Optimizing Lead Conversion in EdTech

Classification and Hypothesis Testing

Date: Jul 28, 2025

Taha Asim

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

Problem

How can we help ExtraaLearn identify which leads are more likely to convert to paying customers, so that marketing and sales efforts can be prioritized more effectively?

Objective

To build a data-driven model that can help find the likelihood of lead conversion and catering to all leads to could waste time and resources.

The path that needs to be taken is to recognize high conversion leads early.

Data Overview

#	Column	Non-Null Count	Dtype
0	ID	4612 non-null	object
1	age	4612 non-null	int64
2	current_occupation	4612 non-null	object
3	first_interaction	4612 non-null	object
4	profile_completed	4612 non-null	object
5	website_visits	4612 non-null	int64
6	time_spent_on_website	4612 non-null	int64
7	page_views_per_visit	4612 non-null	float64
8	last_activity	4612 non-null	object
9	print_media_type1	4612 non-null	object
10	print_media_type2	4612 non-null	object
11	digital_media	4612 non-null	object
12	educational_channels	4612 non-null	object
13	referral	4612 non-null	object
14	status	4612 non-null	int64

Data Overview

- 15 different data types
- No duplicate data found
- 4612 unique records
- No null values

Feature Categories

- **Demographics:** age, occupation
- **Engagement Metrics:** website visit, time spent, page views
- **Marketing Channels:** digital, print, referral
- **User Activity:** email, phone, website interaction
- **Target Variable:** status

EDA Results

- Provide comments on the visualization such as range of attributes, outliers of various attributes.
- Provide comments on the distribution of the variables
- Use appropriate visualizations to identify the patterns and insights
- Key meaningful observations on individual variables and the relationship between variables

