ĐẠI HỌC QUỐC GIA HÀ NỘI

KHOA QUẢN TRỊ VÀ KINH DOANH

FINAL REPORT
NHÓM 8

Potential Customer

LỚP	MAS1
THÀNH VIÊN	Nguyễn Quang Huy Nguyễn Trọng Vinh Hoàng Quang Minh

EXCUTIVE SUMMARY

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 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.

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.

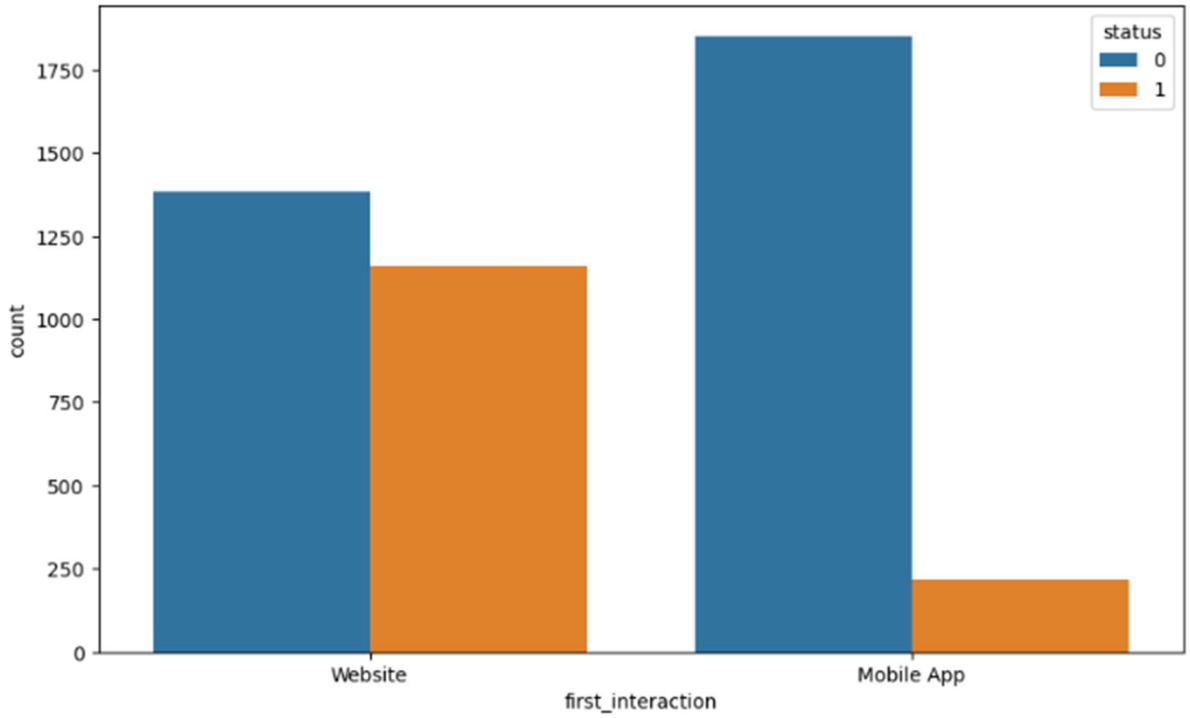
	age	current_occupation	first_interaction	profile_completed	website_visits	time_spent_on_website	page_views_per_visit	last_activity	print_media_t
0	57	Unemployed	Website	High	7	1639	1.861	Website Activity	
1	56	Professional	Mobile App	Medium	2	83	0.320	Website Activity	
2	52	Professional	Website	Medium	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	

