ExtraaLearn Customer Conversion Analysis

Practical Data Science

Stephanie Jennings 5/27/24

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

This report is to help EdTech company, ExtraaLearn, determine best practices for converting potential customers, or leads, into paying customers. The report includes statistical summaries of lead data, as well as a constructed decision tree model to predict what factors are most relevant in converting leads to customers.

In order to do this, data about the leads was taken to find the most important variables about each potential customer. This includes their age, first interactions with the company, their experiences with different advertising strategies, the extent of their involvement with the website or app, among others. The data were taken and used to create both a decision tree and a random forest model to predict the best outcomes. After evaluating the models, the random forest was a better fit for this problem because we want to reduce the number of false negatives detected in the data.

Business insights and recommendations include focusing on improving the website, in addition to offering one free lesson for leads to fully experience ExtraaLearn before deciding to pay for a course. Also, ExtraaLearn should focus their advertising strategies to target older people and remove print ads.

Data Overview

Brief overview of the data:

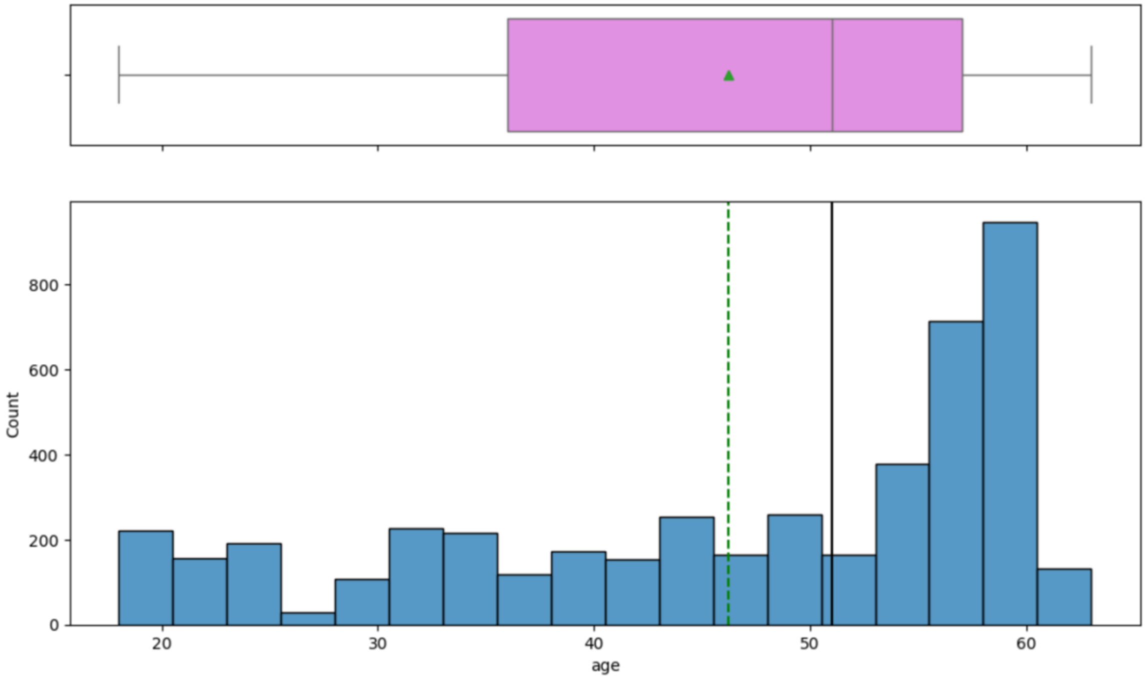
1. The data includes 4612 rows and 15 columns. The columns include information about the leads such as: ID, age, current occupation (professional, unemployed, or student), first interaction (how leads first interacted with ExtraaLearn), profile completed (percentage of the profile that is complete), website visits (number of times a lead visited the website), time spent on the website (in seconds), page views per visit (average number of pages viewed per visit), and last activity (the last interaction with the lead and the company). The other columns include flags about how the leads found ExtraaLearn: print media type 1 (newspaper), print media type 2 (magazine), digital media, educational channels, or referral. The final column is status, which tells us if a lead was converted to a paying customer or not.
2. Columns 1, 5, 6, and 14 are integer values; columns 0, 2, 3, 4, 8, 9, 10, 11, 12, and 13 are object values; and column 7 is a float value.
3. There are no missing (null) values in the data.

EDA Results

Based on certain attributes of the leads, there are some trends in the types of people viewing the website.

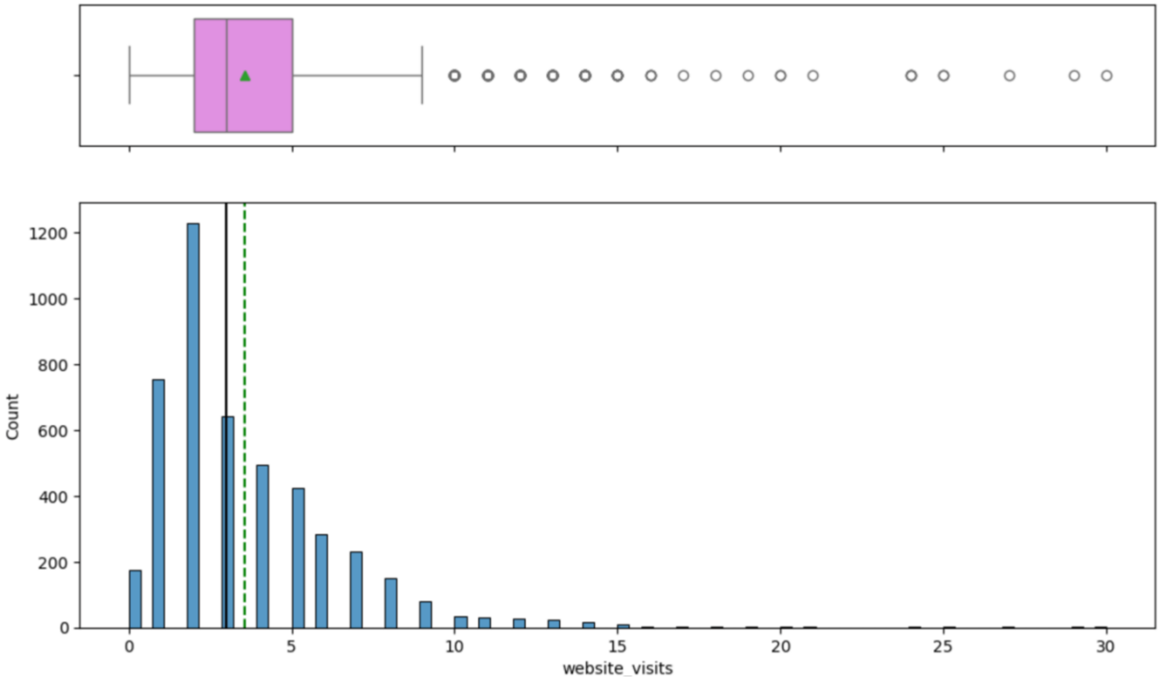
Age

- Most of the leads are over 40 years old, with the average age being 46.2. The youngest person is 18 and the oldest is 63. There were no outliers in the data but the data is skewed left.