***Note**: You can use more than one slide if needed*

EDA - Data Statistics

	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

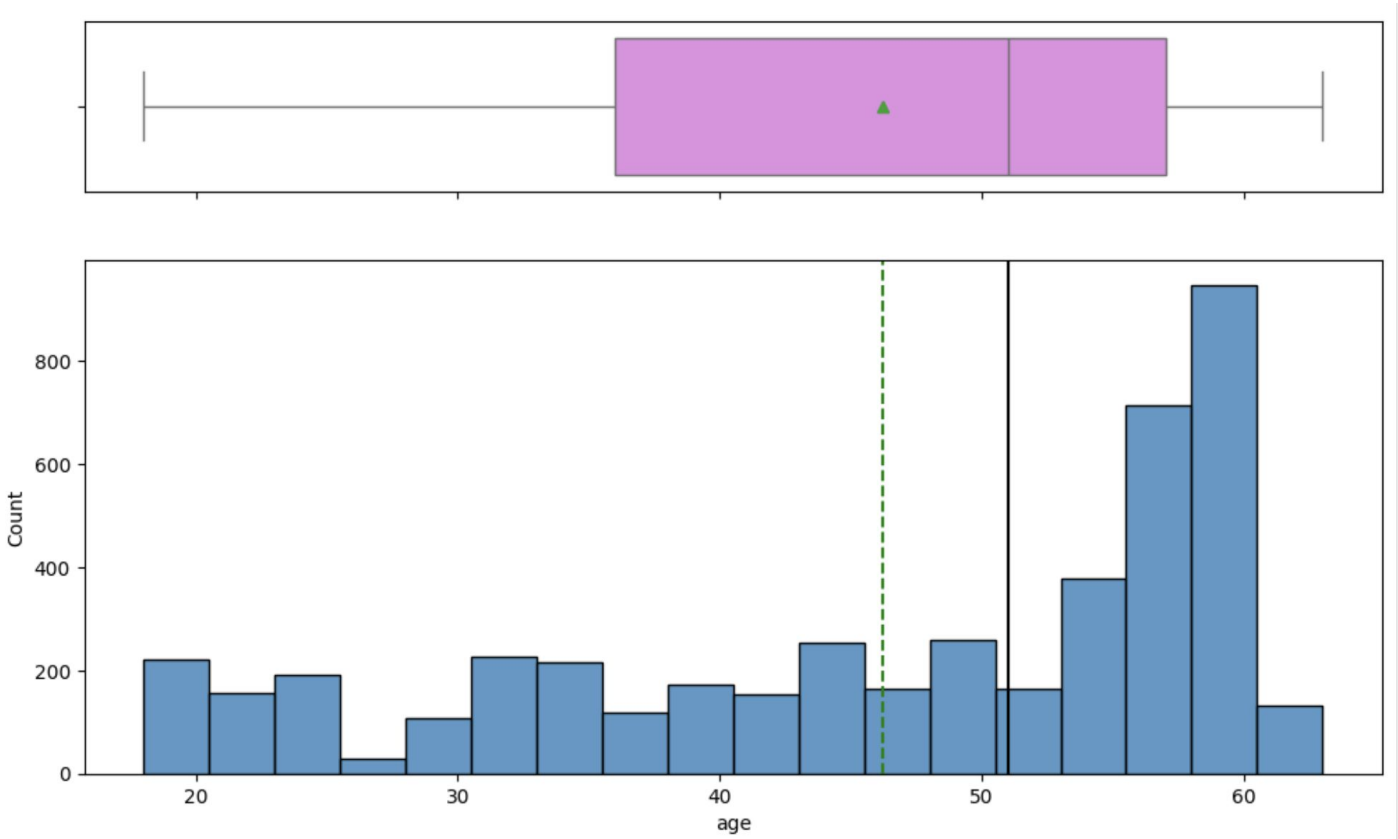
EDA - Unique Values Count

- This table shows the distribution for key categorical features

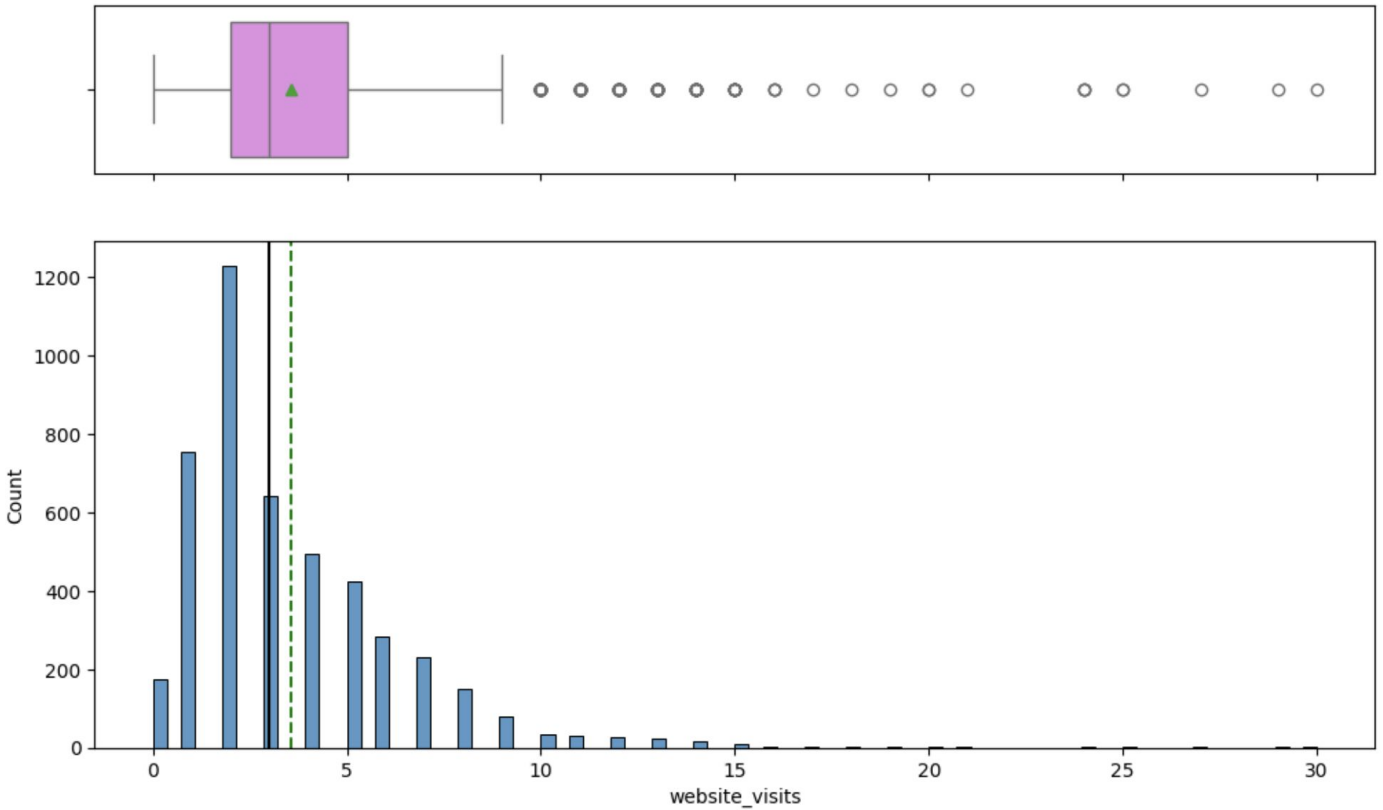
```
Name: count, Length: 4612, d
------------------------------
current_occupation
Professional    2616
Unemployed      1441
Student          555
Name: count, dtype: int64
------------------------------
first_interaction
Website       2542
Mobile App    2070
Name: count, dtype: int64
------------------------------
profile_completed
High      2264
Medium    2241
Low        107
Name: count, dtype: int64
------------------------------
last_activity
Email Activity      2278
Phone Activity      1234
Website Activity    1100
Name: count, dtype: int64
------------------------------
print_media_type1
No     4115
Yes     497
Name: count, dtype: int64
```

```
------------------------------
digital_media
No     4085
Yes     527
Name: count, dtype: int64
------------------------------
educational_channels
No     3907
Yes     705
Name: count, dtype: int64
------------------------------
referral
No     4519
Yes      93
Name: count, dtype: int64
------------------------------
```

