

ExtraaLearn Customer lead prediction Using ML Classification methods

Suhani Patel

3/14/2023

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 is an EdTech company that wants to capitalize on the Online Education market which is worth \$286.62bn growing at a CAGR of 10.26% from 2018 to 2023. Potential customer leads found from digital marketing resources/database, need to be converted to actual customers to increase revenue.

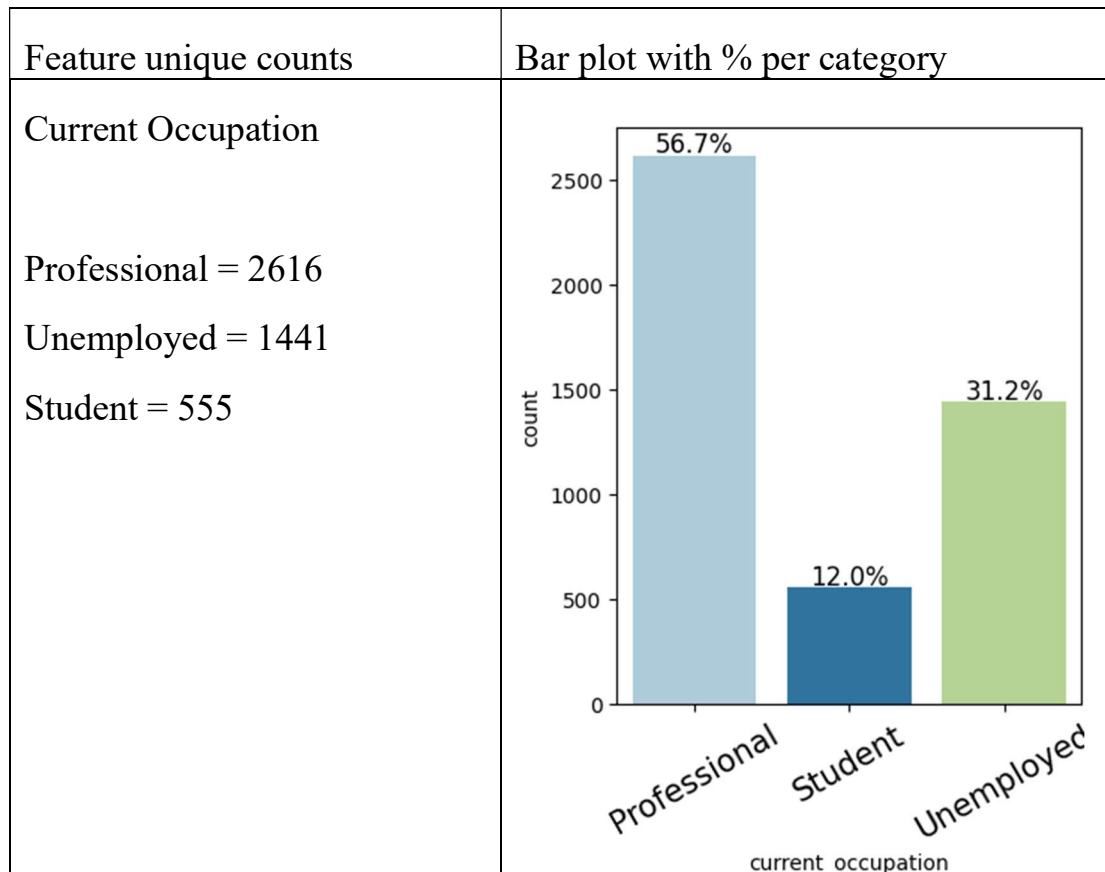
Using ML classification methods, we are able to learn the features that influence how a lead converts to a paid customer. We can then develop models, such as Logistic regression, Decision tree, Random forest. A suitable model can be chosen for deployment based on its performance characteristics, and business value. Chosen model can be deployed such that it is integrated with Sales/Customer Service/Program Management teams and systems to provide real-time understanding of when a lead can convert to a potential customer.

Data Overview

The data contains different attributes of leads and their interaction details. The data consists of 15 columns and 4612 rows.

Columns are attributes such as customer ID, age, occupation, type of first interaction with ExtraaLearn, % of profile completed, # of website visits, time spent on website, # of page views per visit, type of last interaction, type of print media ad, digital media ad, or educational channels, any past referral, current status of lead (if converted to paid customer or not).

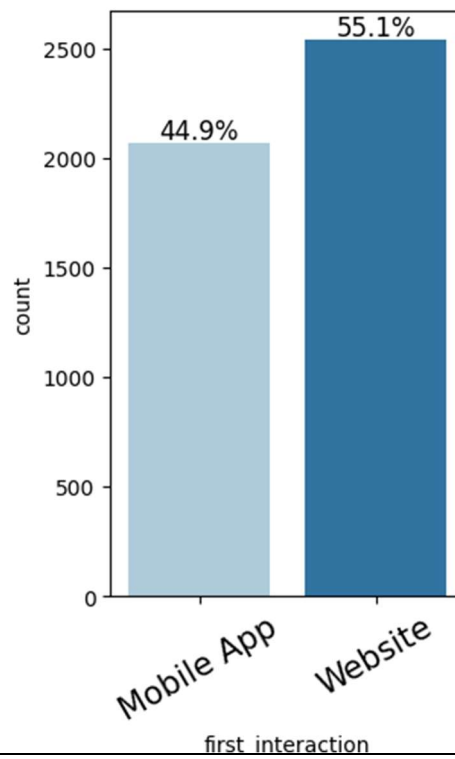
There are no null values in any column, and no duplicate rows. So, there is no need to impute missing values, or delete any duplicated rows.