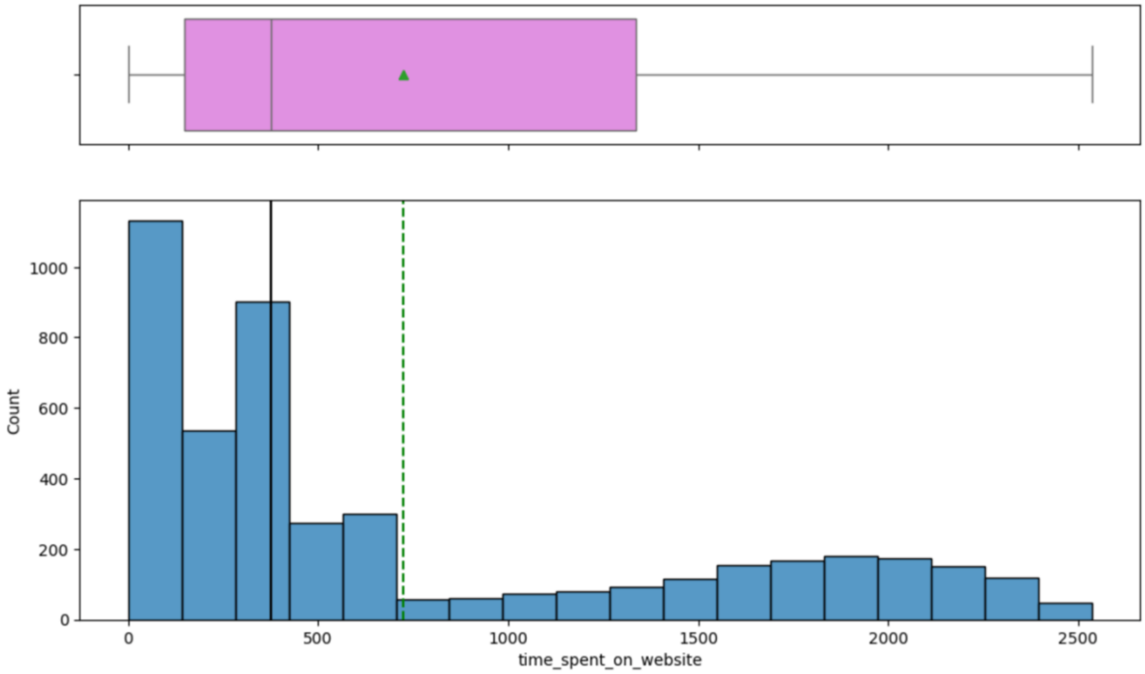
Website Visits

- The average amount of website visits is 3.6 visits, while the most visits is 30. The lowest number of visits is 0, but most tend to be between 1 and 5 visits. There are many outliers towards the high end of website visits, making this data skewed right, and 174 people have visited the website 0 times.



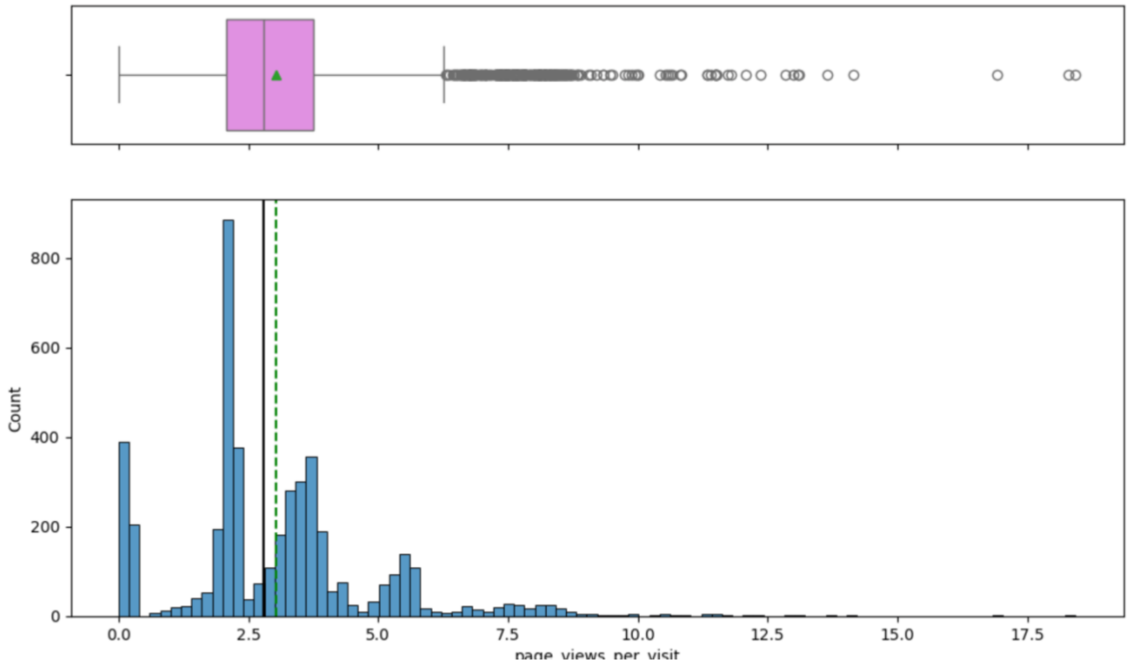
Time Spent on Website

- The average amount of time spent on the website is 724 seconds, while the minimum is 0 and the maximum is 2537 seconds. The median amount of time spend on the website is about 376 seconds, and the data is skewed right.