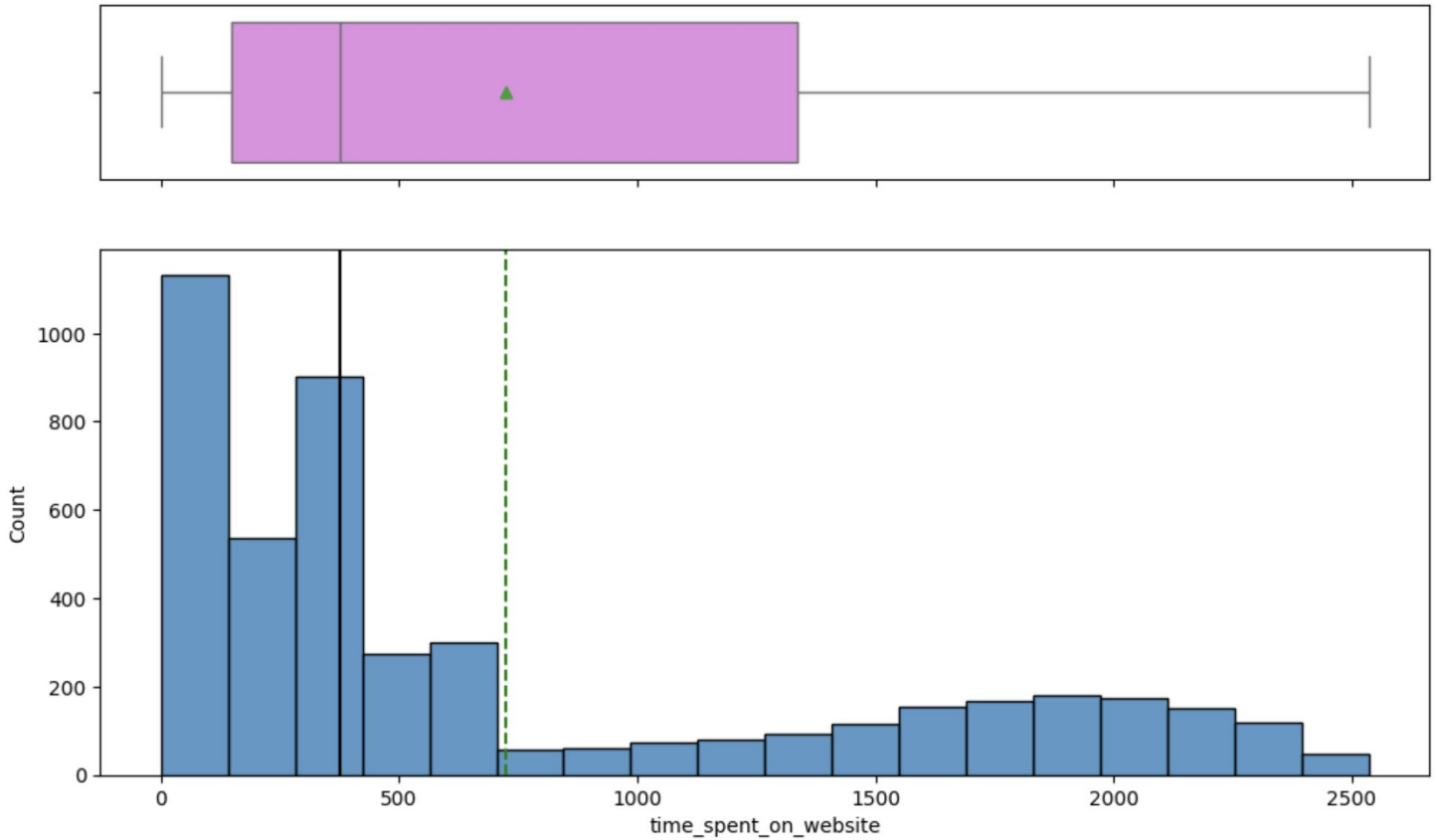
EDA - Boxplot for Age



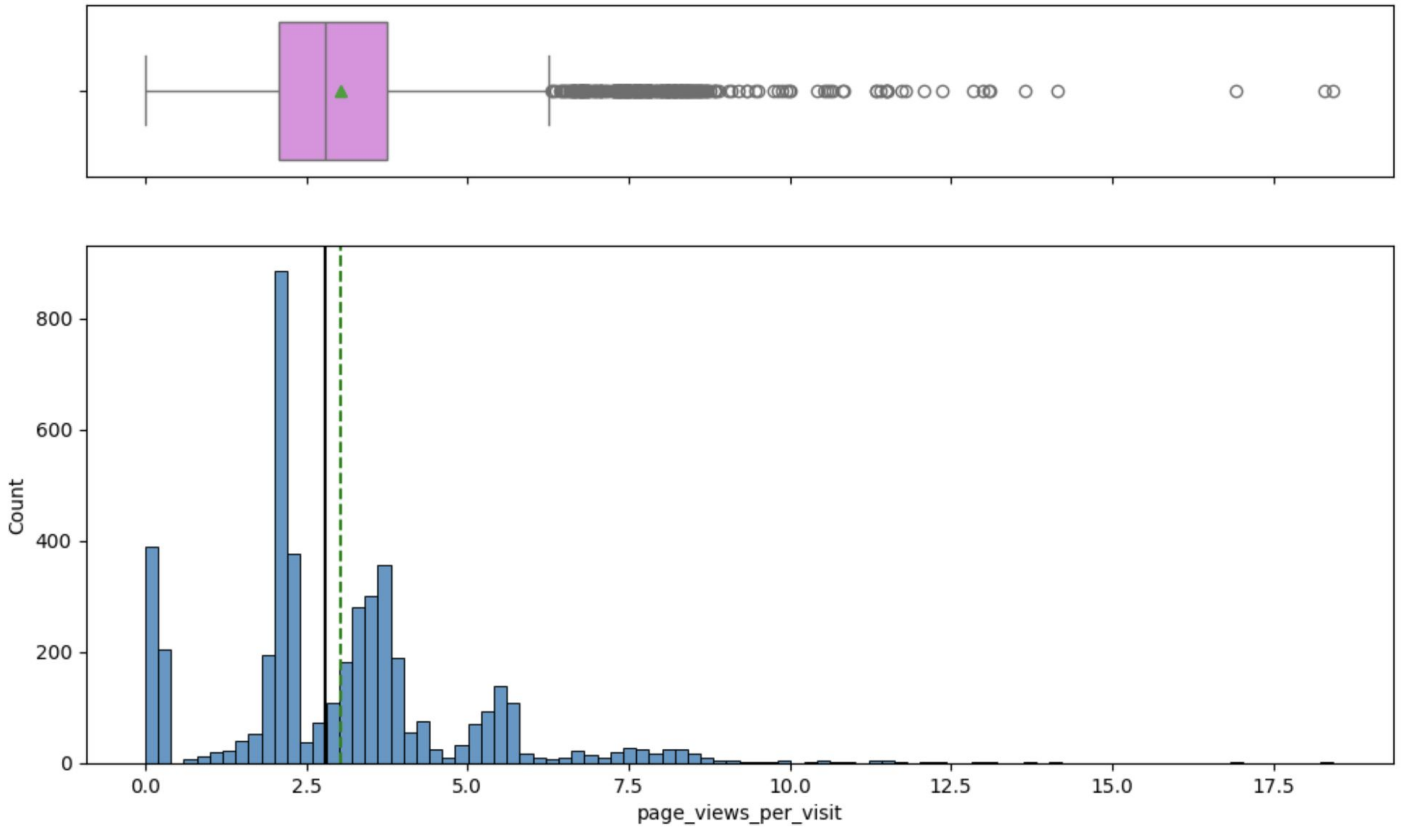
EDA - Boxplot for website visits



EDA - Boxplot for time spent on the website



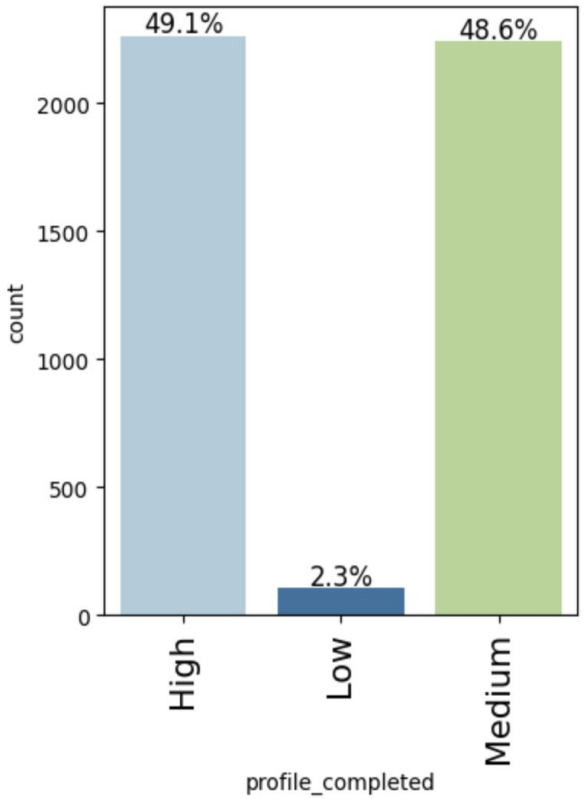
EDA - Boxplot for number of pages on website viewed

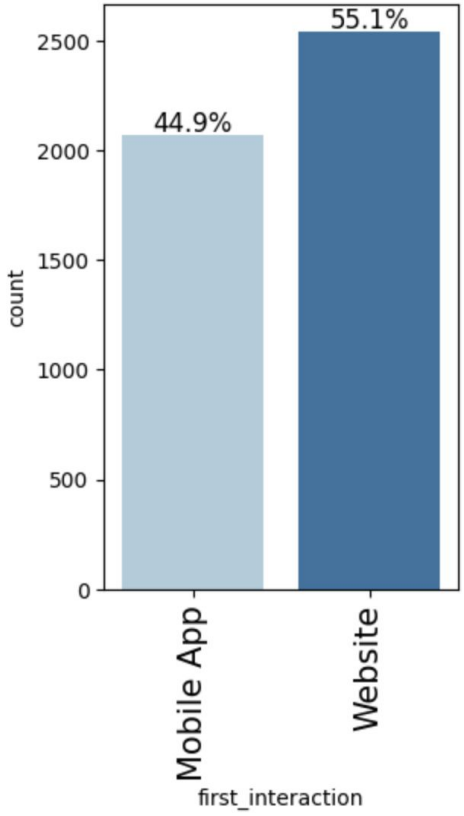


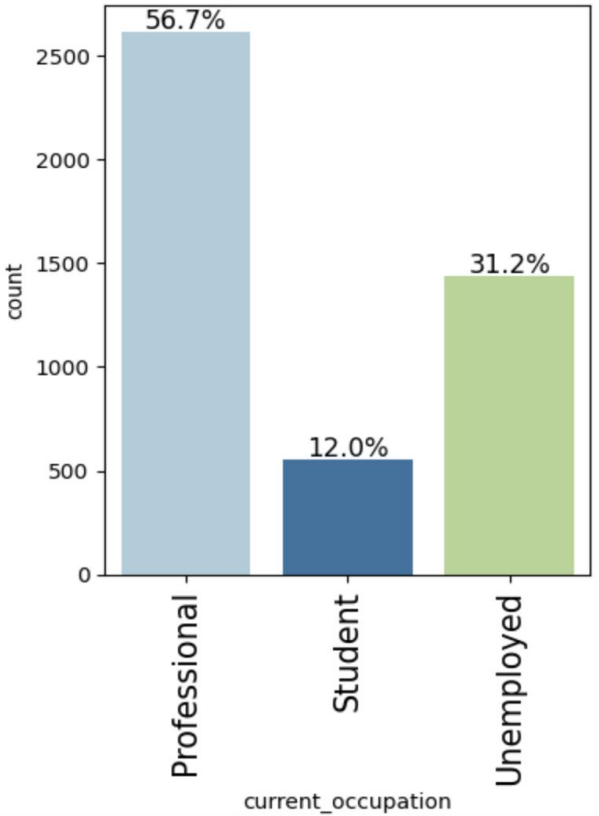
EDA - Insights gained so far

- 174 leads have NOT visited the website
- Median age is 50 but older users appear more engaged.
- Most leads visit 2 to 4 times which indicates that it may be of genuine interest.
- Skewed data and most people spent about 10 mins on the website. This could be used to identify the lead quality.
- Multiple outliers found and high numbers of views per visit may indicate deep interest in offerings.