4612 rows × 14 columns

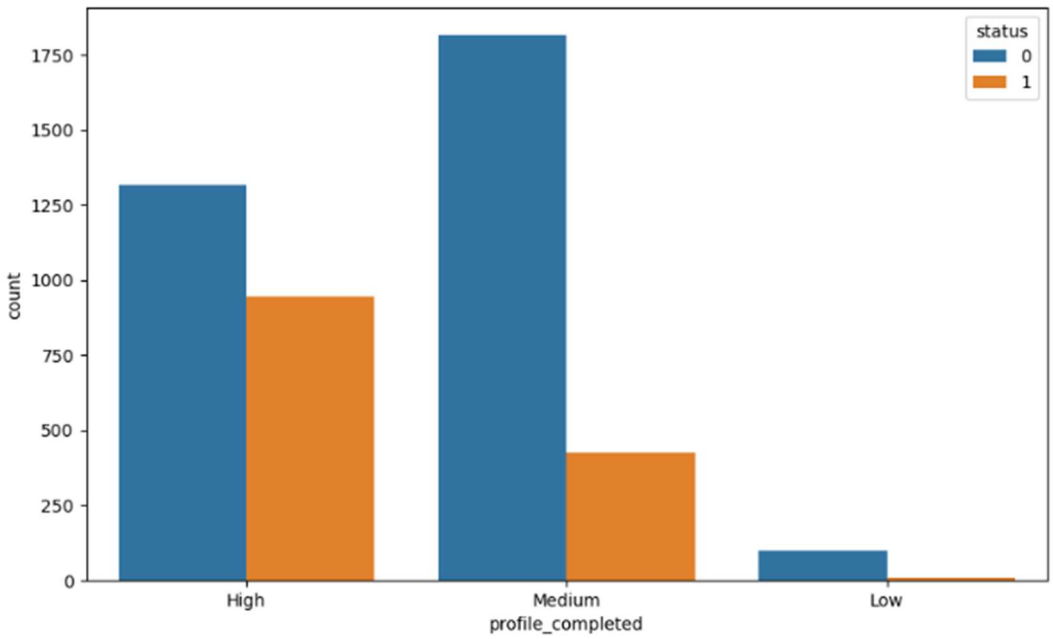
visits	time_spent_on_website	page_views_per_visit	last_activity	print_media_type1	print_media_type2	digital_media	educational_channels	referral	status
7	1639	1.861	Website Activity	Yes	No	Yes	No	No	1
2	83	0.320	Website Activity	No	No	No	Yes	No	0
3	330	0.074	Website Activity	No	No	Yes	No	No	0
4	464	2.057	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
4	2290	2.075	Phone Activity	No	No	No	No	No	0

We performed 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. We used GridSearchCV for hyperparameter tuning and evaluated the model using various metrics like confusion matrix, classification report, and recall score.

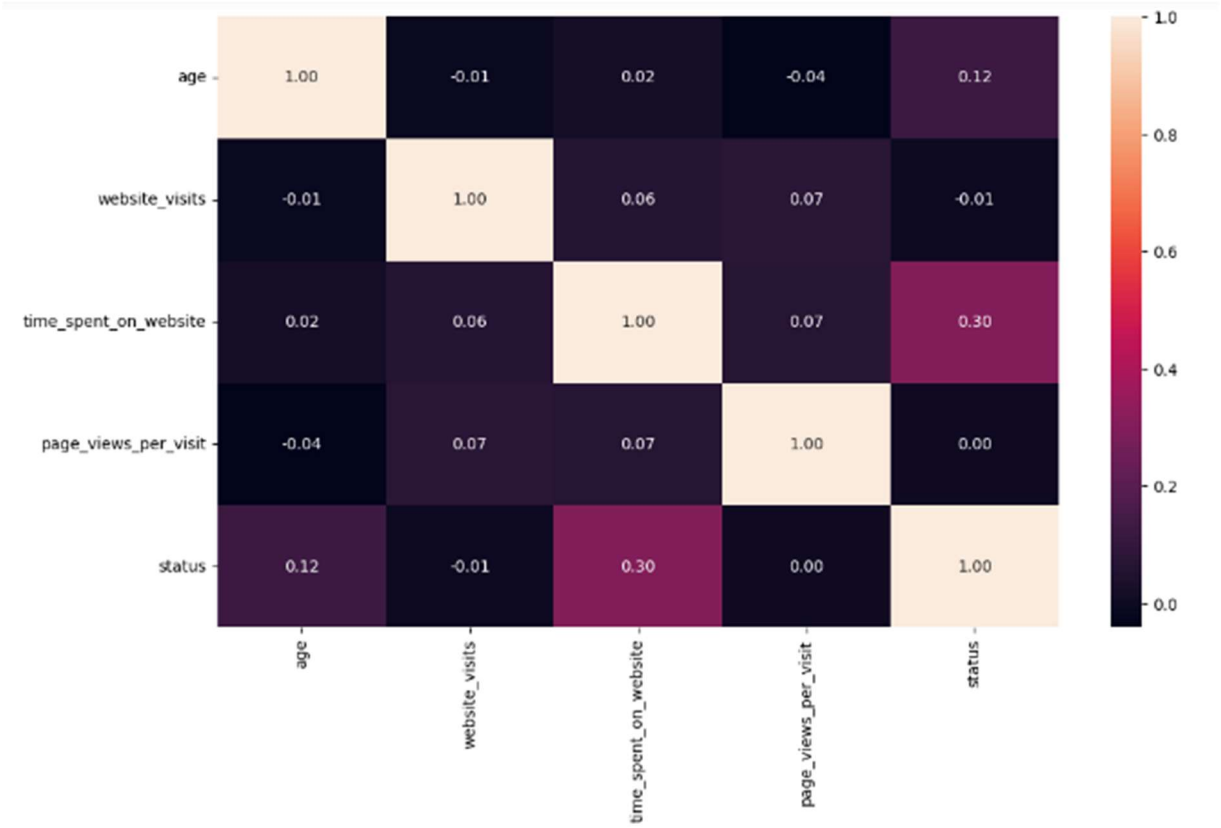
Compared between website and mobile app through status first interaction