First interaction with
Extraalearn

Website = 2542

Mobile app = 2070

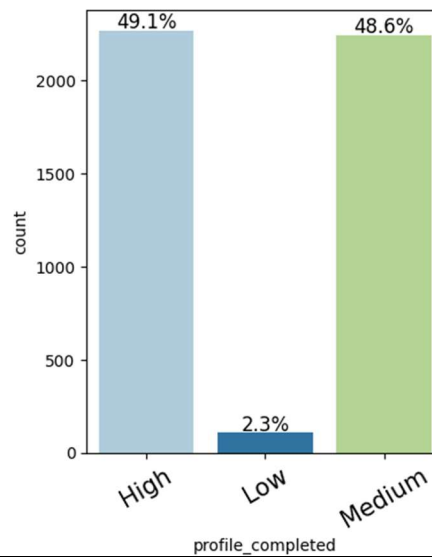


% of profile completed

High = 2264

Medium = 2241

Low = 107

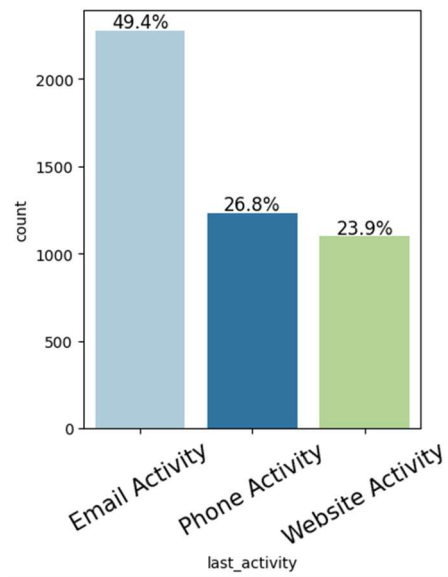


Last activity

Email = 2278

Phone = 1234

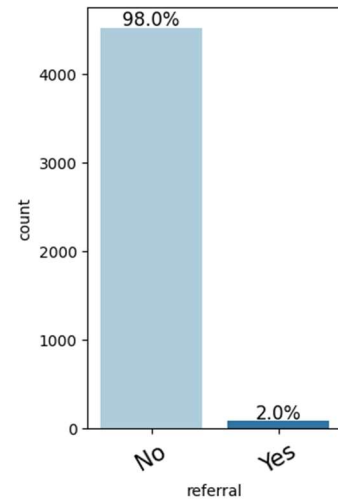
Website = 1100



Referral

Yes = 93

No = 4519



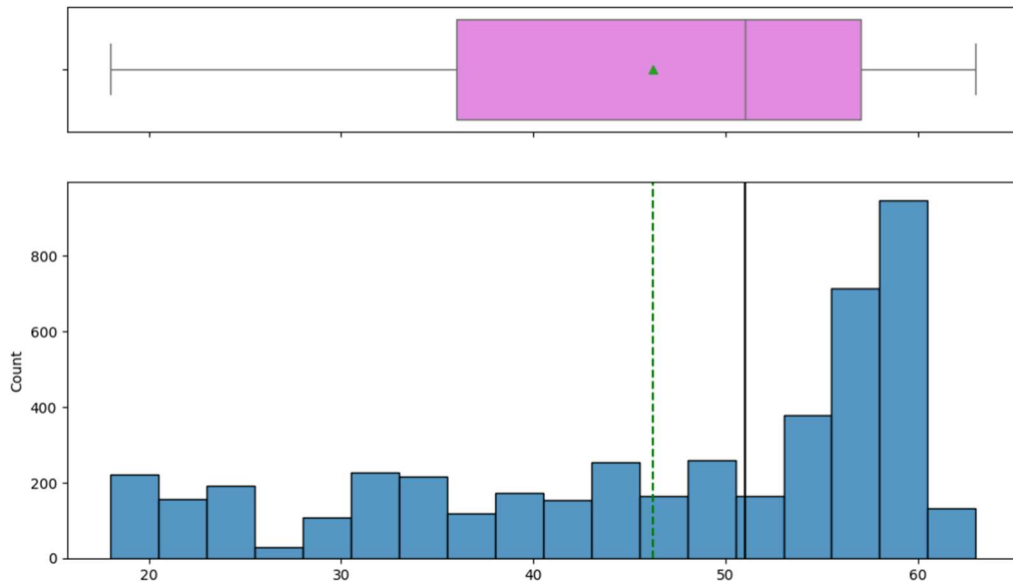
Statistical summary of the data is as below

	age	website_visits	time_spent_on_website	page_views_per_visit	status
count	4612.00000	4612.00000	4612.00000	4612.00000	4612.00000
mean	46.20121	3.56678	724.01127	3.02613	0.29857
std	13.16145	2.82913	743.82868	1.96812	0.45768
min	18.00000	0.00000	0.00000	0.00000	0.00000
25%	36.00000	2.00000	148.75000	2.07775	0.00000
50%	51.00000	3.00000	376.00000	2.79200	0.00000
75%	57.00000	5.00000	1336.75000	3.75625	1.00000
max	63.00000	30.00000	2537.00000	18.43400	1.00000

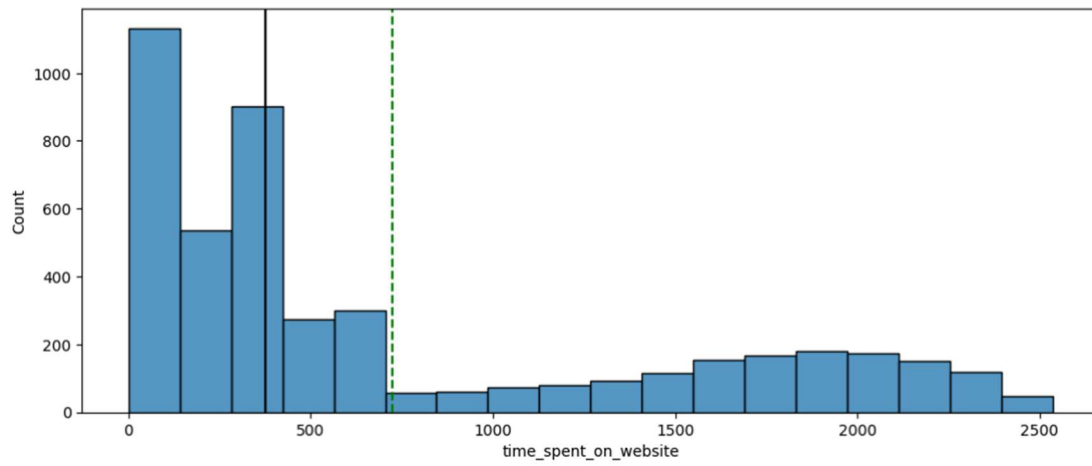
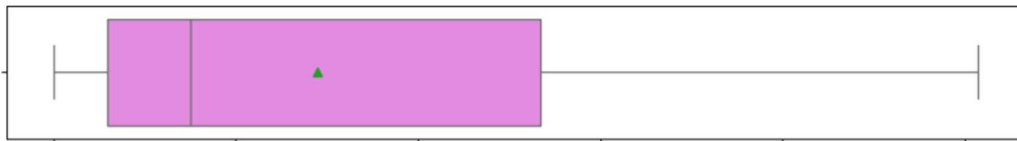
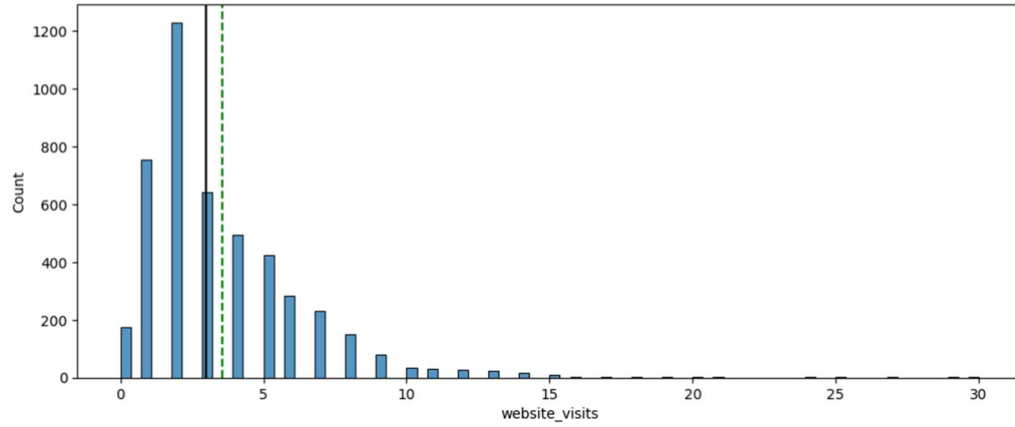
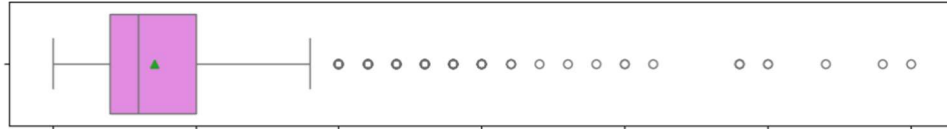
EDA Results