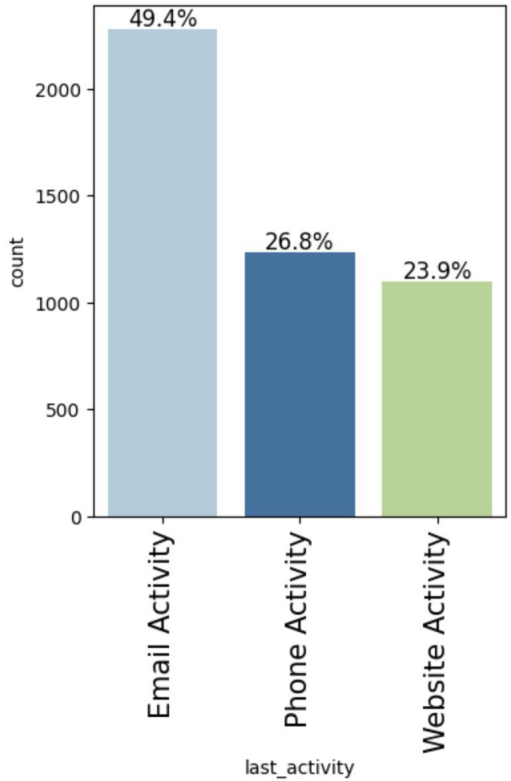
EDA - barplots on profile filled on phone, first interaction and occupation status

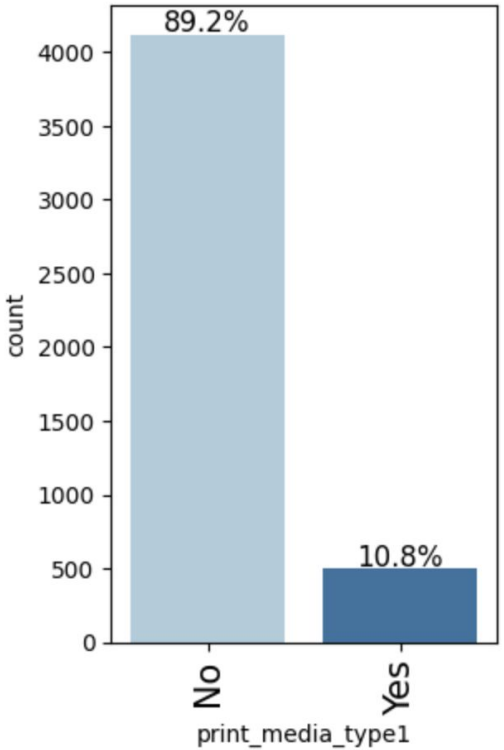


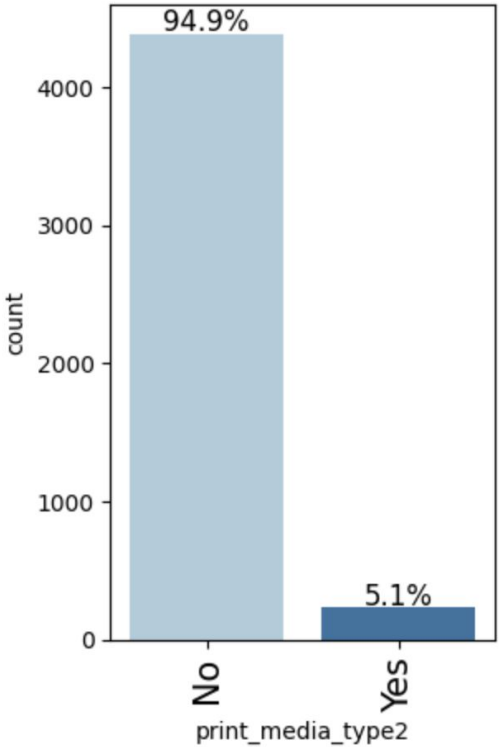




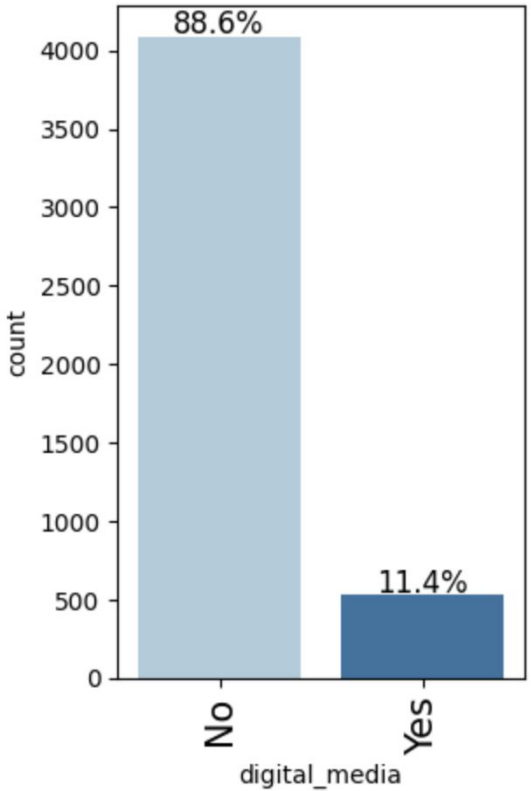
EDA - barplot for last activity, print media 1 and 2