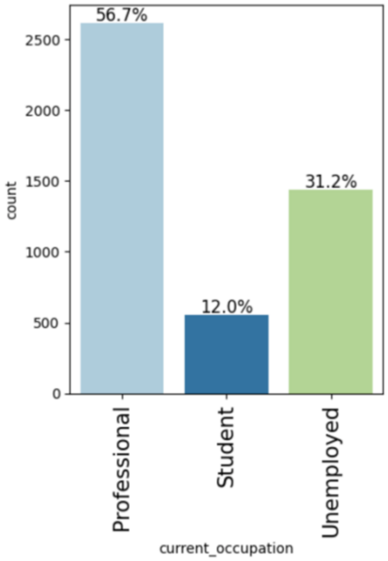
Page Views Per Visit

- The average amount of pages viewed per visit is about 3 and the median is similar. There are many outliers here on the high end and the most pages visited is 18.43.



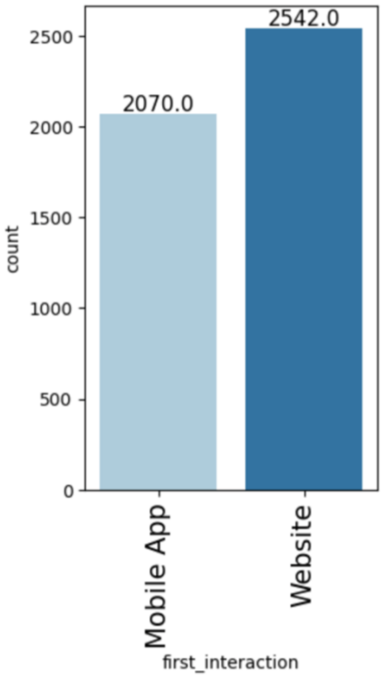
Occupation

- Most leads are professionals (56.7%), with unemployed being the next highest category (31.2%). Student is the lowest with 12% of people identifying as students.



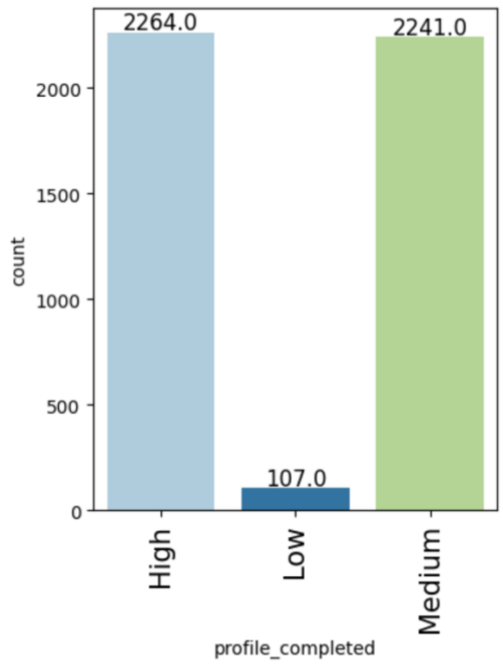
First Interaction

- There are more first interactions through the website than the mobile app, but they are relatively close.



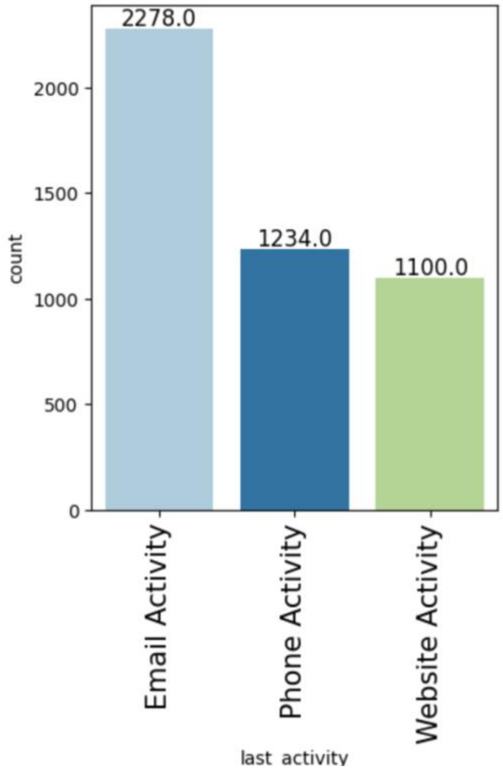
Profile Completed

- Most people have completed over at least half of their profile, with only 107 out of 4612 people completing less than 50%. The distribution of people who have completed 50-75% and 75-100% is about the same.



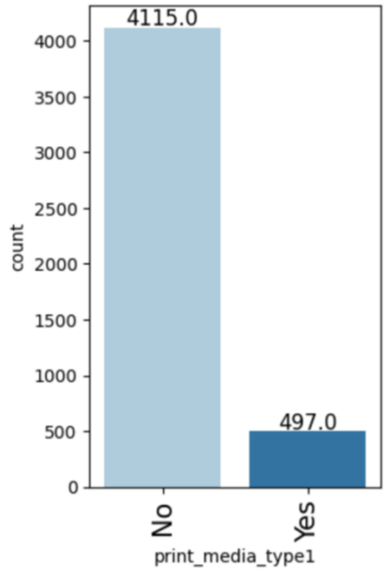
Last Activity

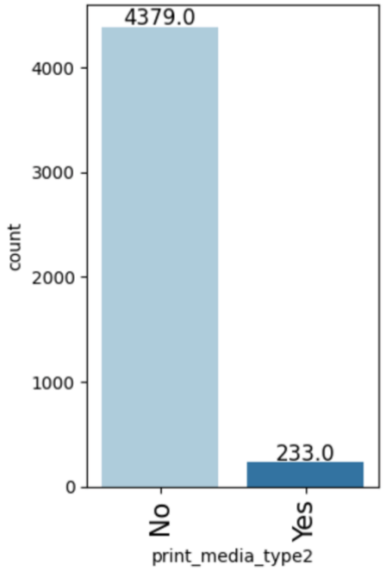
- The last interaction most people had was via email, while website and phone activity are about evenly split.