Profile completed and status



Correlation heatmap



PROBLEM AND SOLUTION SUMMARY

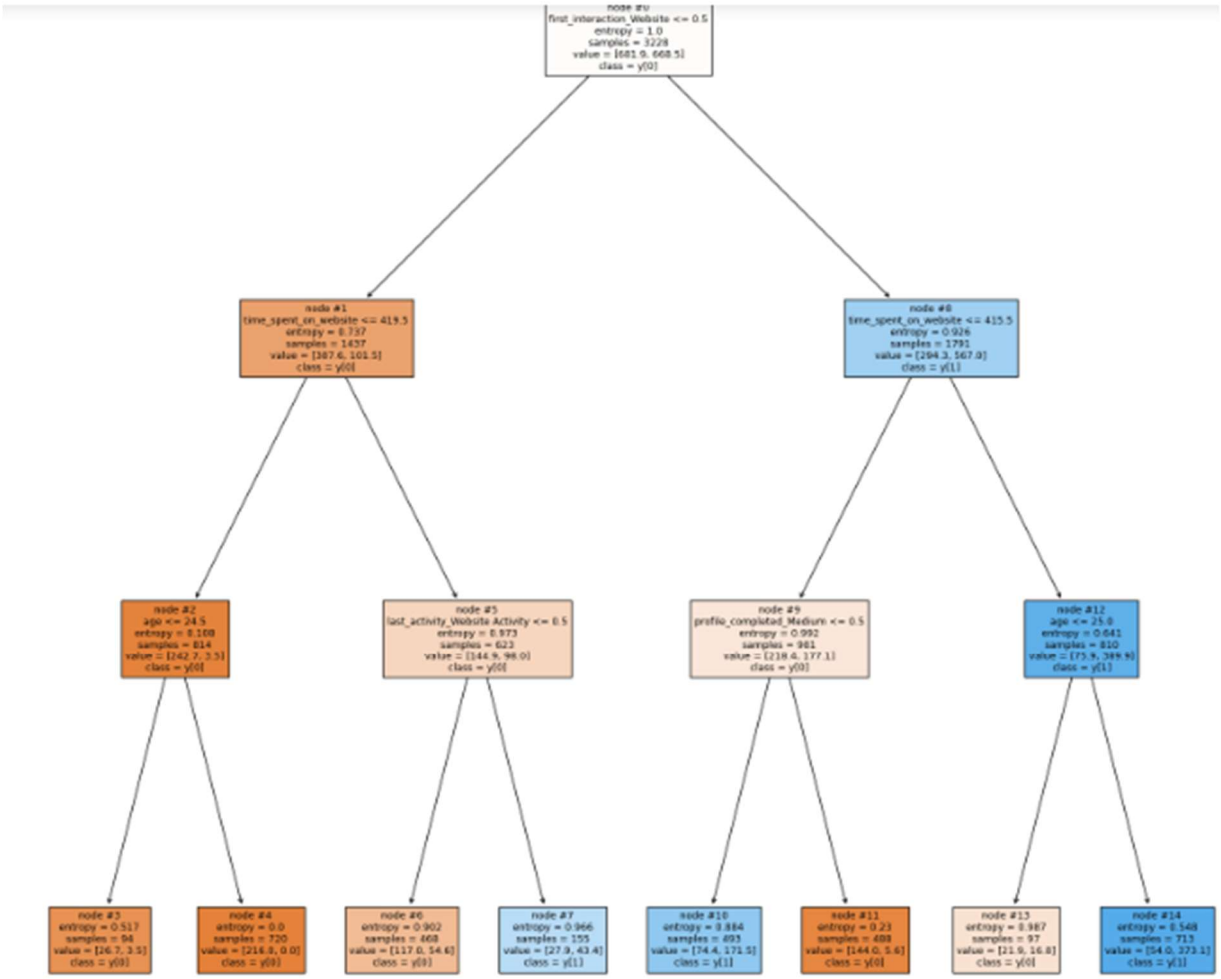
Maximizing the Recall value is our aim. The probability of reducing False Negatives increases with Recall score. the greater likelihood of forecasting a lead's conversion rate.
Before tuning

	precision	recall	f1-score	support
0	0.87	0.86	0.87	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

After tuning:

	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

We have developed tree-based models that the business may use to identify prospects that are most likely to convert and adjust their marketing tactics accordingly.

