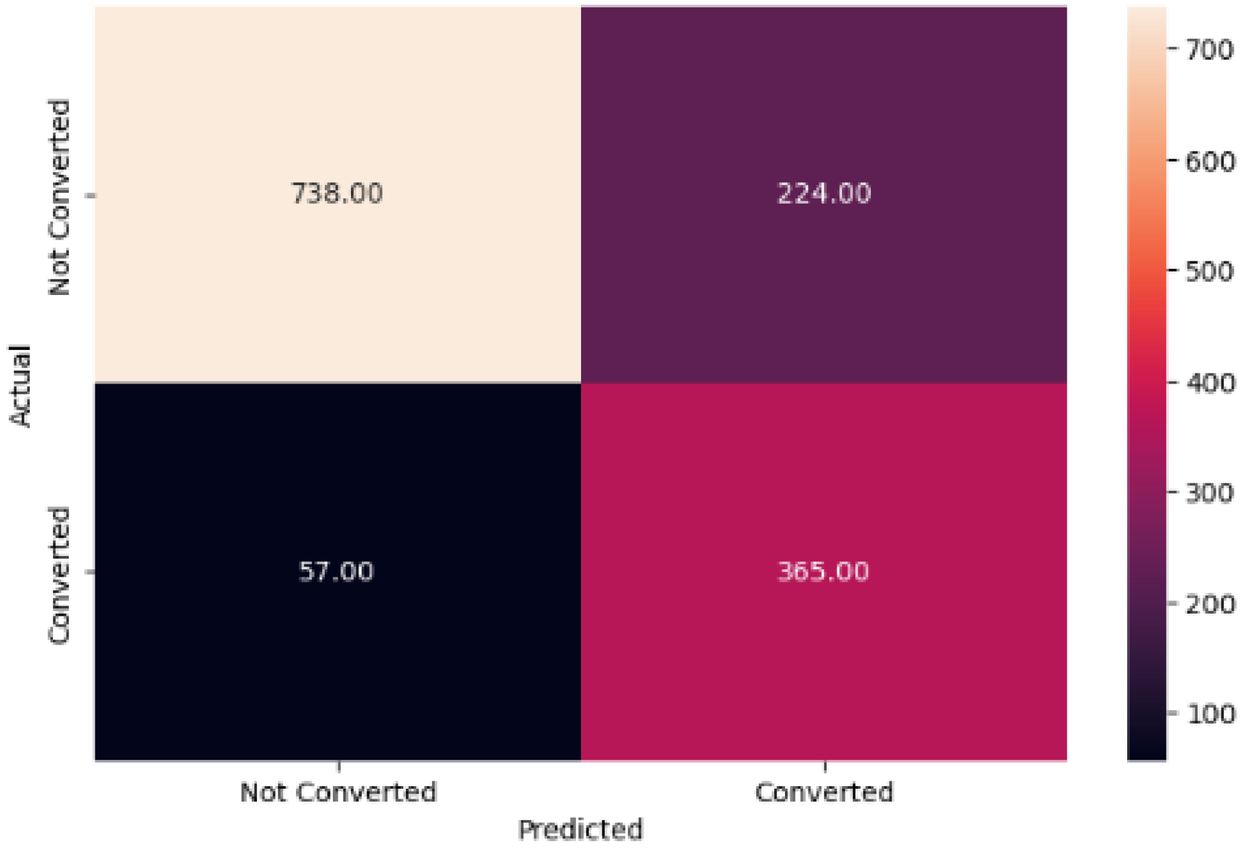
Random forest feature importance

- 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.