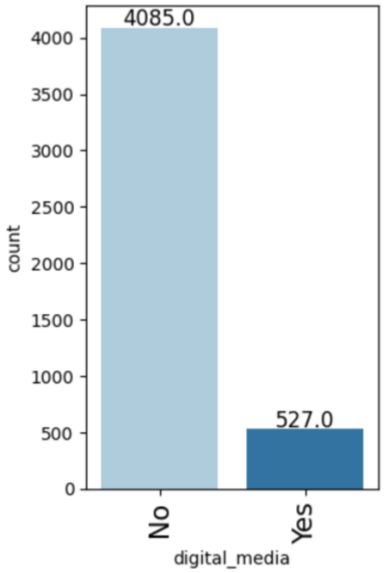


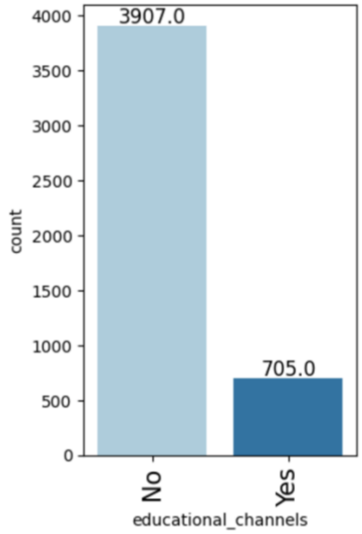
Advertising

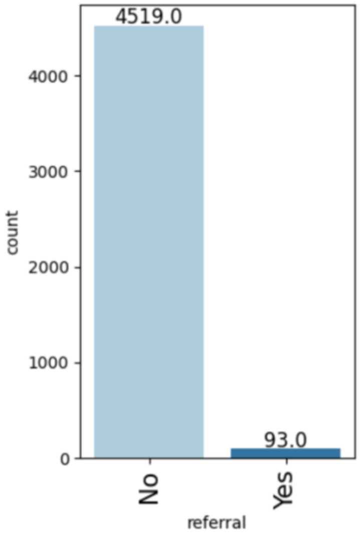
- For all media types (newspapers, magazines, and digital), the large majority of people have not interacted with them. Of the three media types, digital media has the most interactions with 527. Educational channels have more reach than the others, with 705 people hearing about ExtraaLearn from there. Referrals gain the least amount of leads with 93.





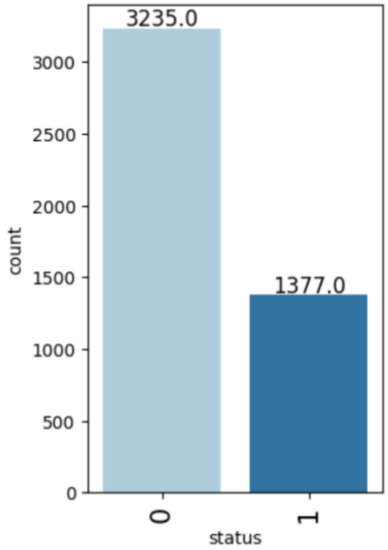






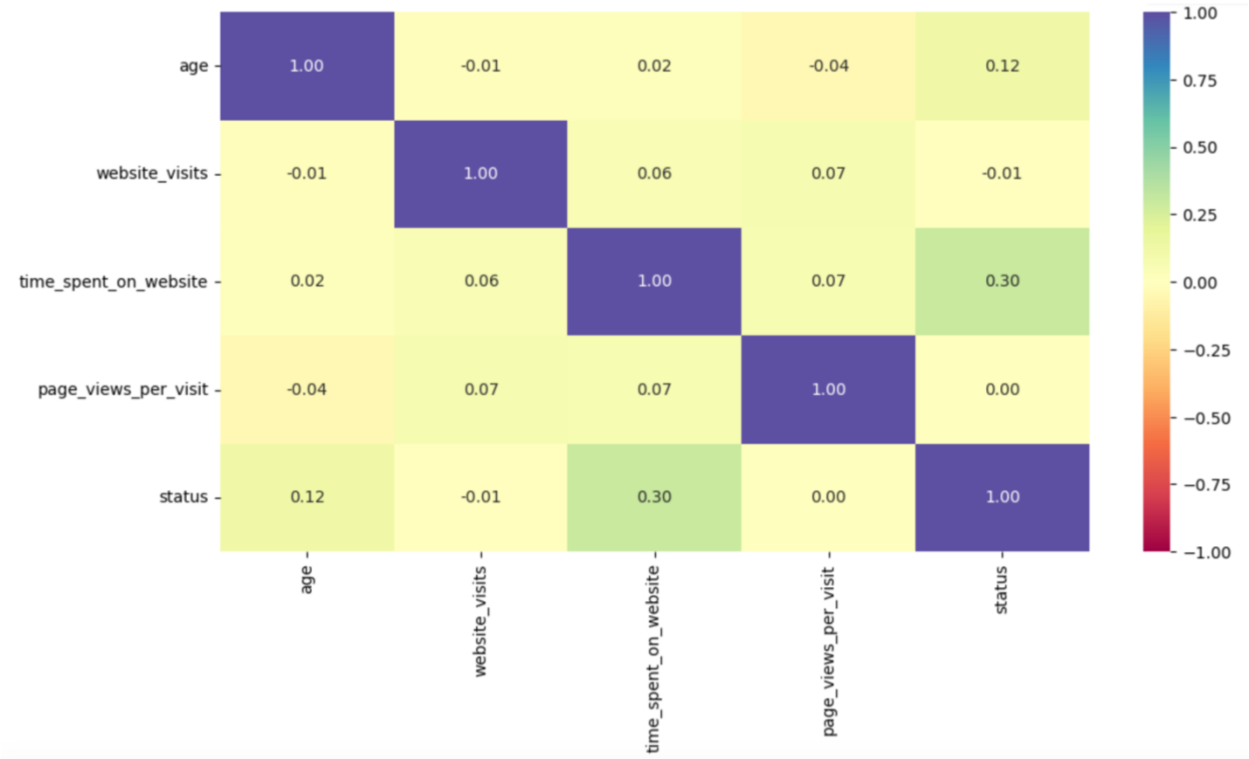
Status

- Out of the 4612 leads, 1377 of them were converted into paying customers.