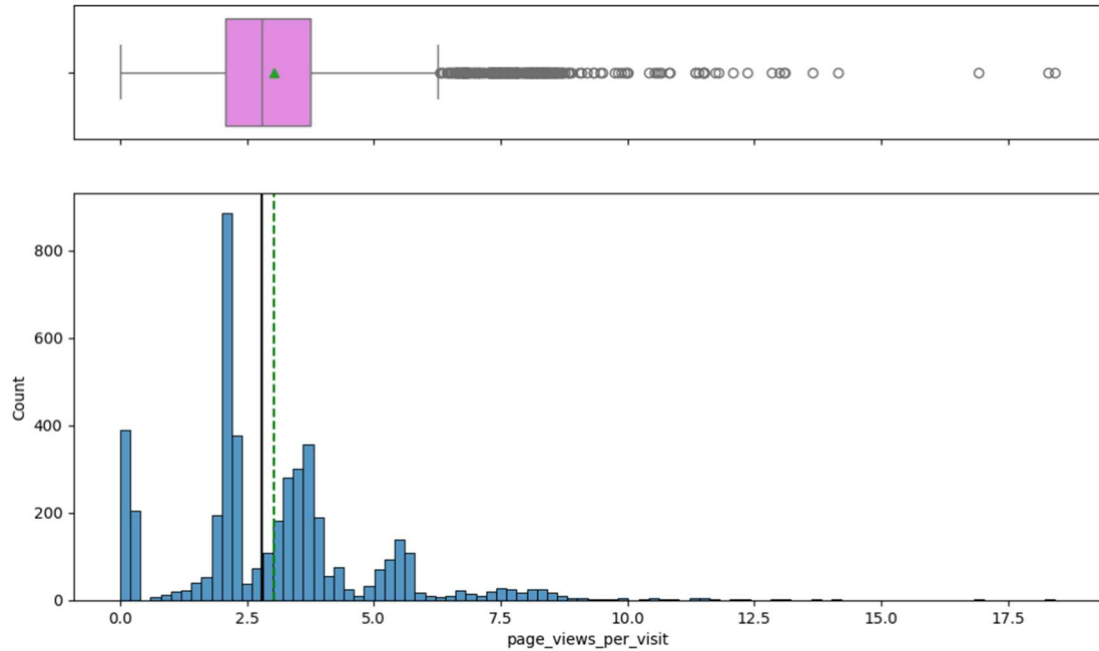
UNIVARIATE ANALYSIS

- Histogram and box plot of Age shows that the distribution is skewed to the right, and has a larger population of higher aged people. Median is 51 years.



- Similarly, distributions of other variables are as shown below



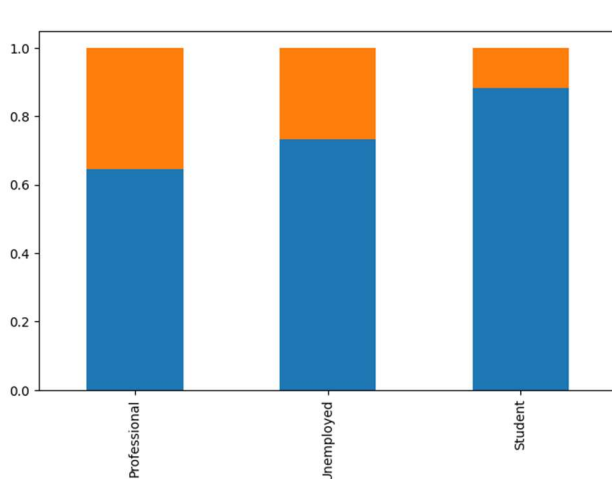


These distributions show that the data features are multimodal.

- Heat map of the correlation between variables shows that there is not any significant correlation. The highest correlated variables are time spent on website and status.



- Since this project is trying to predict the final status of the lead as a paid customer, it is important to visualize the relationship between different categorical features and the status.
 - If we compare current occupation type with status of 1 (i.e. paid customer), it is seen that professionals have a higher rate of conversion to status =1.

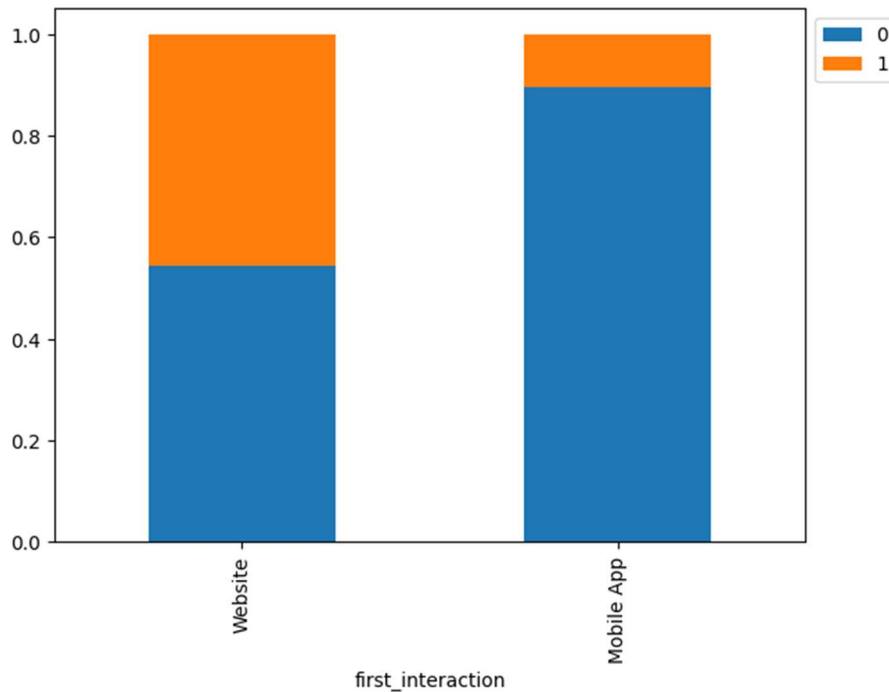


status	0	1	All
current_occupation			
All	3235	1377	4612
Professional	1687	929	2616
Unemployed	1058	383	1441
Student	490	65	555