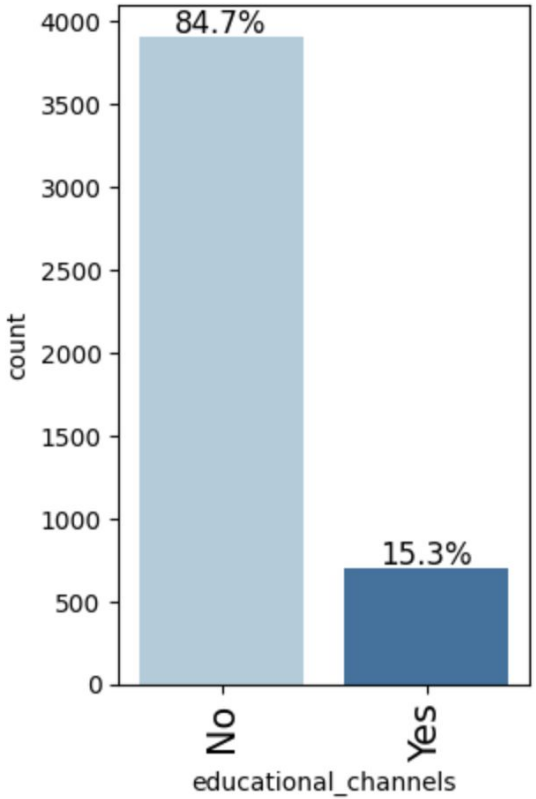


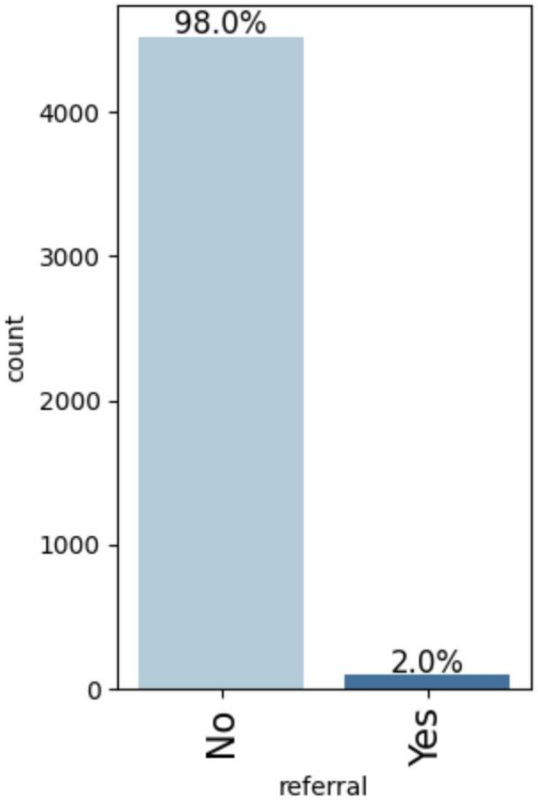




EDA - digital media, educational channels awareness and referral

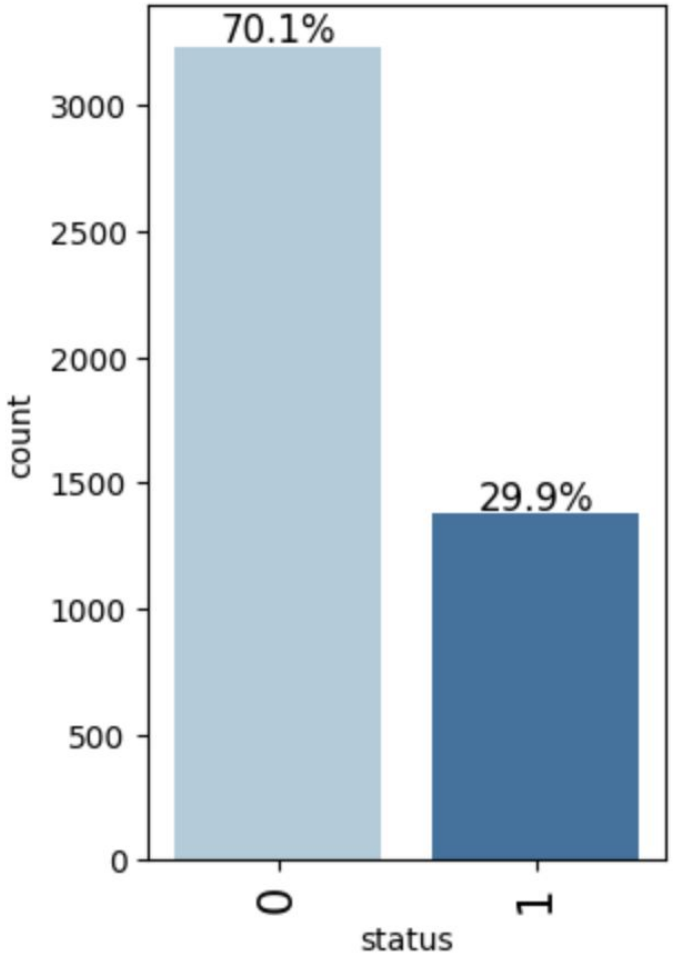






EDA - lead conversion status

1 = Paid Customers
2 = Unpaid Customers



EDA - Insights gained so far

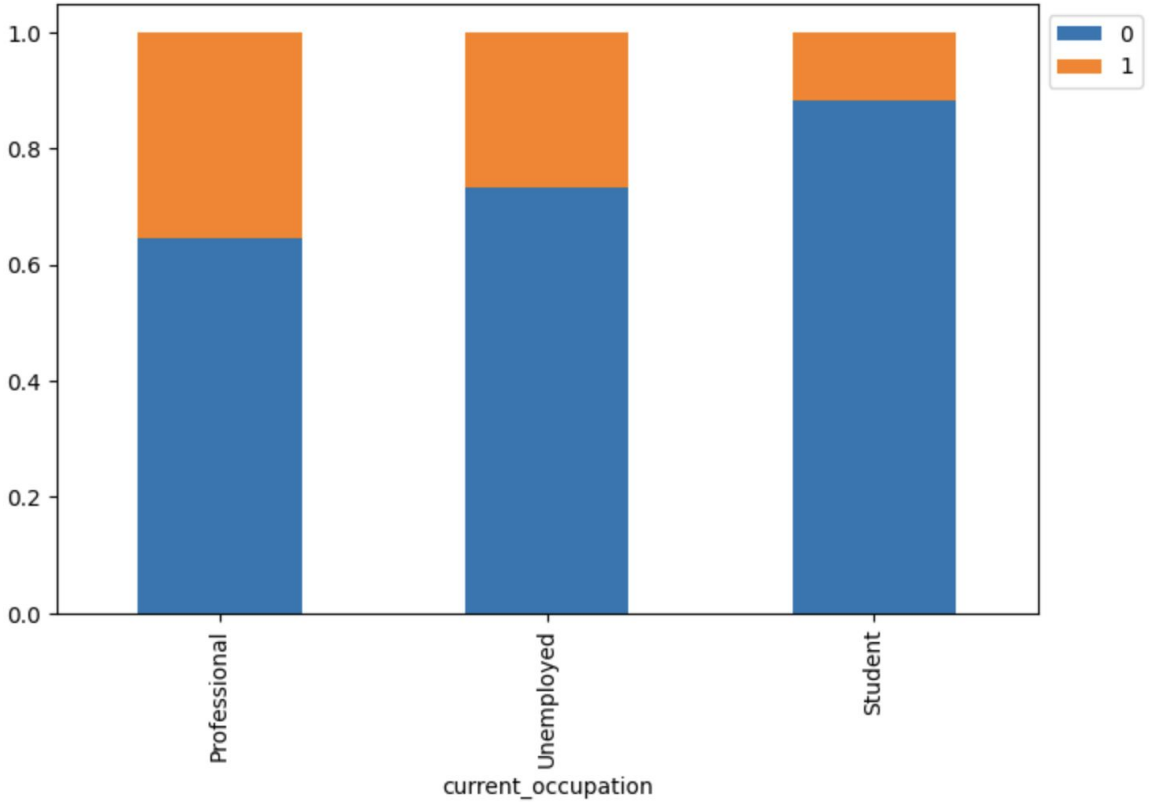
- About 98% of the leads have medium or high profile completion showing strong initial interest.
- Website is the most common first touchpoint at about 55% suggesting it is a key conversion asset.
- Majority of leads are professionals at about 56.7% implying that professionals are a major target segment.
- Compared to print media, digital platforms so we need to put more focus on email and digital platforms.