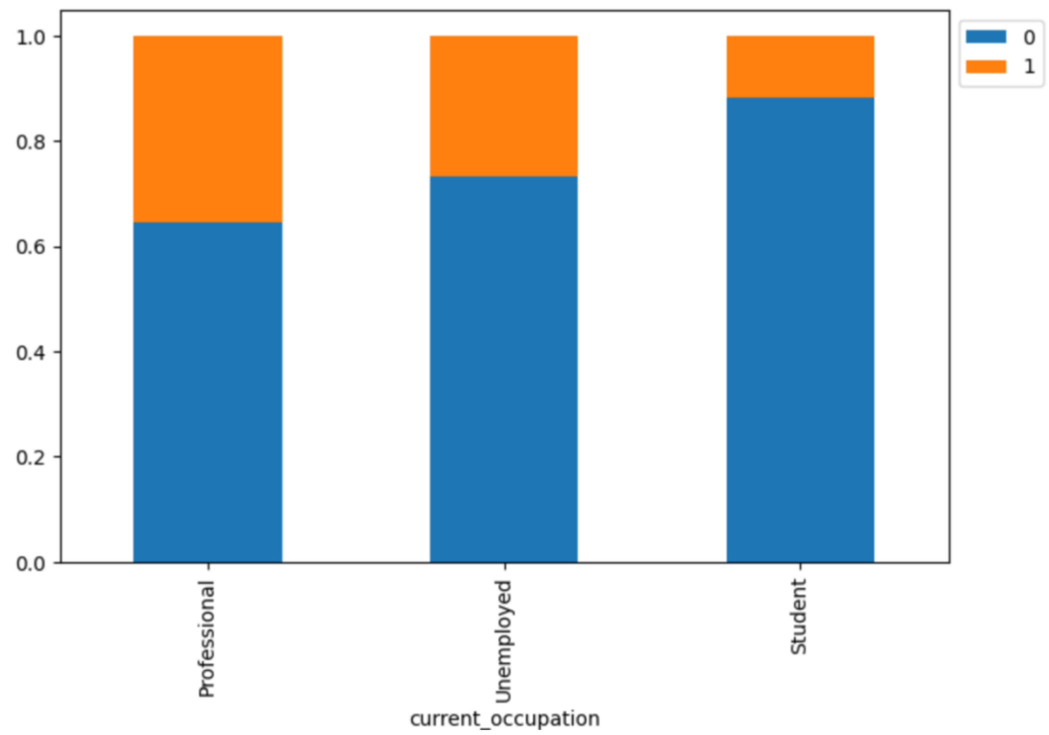
Bivariate Analysis

- The heat map below shows that not many numerical variables are highly correlated to each other, either positively or negatively. The highest positive correlation is between time spent on the website and status, meaning that more time spent on the website was correlated to a higher lead conversion. All other correlations are insignificant.



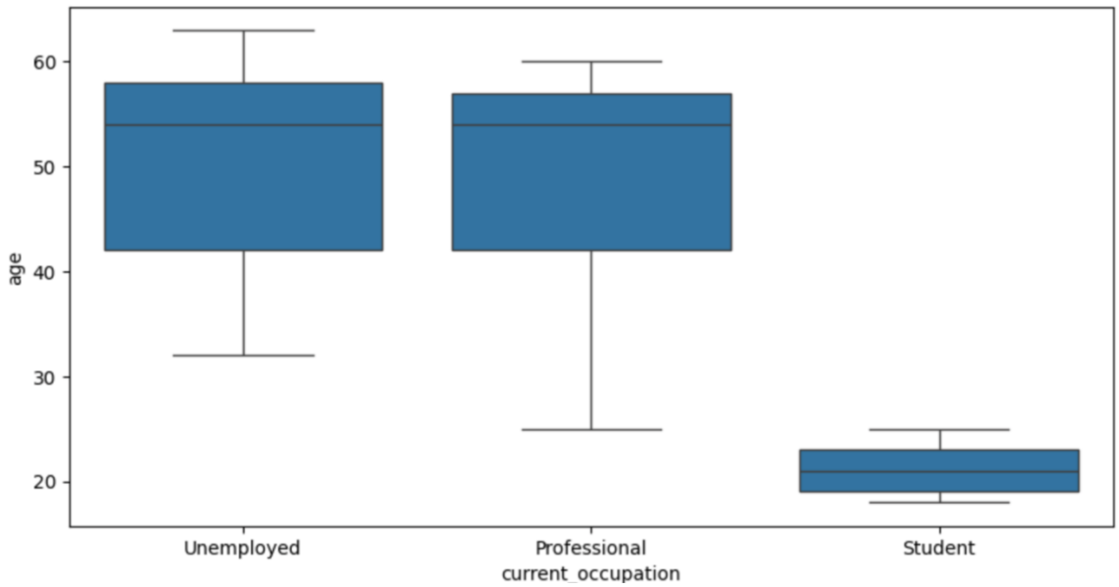
Occupation vs. Status

- The distribution of people that became paying customers based on their occupation is shown below. More people that are paying customers are categorized as a "Professional" and the next highest category is people that are unemployed. Most students did not become paying customers.



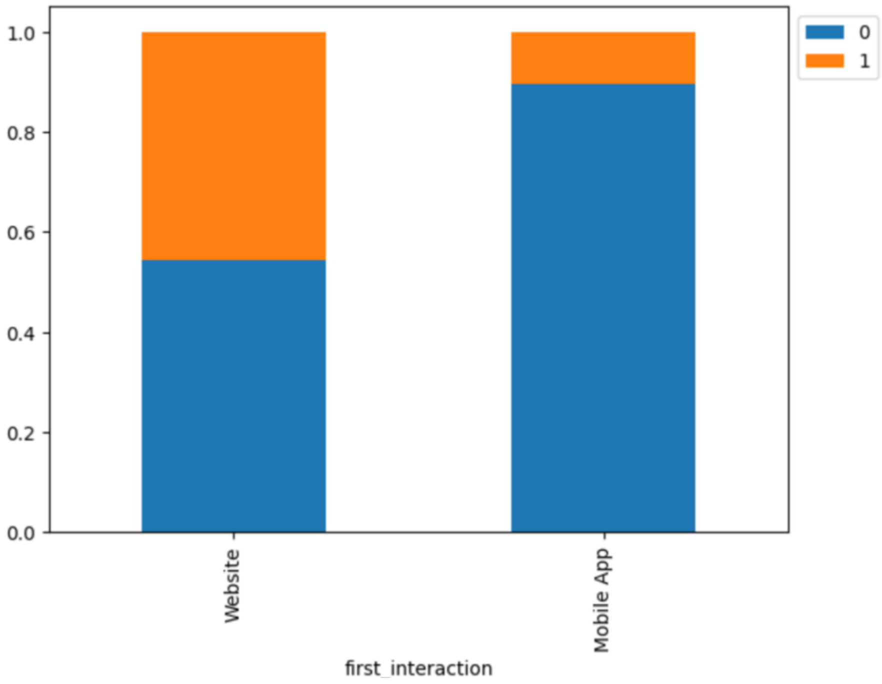
Occupation vs. Age

- People who are working or unemployed are older, while those that are students are significantly younger. That means that people that are younger are less likely to become paying customers of ExtraaLearn based on the previous comparison of occupation and status.