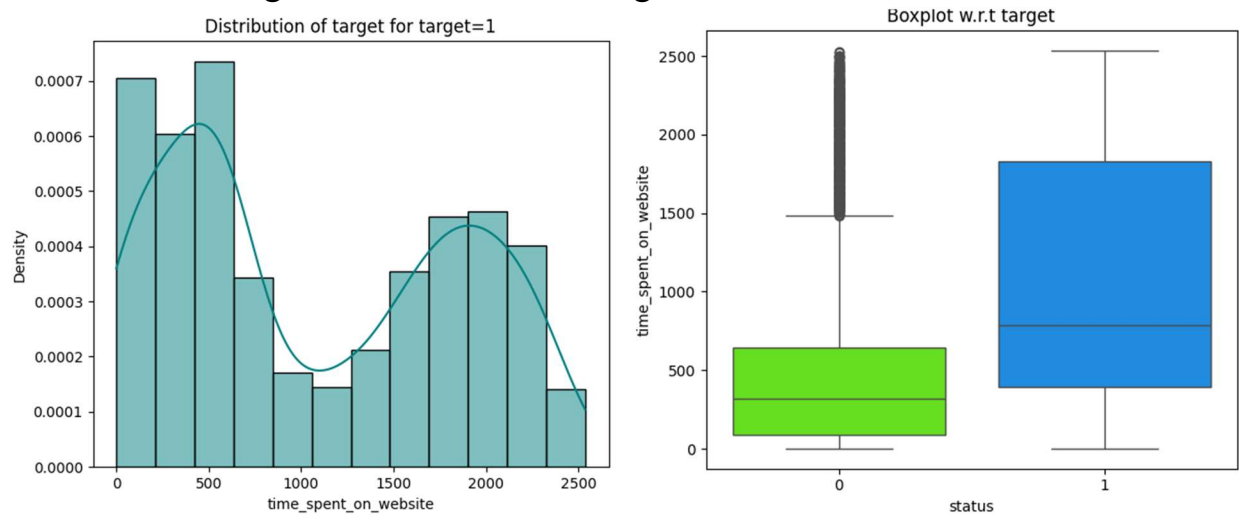
○

- Website interactions have a higher conversion rate than mobile app interactions per bar plot below.

status	0	1	All
first_interaction			
All	3235	1377	4612
Website	1383	1159	2542
Mobile App	1852	218	2070



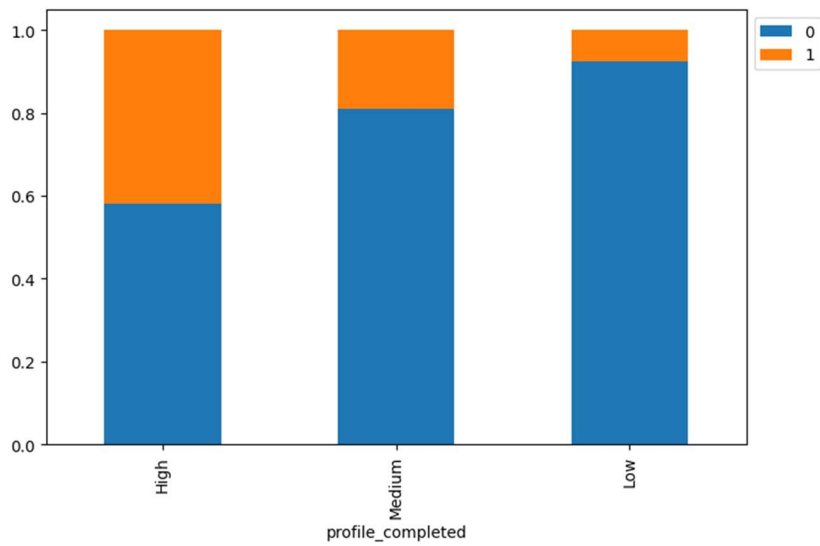
- Charts below show that customers that spent longer time on the website had a higher chance of converting to status = 1.



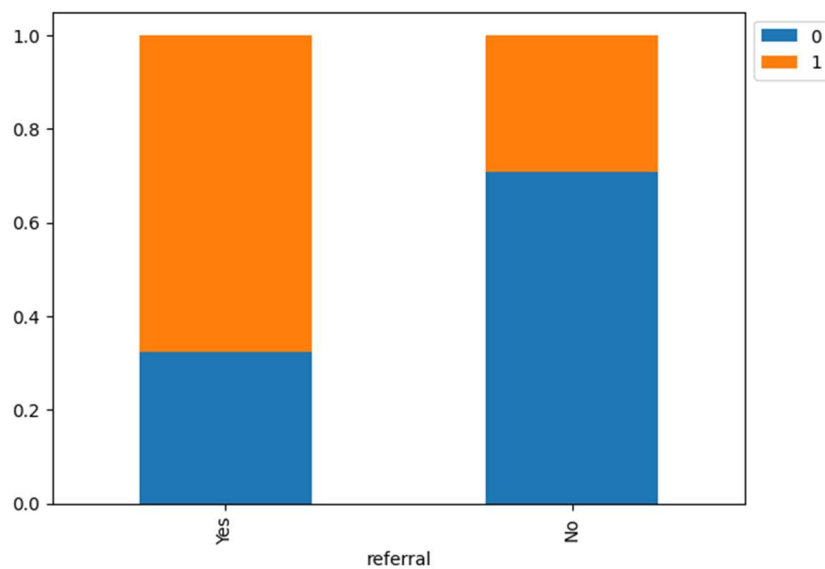
Variables such as # of website visits and page views per visit do not show much difference between status of 1 and 0. Boxplots and distribution profiles are similar.