EDA - Correlation heat map

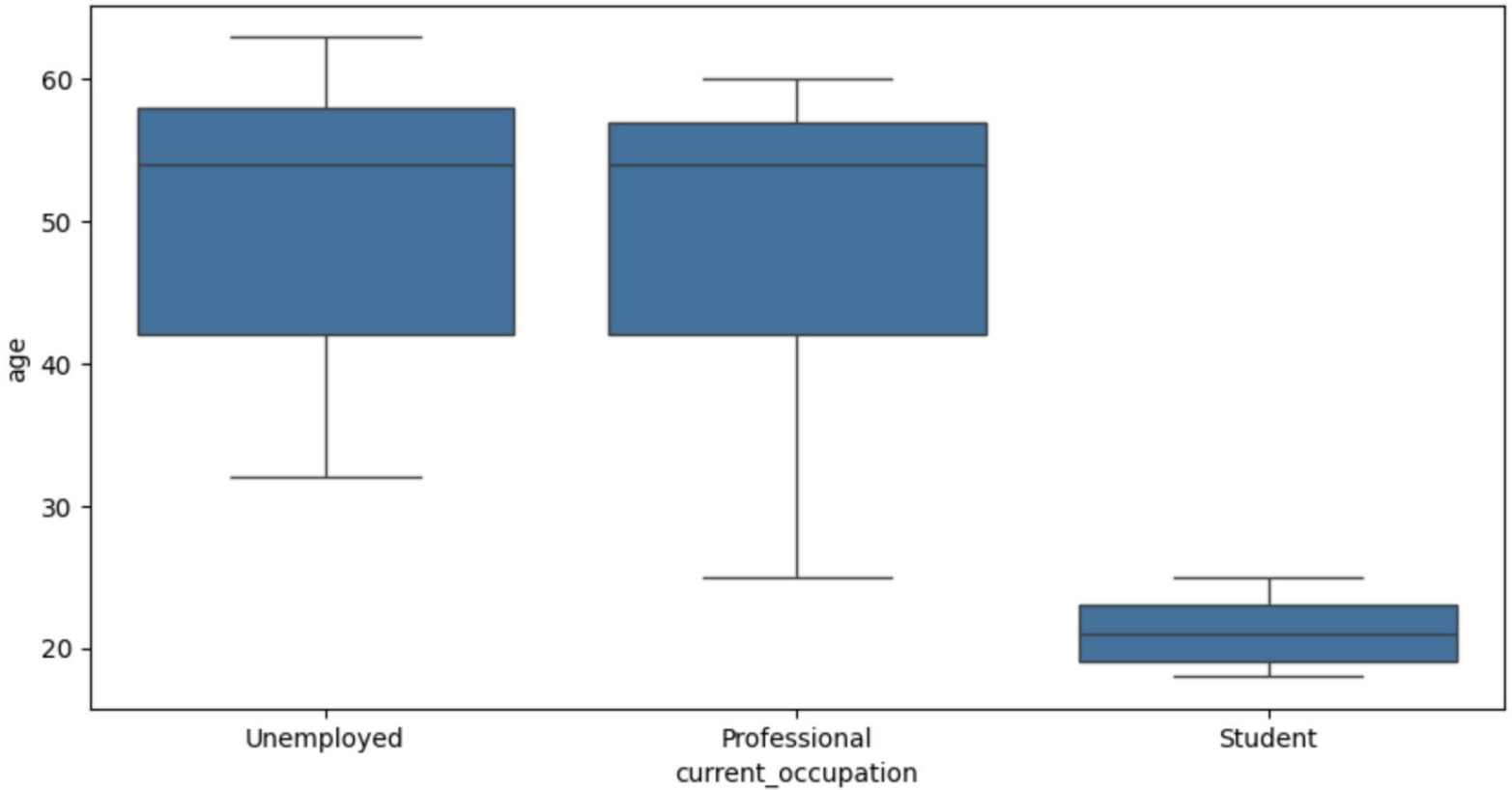
EDA - current occupation vs conversion

Does the channel of the first interaction impact lead conversion?

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



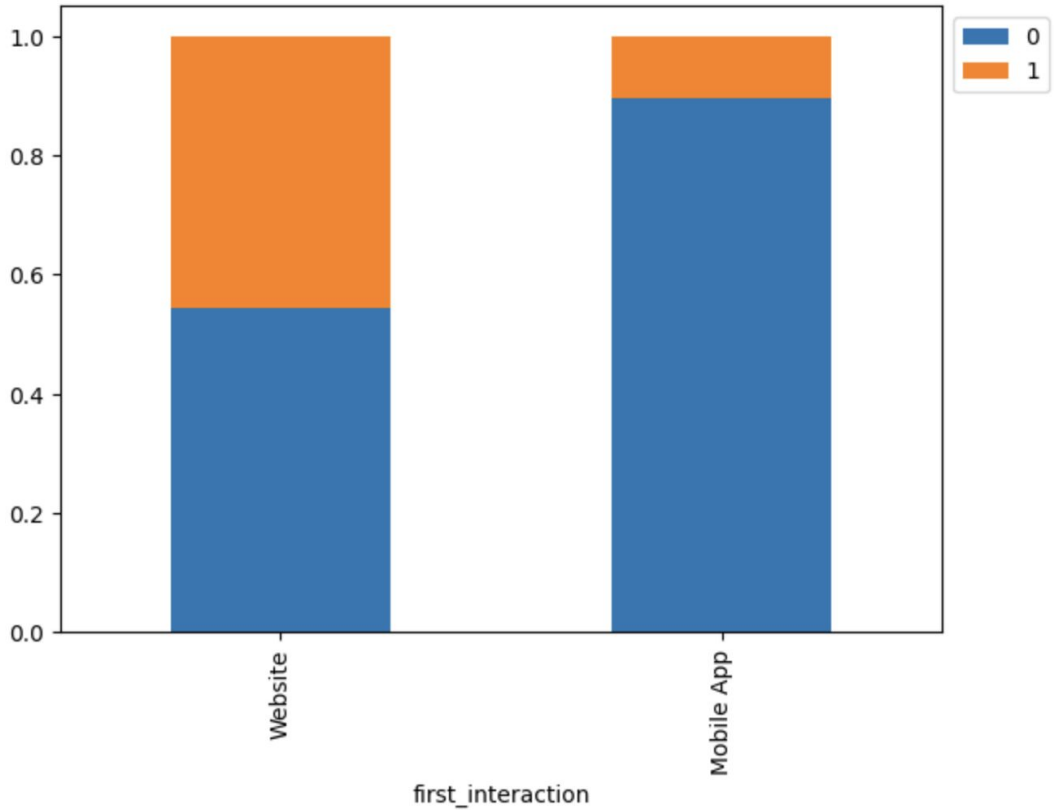
EDA - current occupation vs age



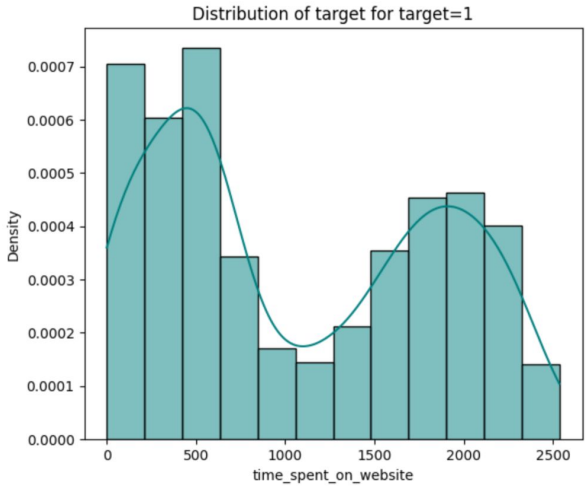
EDA - First interaction vs Status

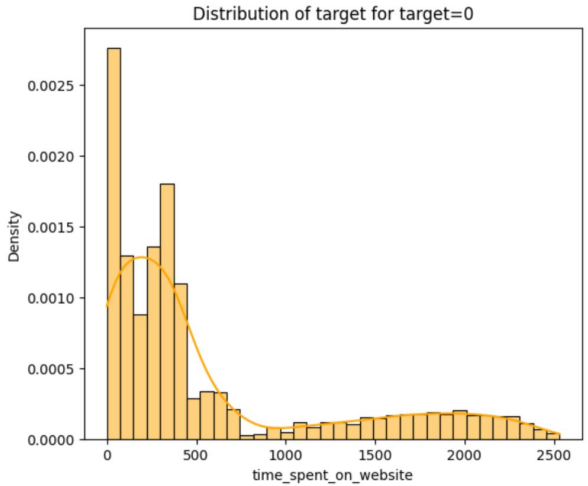
Does the channel of first interaction impact lead conversion

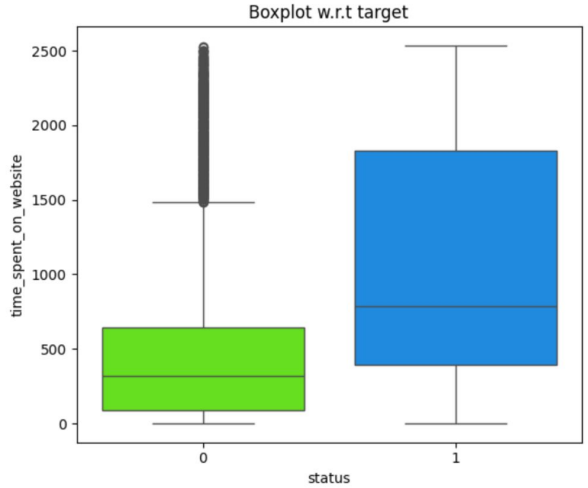
status	0	1	All
first_interaction			
All	3235	1377	4612
Website	1383	1159	2542
Mobile App	1852	218	2070