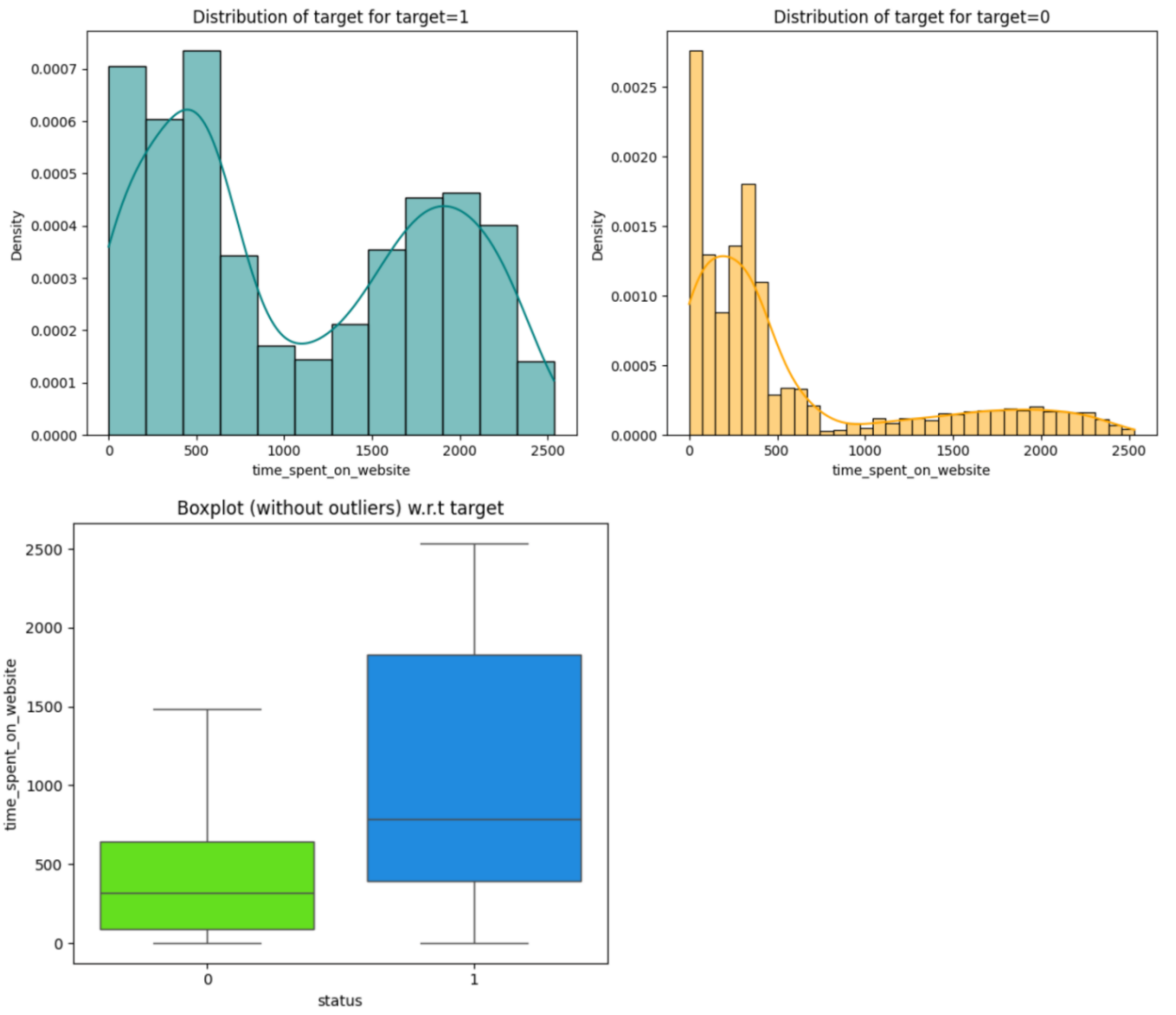
First interaction vs. Status

- Most people who did become paying customers had their first interaction with the company on the website and not on the mobile app. This could mean many things- either the app is not appealing enough or the website has a better interface to give someone a more comprehensive experience.



Time Spent on Website vs. Status

- By dividing leads up by their status, we can see the distribution of the time spent on the website. The blue graph shows leads that were successfully converted and the amount of time spent on the website for this group is bimodal and more spread out over varying amounts of time. The group of leads that did not convert (in yellow) tends to spend much less time on the website and is skewed right.

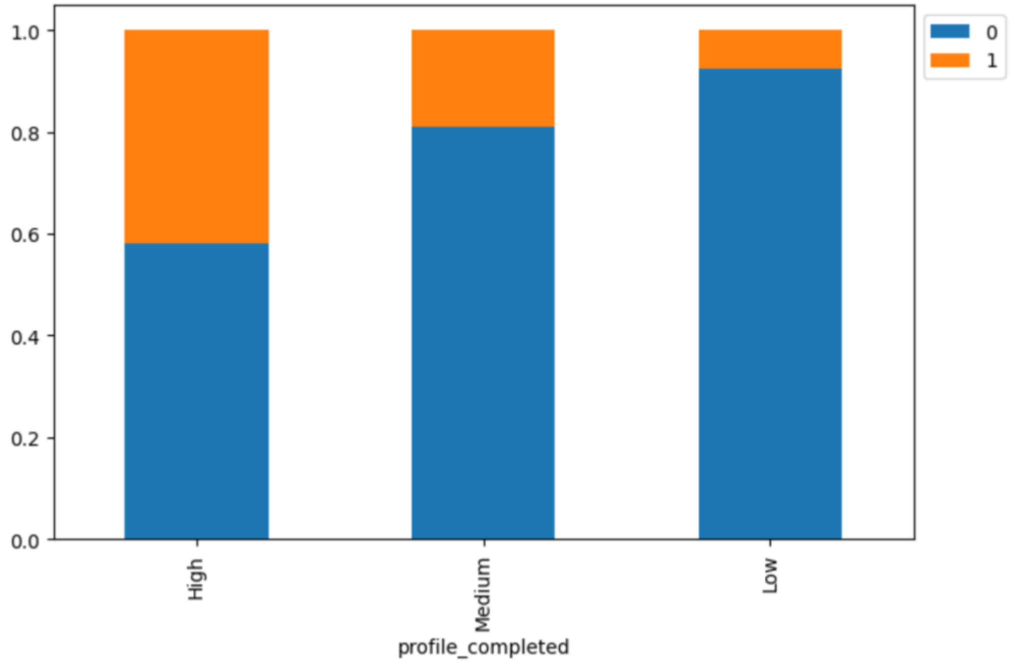


Other Website Variables vs. Status

- When comparing website visits and status, there is no significant difference in the distribution of visits based on if the lead was converted or not. Similarly, there is no significant difference in the distribution of page views per visit based on if the lead was converted or not.