- Website interactions and referrals have a higher conversion rate per bar plots below

status	0	1	All
profile_completed			
All	3235	1377	4612
High	1318	946	2264
Medium	1818	423	2241
Low	99	8	107



status	0	1	All
referral			
All	3235	1377	4612
No	3205	1314	4519
Yes	30	63	93



Model Building

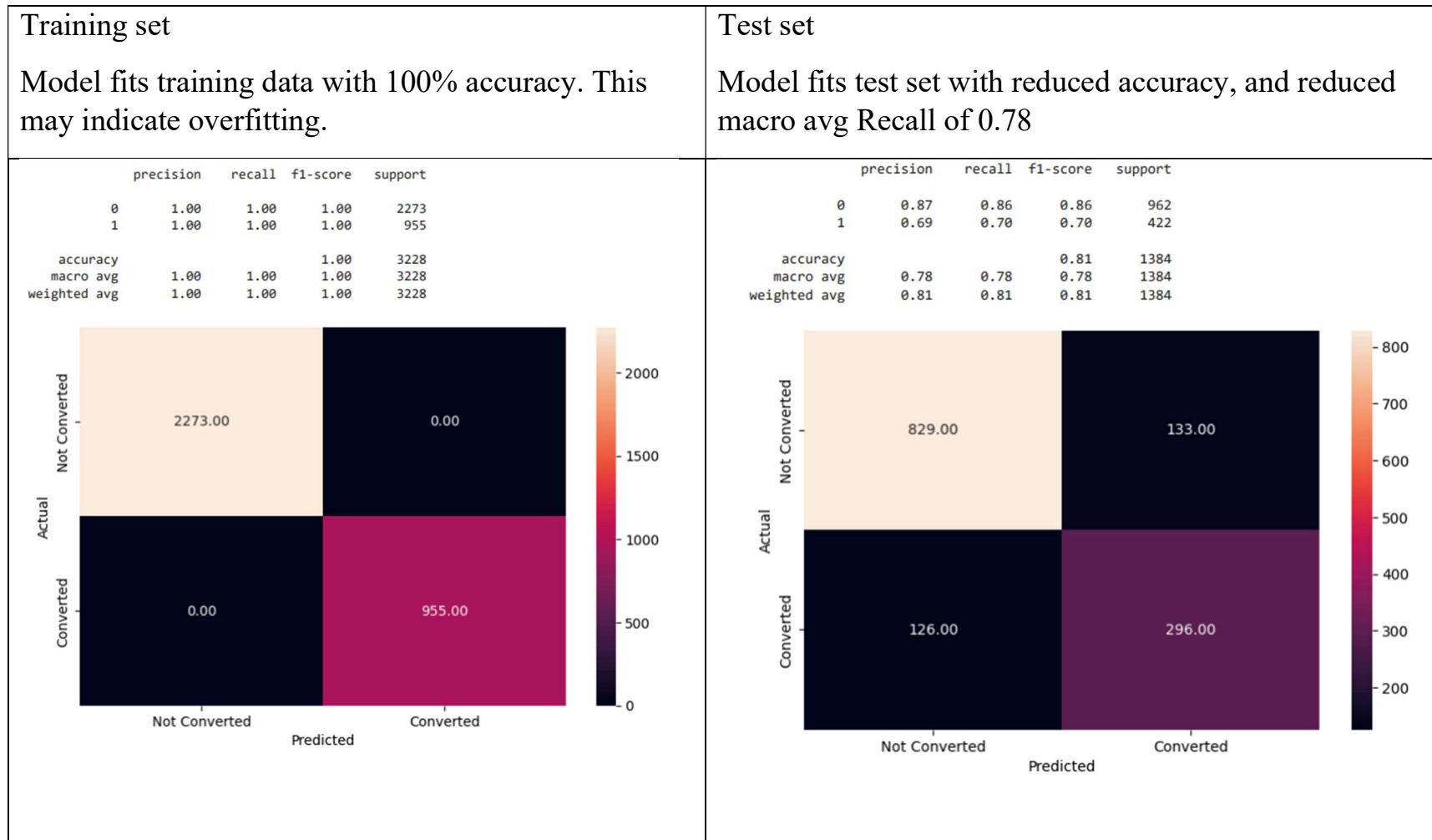
- Data was split into training and test set as a 70:30 ratio. Percentage of classes in the dataset are as shown below, and are similar for both training and test set. This will reduce any bias.

```
Shape of Training set : (3228, 16)
Shape of test set : (1384, 16)
Percentage of classes in training set:
0    0.70415
1    0.29585
Name: status, dtype: float64
Percentage of classes in test set:
0    0.69509
1    0.30491
Name: status, dtype: float64
```

- Amongst the different model performance metrics, maximizing Recall value is more beneficial. Recall value measures probability that a lead will not get converted based on model but actually the lead gets converted. This will result in losing a customer which is a greater loss for the company.

- **DECISION TREE**

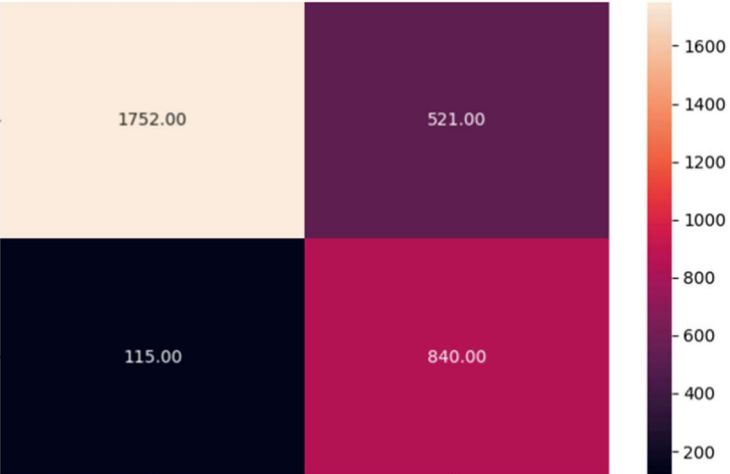

A decision tree model was developed, and the performance between Training and Test set can be seen below.