Decision Tree

	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

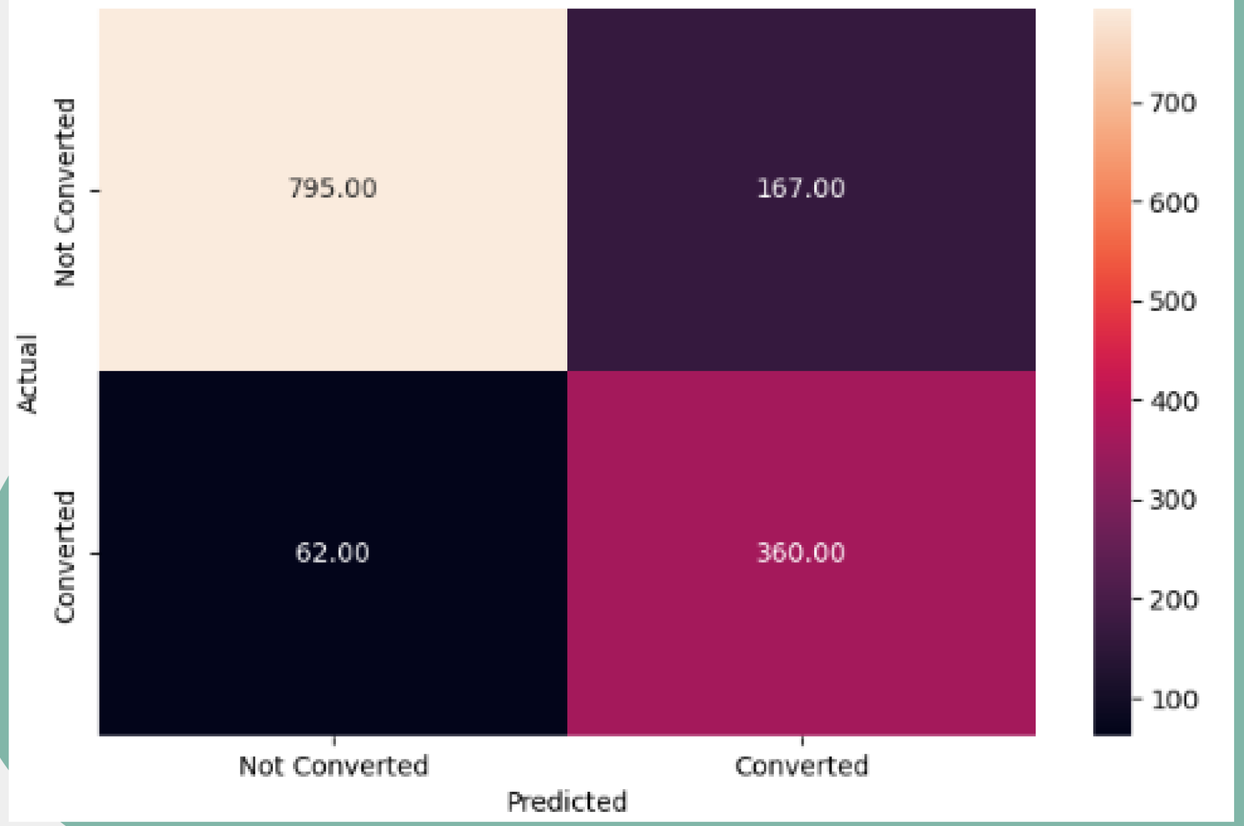


- While the precision is decreased the recall is increased.

- There has been less overfitting. On both training and testing data, the model's performance is becoming more broadly applicable.

Random forest

	precision	recall	f1-score	support
0	0.93	0.83	0.87	962
1	0.68	0.85	0.76	422
accuracy			0.83	1384
macro avg	0.81	0.84	0.82	1384
weighted avg	0.85	0.83	0.84	1384



For the random forest with default parameters, the precision, recall, and f1-score are lower.

The adjusted random forest models are a better prospective model for determining which leads will be converted since they have higher recall, f1-score, and precision.

Conclusions:

- We have been able to build tree-based models that can be used by the company to predict which leads are likely to be converted and plan their marketing strategies accordingly.
- Our goal is to maximize the Recall value. The greater the Recall score, higher the chances of minimizing False Negatives. The higher chance of predicting whether a lead will convert or not.

- Examining significant variables, we may conclude that the model's primary components are age, profile completion, time spent on the website, and time spent on the website depending on feature importances

- 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. So the random forest is better

Business Recommendations

- The time spent on the website can play a key role in identifying if a lead will be converted or not. We observed that the more time a lead spends on the website the more likely they are to convert.
- The company should spend time and money developing their website to attract and keep the attention of leads.
- Leads that filled out the profile had a higher chance of converting.
- Encouraging leads to fill out the profile with a discount might improve the number of profiles completed.
- A lead had a higher conversion rate is the age between the ages of 55 and 65.
- ·The company should market to older adults, to attract more of these types of leads.

Thank you for listening

Do you have
any question?