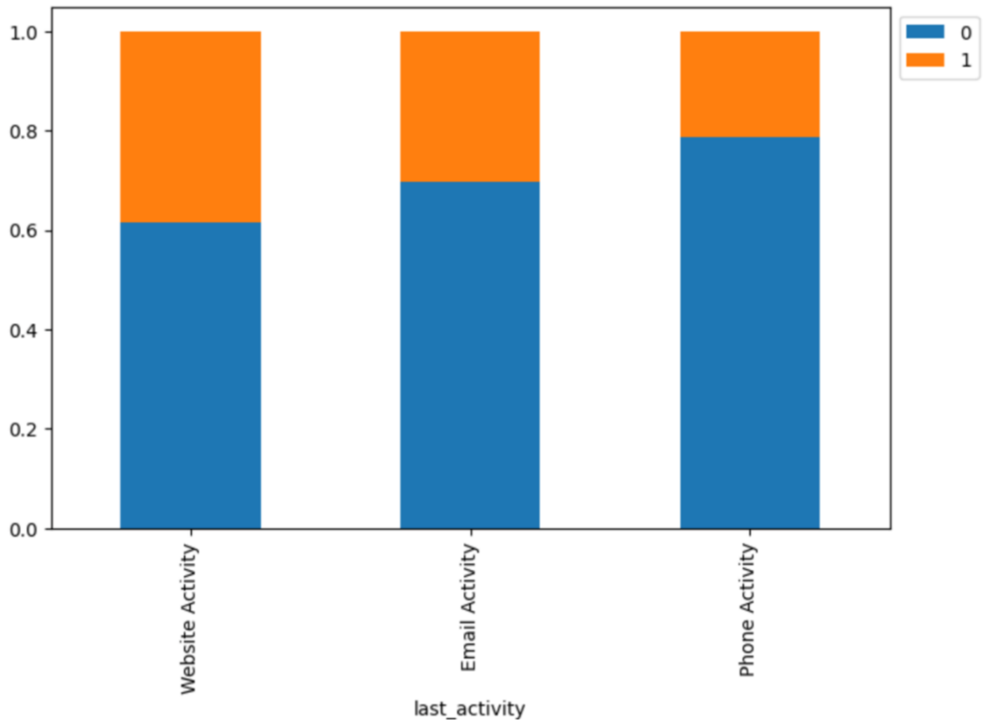
Profile Completed vs. Status

- People that completed 75-100% of their profile are most likely to be a converted lead compared to those that completed less of their profile. Those that completed 50-75% were more likely than those that completed less than 50% of their profile to be converted, which shows a correlation between completing more of the profile and being a successfully converted lead.



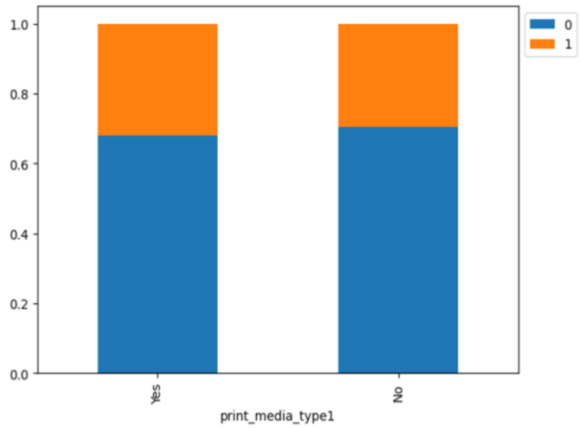
Last Activity vs. Status

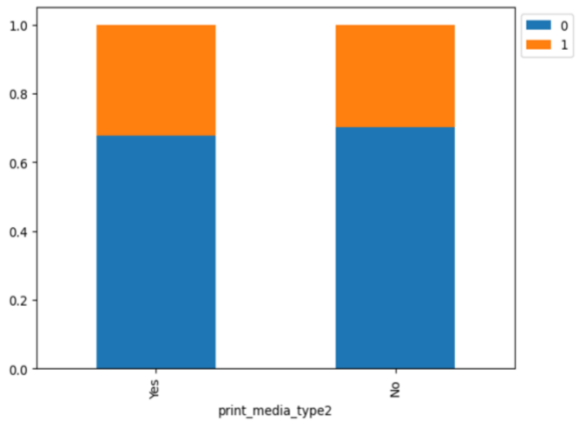
- Leads that last communicated via the website are more likely to be converted than via email or phone. Phone activity has the least amount of successful lead conversions, which means both the website and email are better ways to communicate with leads in order to convert them into paying customers.

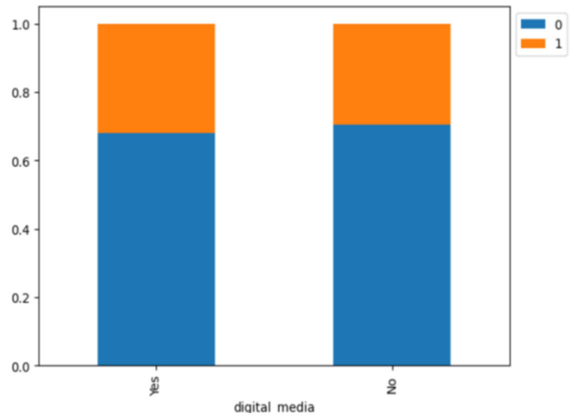


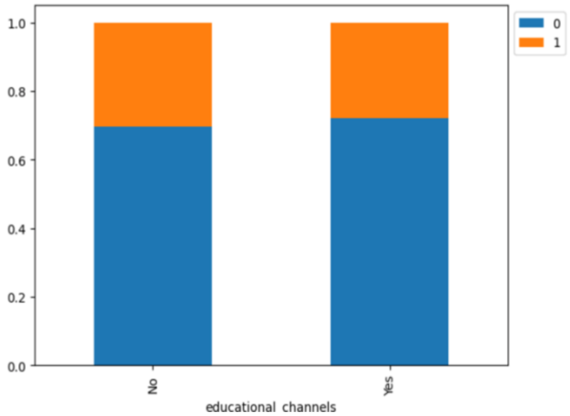
Advertising vs. Status

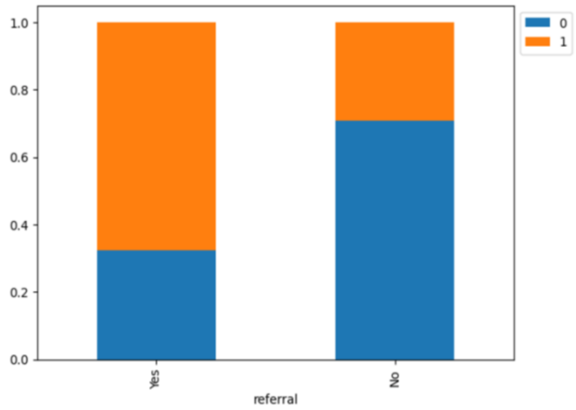
- All advertising channels seem to have little correlation to the amount of leads that are converted successfully. Both print media types, digital media, and educational channels all seem to have about the same amount of successful conversions based on if someone saw the ad. The only type of advertising that has a significant positive effect is referrals, which could mean that people are more likely to become paying customers if they hear about the company from someone they trust.