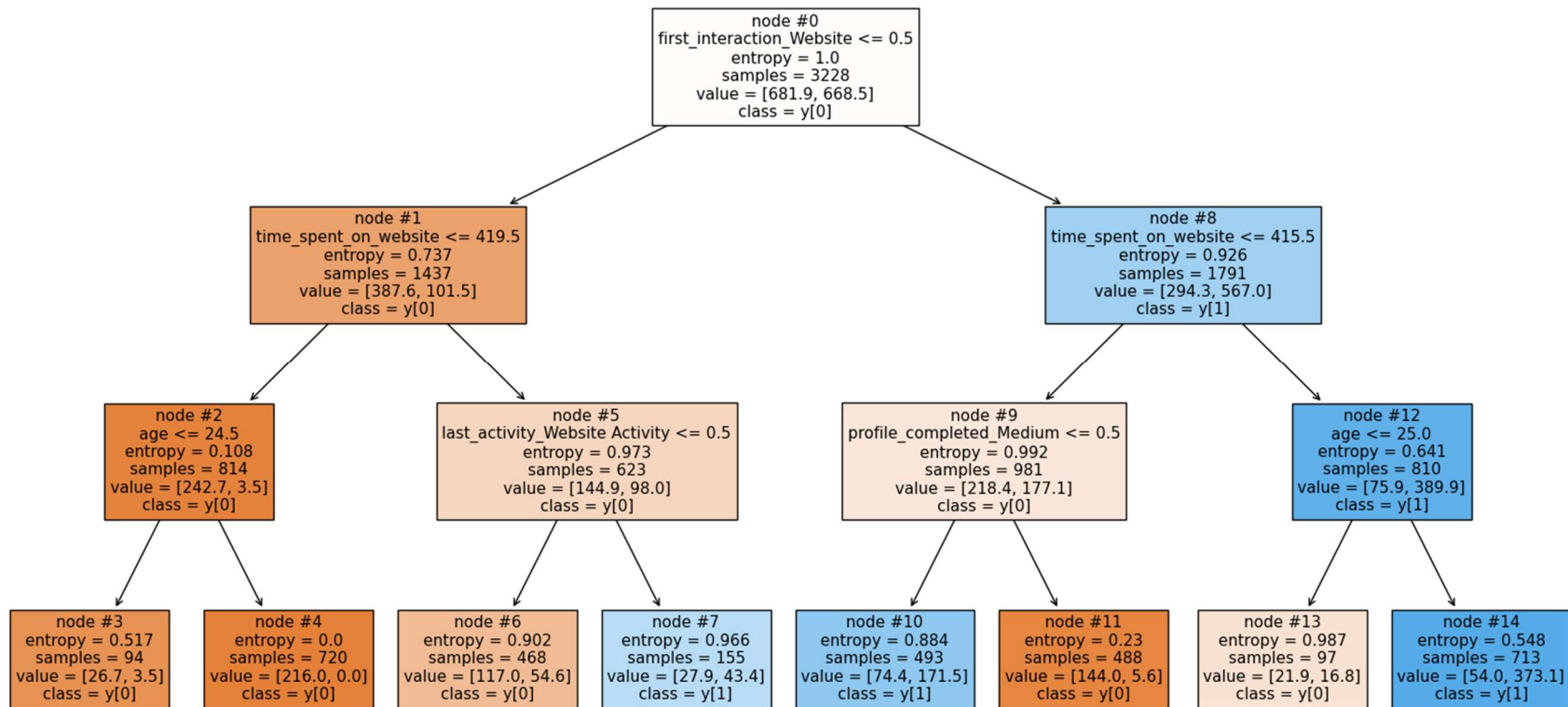
In order to reduce overfitting, model needs hyperparameter tuning. This was done with GridSearchCV. The parameters varied using GridSearchCV are below

Max depth	Criterion	Min Samples Leaf
2 to 10 branches	Gini, Entropy	5, 10, 20, 25

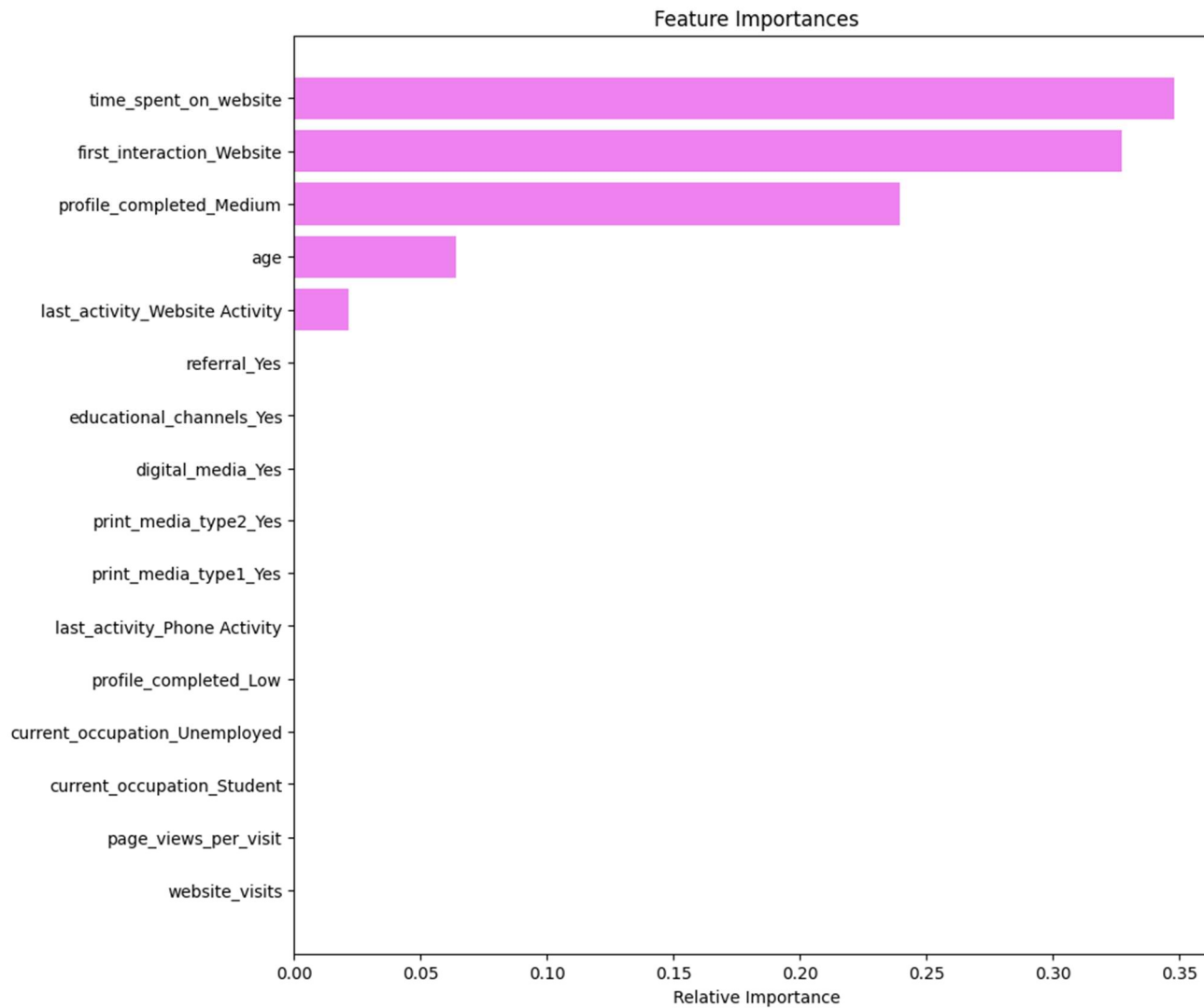
Best estimator model for recall score is decision tree with max depth of 3, criterion = entropy, and min samples leaf = 5.