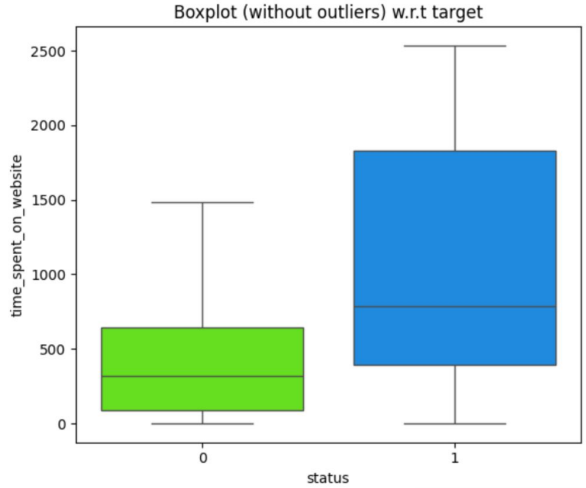


EDA - time spent on website vs Conversion









Further exploratory data analysis

- We further analyzed 7 categorical variables against the lead conversion status using barplots. They gave us some valuable insights about our leads.
- Leads from educational channels convert at a slightly higher rate than average.
- 68% of the referred leads converted to customers.
- Phone activity results in lowest conversion rate, followed by email activity and while website activity shows the highest conversion share.

Outlier Check

- No significant outliers were detected in age.
- Website visit had several outliers with more than 10 visits
- Time spent on website had outliers such as time exceeding 2500 seconds
- Page viewed per visit had outliers that exceeded 10 pages.
- Outliers were not removed from the database as they may represent highly engaged leads.

EDA - Insights gained so far

- Professionals and unemployed individuals form the core age demographic of leads with a median age of 54, while students are significantly younger with median age being 21
- Students are the least likely to convert followed by unemployed people.
- Leads who converted were spending a median of 13 mins on the website.
- We also see that non-converters drop out of being customers early on while converting audience have significantly higher media time spent of the website

Model Building

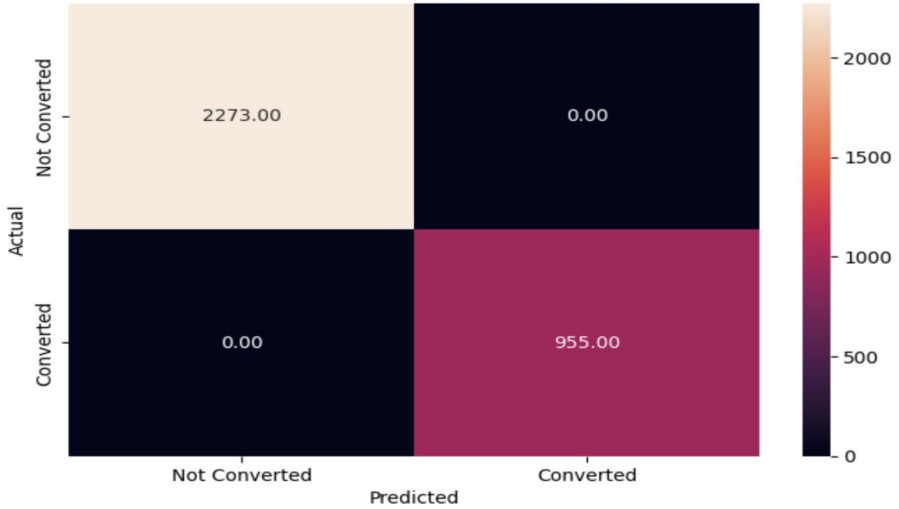
- Provide insights on the performance of different models.
- Provide comments about model performance after tuning the hyperparameter using GridSearchCV.
- Choose the model performance metric and provide reasoning for the same.

***Note**: You can use more than one slide if needed*

Model Building

Model Training

	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



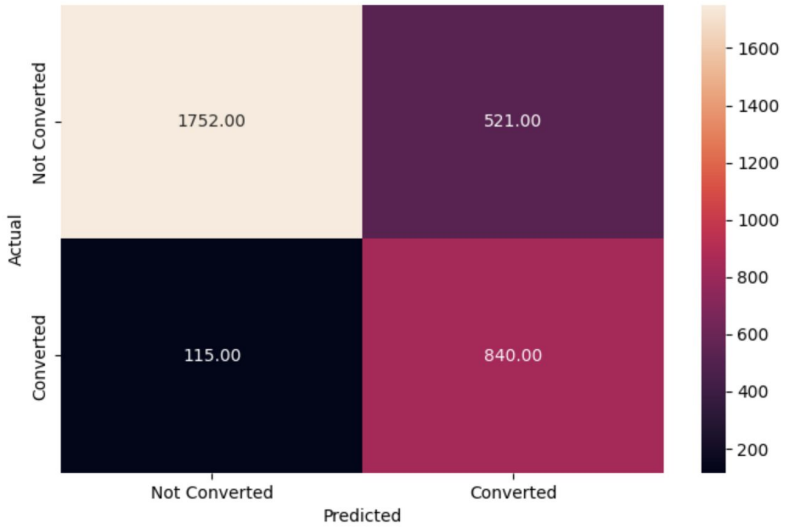
Model Testing

	precision	recall	f1-score	support
0	0.88	0.90	0.89	962
1	0.75	0.71	0.73	422
accuracy			0.84	1384
macro avg	0.81	0.81	0.81	1384
weighted avg	0.84	0.84	0.84	1384