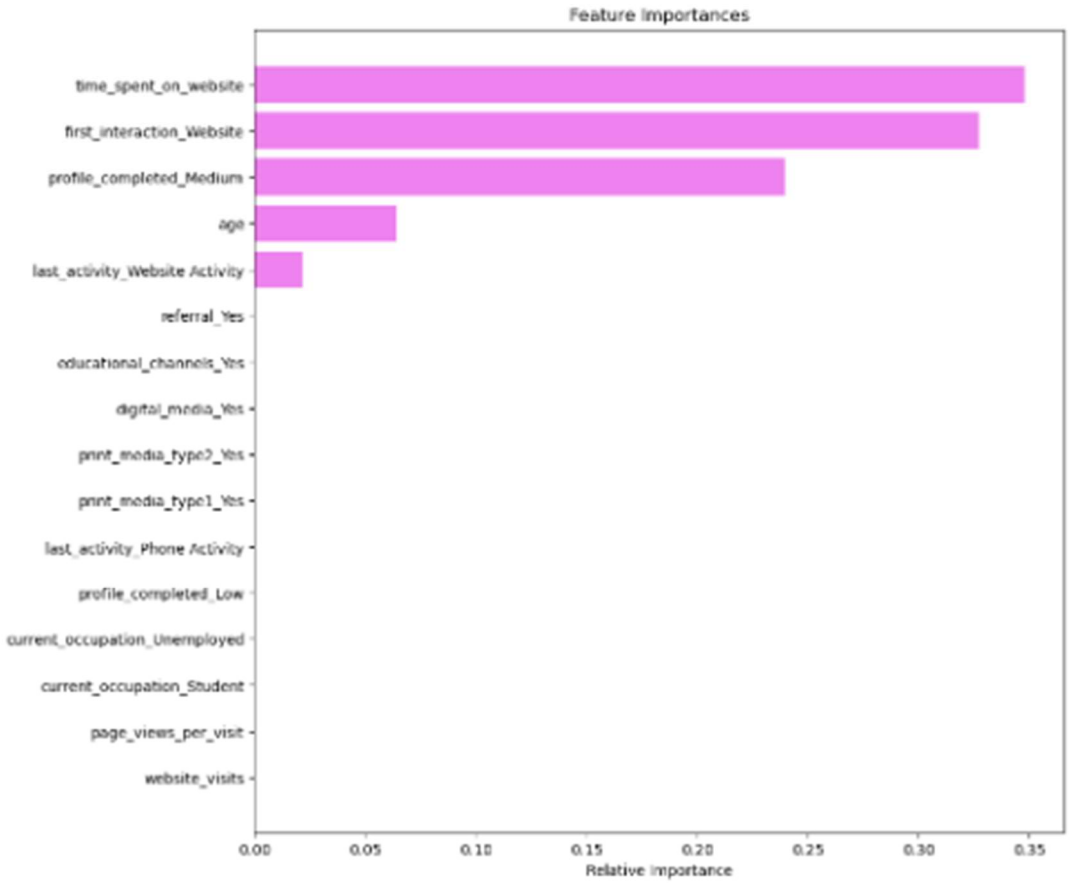


With a macro average of 84% and the highest recall score of 85% on the test data, the weaked random forest model performs well.

	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

In contrast to the decision tree, the random forest assigns a certain amount of weight to other variables such as occupation and page views per visit. This suggests that compared to the decision tree, the random forest assigns greater weight to more criteria.

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



The Decision tree model yields an 86% recall score on the test data, despite significantly overfitting the training set. With both training and testing data, the adjusted Decision Tree model produces more generalized findings and is highly balanced.

	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

The business will be able to forecast which leads will and won't be converted by using the tuned random forest model, which has an F1 score and an 85% Recall score with respectable precision. The business will have access to market to prospective customers who will eventually convert.

RECOMMENDATIONS FOR IMPLEMENTATION

One of the main factors in determining whether or not a lead will convert on the website is the amount of time they spend there. Our findings show that leads are more likely to convert the longer they spend on the website.

- In order to draw in and retain leads, the business needs to invest time and resources into establishing their website

The likelihood of conversion was higher for leads who completed the profile.

- Offering rewards to leads who complete the profile could increase the quantity of profiles that are finished.

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.