Training set					Test set				
Model fits training data with recall 0.83					Model fits test set with Recall of 0.82				
	precision	recall	f1-score	support		precision	recall	f1-score	support
0	0.94	0.77	0.85	2273	0	0.93	0.77	0.84	962
1	0.62	0.88	0.73	955	1	0.62	0.86	0.72	422
accuracy			0.80	3228	accuracy			0.80	1384
macro avg	0.78	0.83	0.79	3228	macro avg	0.77	0.82	0.78	1384
weighted avg	0.84	0.80	0.81	3228	weighted avg	0.83	0.80	0.80	1384
Actual					Actual				
	Not Converted	1752.00	521.00	Converted					
Converted	115.00	840.00	Converted		Not Converted	738.00	224.00	Converted	
	Predicted					Predicted			

The tuned decision tree can be visualized as below. Following the nodes that are blue will show us the path where lead will convert to a paid customer at a higher rate.

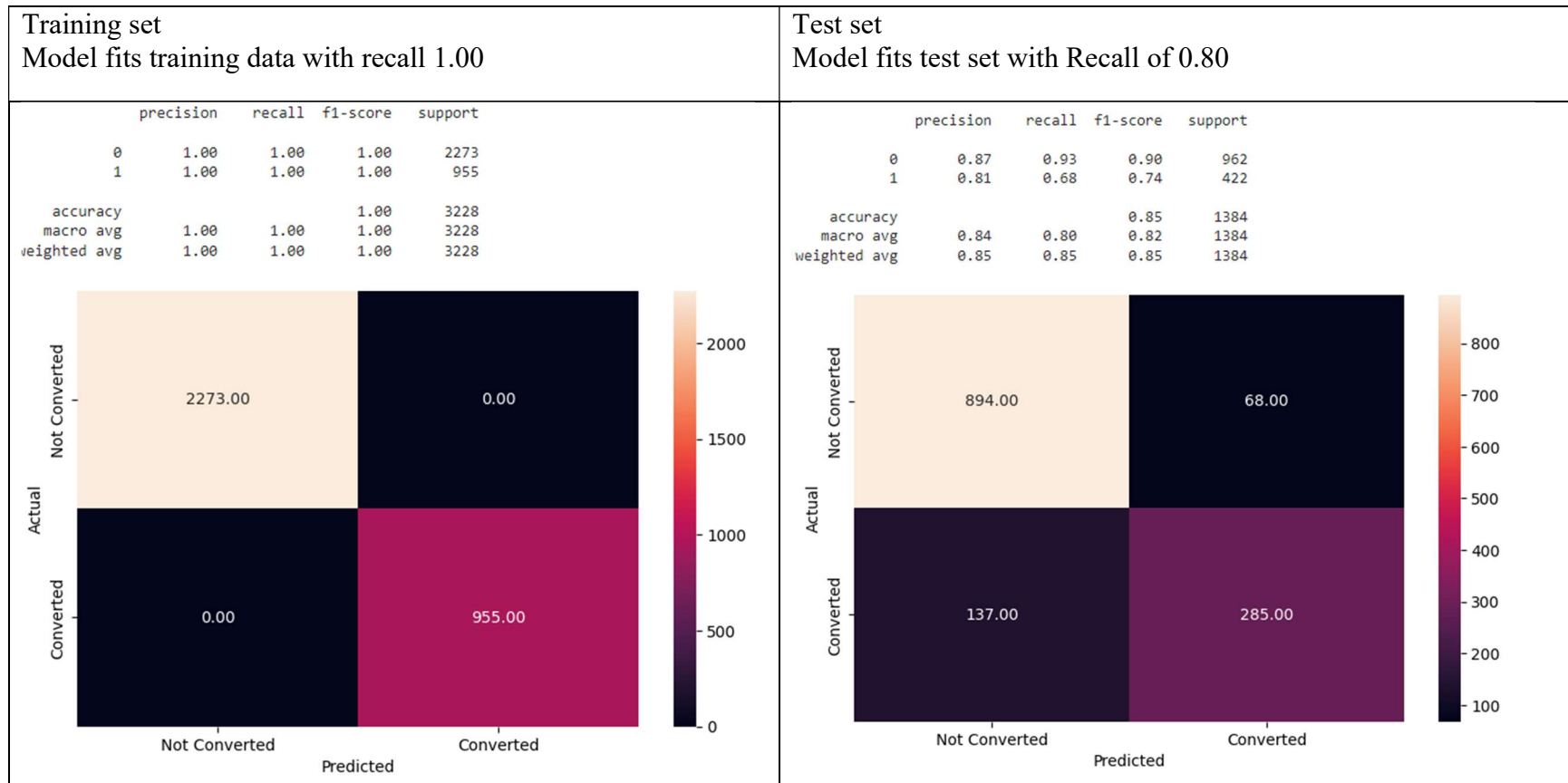


Feature importances can also be visualized and based on image below top 5 features are most relevant.



- **RANDOM FOREST**

A Random Forest model was developed, and the performance between Training and Test set can be seen below.