Model Building - After tuning

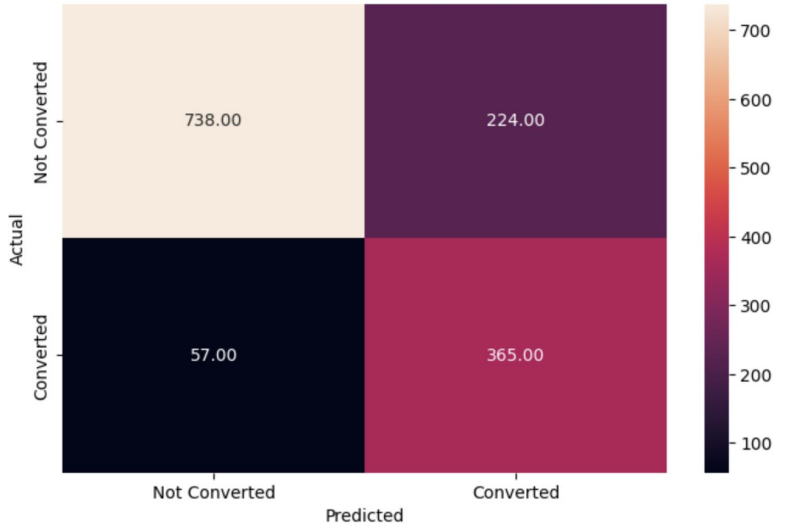
Model Training

	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



Model Testing

	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

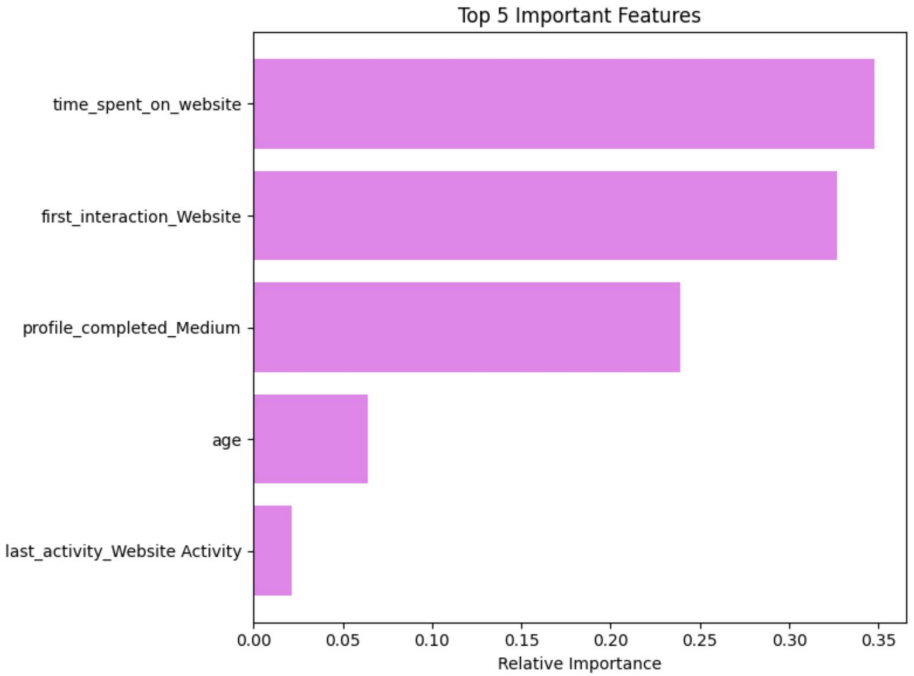


Model Building

- Decision trees overfit easily. Model tuning is essential to improve generalization and ensure the model can reliably identify high-converting leads on new data.
- As seen, tuning the decision tree greatly improved the model's ability to identify converting leads without overfitting. This could potentially be valuable to the business by minimizing the risk of lost potential leads.

Feature Importance - Decision tree

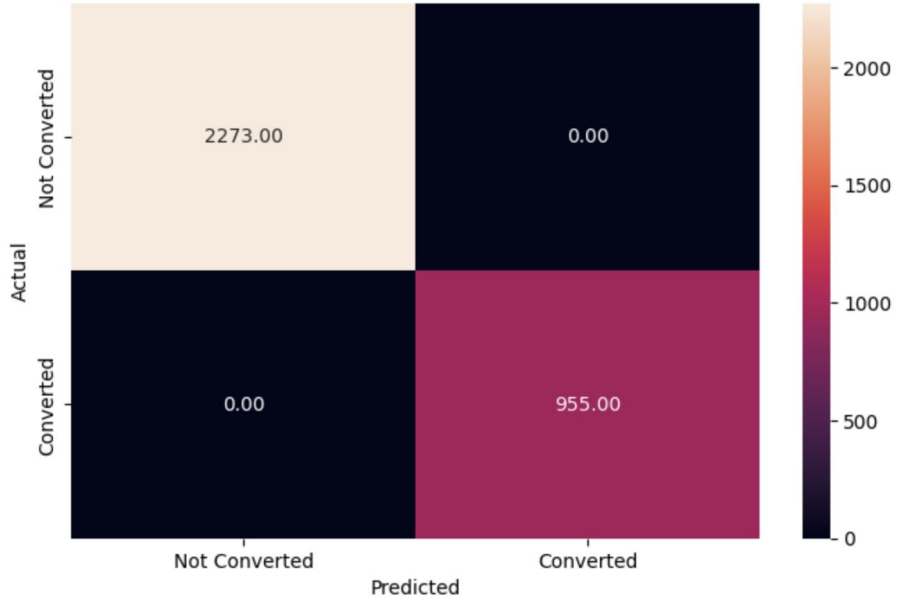
- Time spent on the website and first interaction are most important followed by profile completed, age and then last activity at the end.
- The decision tree generated gave some simple suggestions such as:
  - Suggest how the lead first interacts with the platform play major role in predicting conversion
  - Website activity matters more than email/phone.
- Overall we learn that it does give us an understanding of who might or might not convert but it is best to use it alongside other methods.



Random Forest Classifier

Model Training

	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



Model Testing

	precision	recall	f1-score	support
0	0.86	0.95	0.90	962
1	0.85	0.63	0.72	422
accuracy			0.85	1384
macro avg	0.85	0.79	0.81	1384
weighted avg	0.85	0.85	0.85	1384

Random Forest Classifier

- During the hyperparameter tuning, the execution time was taking a very long time so in order to balance model performance and computation efficiency, we reduced the parameter grid.
- Focusing on the most impactful parameters such as n_estimators, max_depth, min_samples_leaf and class_weight

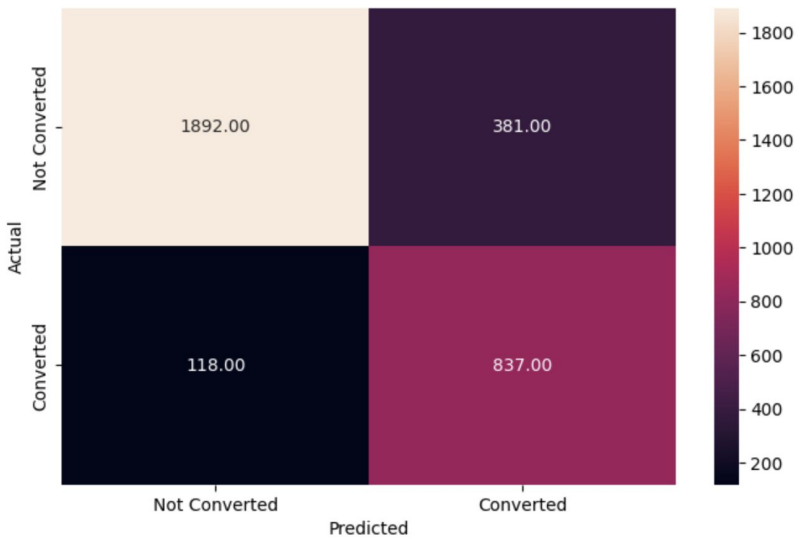
Random Forest Classifier - Hyperparameter tuning

Model Training

	precision	recall	f1-score	support
0	0.94	0.83	0.88	2273
1	0.69	0.88	0.77	955
accuracy			0.85	3228
macro avg	0.81	0.85	0.83	3228
weighted avg	0.87	0.85	0.85	3228

Model Testing

	precision	recall	f1-score	support
0	0.94	0.83	0.88	2273
1	0.69	0.88	0.77	955
accuracy			0.85	3228
macro avg	0.81	0.85	0.83	3228
weighted avg	0.87	0.85	0.85	3228



Model Performance Summary

Model	Accuracy	Recall	Precision	F1-score
Decision Tree (Default)	0.84	0.71	0.75	0.73
Decision Tree (tuned)	0.80	0.86	0.62	0.72
Random Forest (Default)	0.85	0.63	0.85	0.72
Random Forest (Tuned)	0.85	0.88	0.69	0.77

Model Performance Summary

- Default Random Forest performs well in terms of precision and accuracy but has lower recall meaning that it misses true converters.
- Tuned Decision Tree improved the recall significantly but lost some precision.
- Tuned Random Forest provides the best balance of recall, F1 score and accuracy, making it the most reliable model for identifying converters,
- Default Random Forest has high overfitting on the training data, so the gap between train and rest recall makes the model more trustworthy.
- With that understood best model choice is **Tuned Random Forest Classifier** as it:
    - Prioritizes recall which is very important to minimize loss of potential leads as we need to maximize the customers who will become a real customers as compared to pursuing a someone who will not convert.

Business recommendations

- Focus on leads with high website engagement. They are more likely to convert.
- First interaction with the website and recent activity are strong indicators of conversion.
- Encouraging profile completion can significantly improve conversion rates.
- Referred leads showed a high conversion rate so focusing on expanding the referral programs to possibly lead to more converters, could be beneficial.
- Deprioritize phone only leads since they have the lowest conversion rate. Focusing on email or digital follow-ups for these cases to get them to re-engage.

Great Learning

Happy Learning !