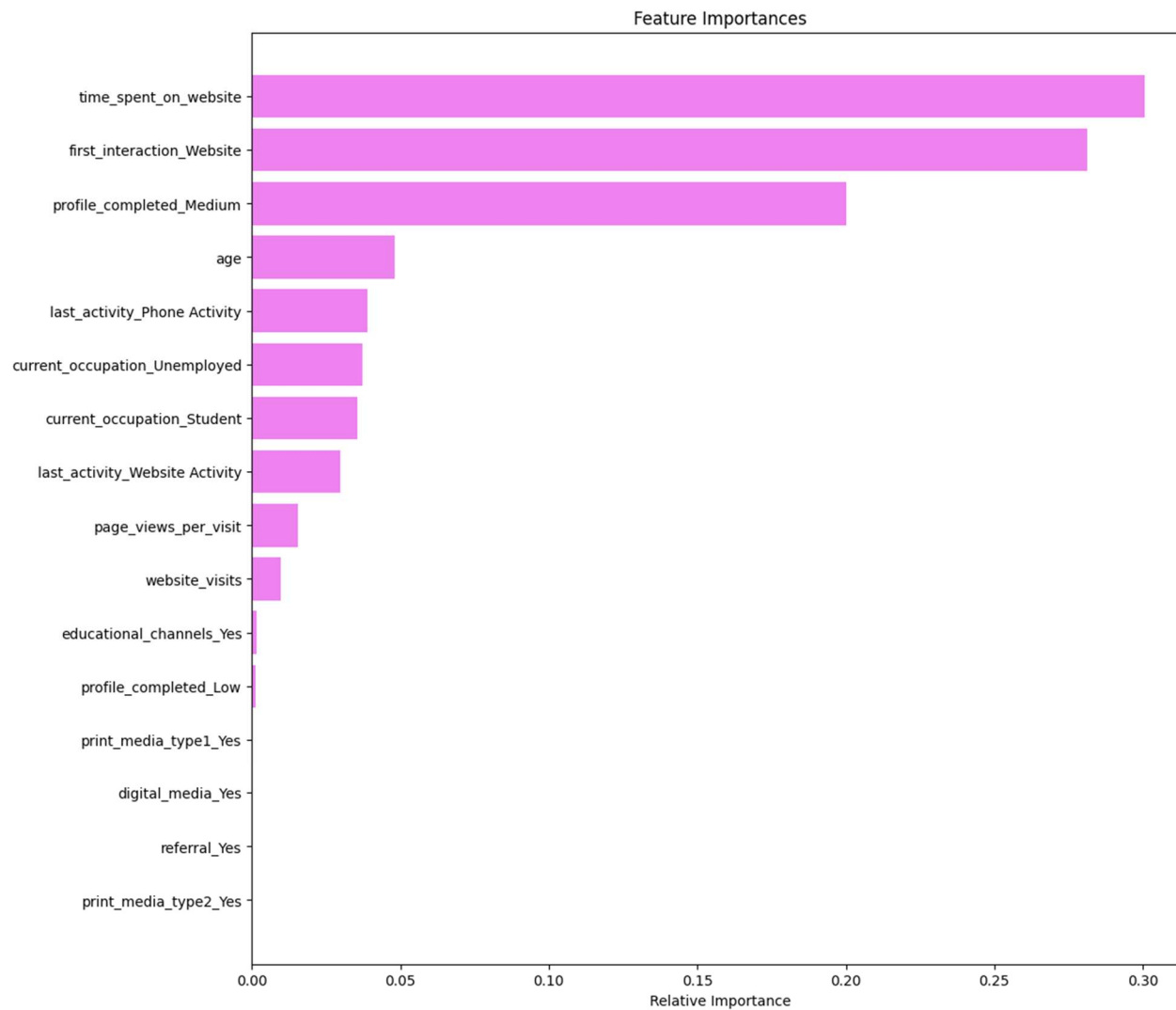
In order to reduce overfitting, Random Forest model needs hyperparameter tuning. This was done with GridSearchCV. The parameters varied using GridSearchCV are below

N_estimators	Max depth	Max features	Min Samples Leaf	Max samples	Class weight
110, 120	6, 7	0.8, 0.9	20, 25	0.9, 1	Balanced, 0.3:0.7

Best estimator model is with max depth = 6, max features = 0.8, max samples = 0.9, min samples leaf = 25, n_estimators = 120, class weight = balanced, criterion = entropy

Training set					Test set				
Model fits training data with recall 0.85					Model fits test set with recall of 0.84				
	precision	recall	f1-score	support		precision	recall	f1-score	support
0	0.94	0.83	0.88	2273	0	0.93	0.83	0.87	962
1	0.68	0.87	0.76	955	1	0.68	0.85	0.76	422
accuracy			0.84	3228	accuracy			0.83	1384
macro avg	0.81	0.85	0.82	3228	macro avg	0.81	0.84	0.82	1384
weighted avg	0.86	0.84	0.84	3228	weighted avg	0.85	0.83	0.84	1384
Actual	Not Converted	1876.00			Actual	Not Converted	795.00		
	Converted	397.00				Converted	167.00		
Predicted	Not Converted	121.00			Predicted	Not Converted	62.00		
	Converted	834.00				Converted	360.00		

Feature importance chart for Random forest model. This chart is different than the decision tree feature chart.



Model Performance Summary

*Logistic regression was also performed and has been included in the Appendix section.

- Table below shows performance comparison of all models on the test set

Model	Recall	Precision	F1 score
Logistic regression	0.75	0.77	0.76
Decision tree	0.78	0.78	0.78
Tuned decision tree	0.82	0.77	0.78
Random Forest	0.80	0.84	0.82
Tuned Random Forest	0.84	0.81	0.82

- The tuned Random Forest model is performing better than the other models in terms of Recall, Precision and F1 score. This model should be utilized to predict if the lead will convert to a paid customer for Extraalearn.
- Factors that affect lead conversion process are
 - Greater time spent on website
 - Profile completed (medium % or higher)
 - Phone activity with lead
- Profile of the leads who are likely to convert are
 - Currently unemployed/student
 - Age > 50 years
 - Referred by another

Conclusion

- In order to increase business revenue for Extraalearn, we should deploy a Tuned Random Forest Model (recall = 0.84) to predict if lead will convert to paid customer.
- If the data is collected in a timely manner, then the model will be able to show sales and program management teams in real-time if the leads are

worth pursuing. Those customers can be provided targeted messaging via emails/phone calls from the sales team.

- By tracking number of successfully converted leads over time, we can understand the \$ revenue generated based on the Random forest model prediction.
- If model needs to be updated based on new data, then it can be revisited in a few months.

Appendix

LOGISTIC REGRESSION

Training set

	precision	recall	f1-score	support
0	0.86	0.90	0.88	2273
1	0.72	0.65	0.68	955
accuracy			0.82	3228
macro avg	0.79	0.77	0.78	3228
weighted avg	0.82	0.82	0.82	3228



Test set

	precision	recall	f1-score	support
0	0.84	0.89	0.86	962
1	0.71	0.61	0.65	422
accuracy			0.80	1384
macro avg	0.77	0.75	0.76	1384
weighted avg	0.80	0.80	0.80	1384



Coefficient table

	θ
first_interaction_Website	2.71973
referral_Yes	0.48239
print_media_type2_Yes	0.21712
last_activity_Website Activity	0.18345
print_media_type1_Yes	0.15063
digital_media_Yes	0.04097
time_spent_on_website	0.00124
educational_channels_Yes	0.00071
age	-0.00883
website_visits	-0.01105
page_views_per_visit	-0.05021
current_occupation_Unemployed	-0.63218
profile_completed_Low	-0.68738
last_activity_Phone Activity	-0.92127
profile_completed_Medium	-1.57449
current_occupation_Student	-2.31646