1.0
0.8
0.6
0.4
0.2
0.0
0
1
Yes
No
print_media_type1


1.0
0.8
0.6
0.4
0.2
0.0
0
1
Yes
No
print_media_type2


1.0
0.8
0.6
0.4
0.2
0.0
0
1
Yes
No
digital_media

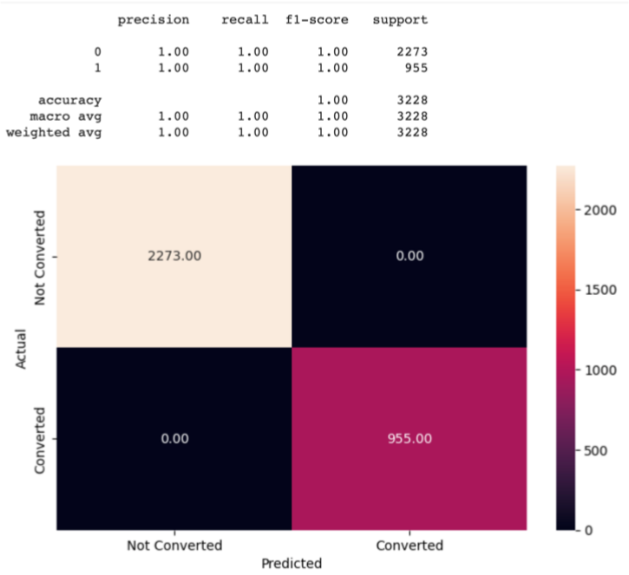

1.0
0.8
0.6
0.4
0.2
0.0
0
1
No
Yes
educational_channels


1.0
0.8
0.6
0.4
0.2
0.0
0
1
Yes
No
referral

Model Building

Decision Tree Model

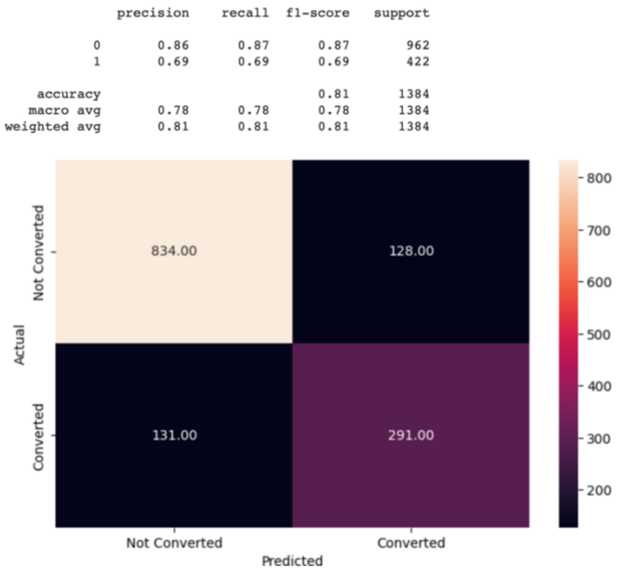
- Before hypertuning the model, the decision tree had a recall of 1 on the training data for both outcomes. On the testing data, the decision tree had a recall of 0.87 for unsuccessful lead conversions and a recall of 0.69 for successful lead conversions. The model has more false negatives than false positives, but is still pretty accurate.



```
              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
```

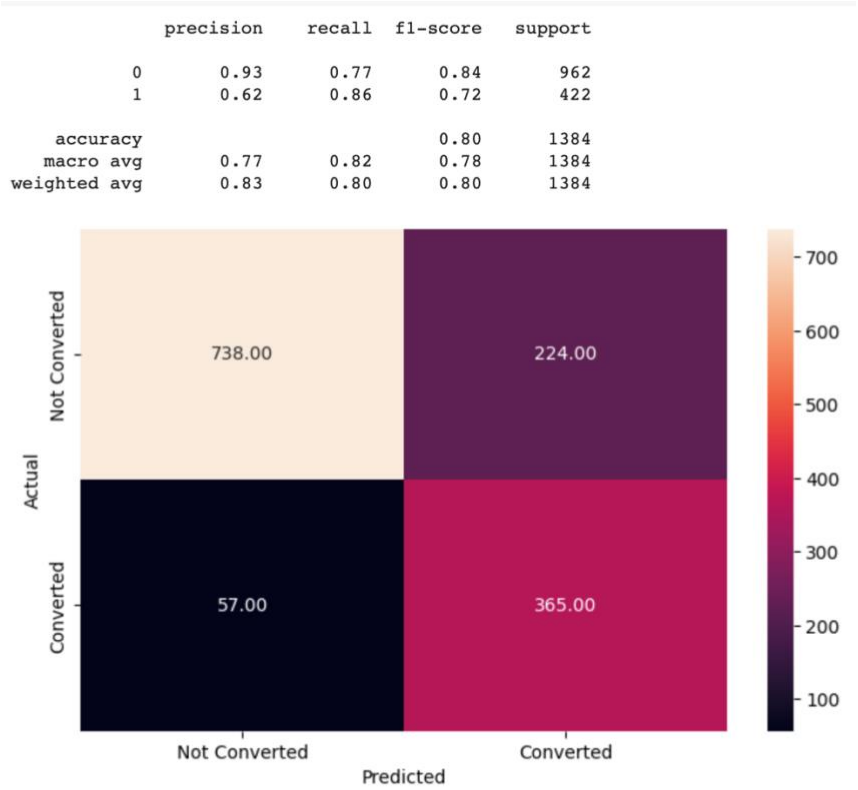


```
              precision    recall  f1-score   support

           0       0.86      0.87      0.87       962
           1       0.69      0.69      0.69       422

    accuracy                           0.81      1384
   macro avg       0.78      0.78      0.78      1384
weighted avg       0.81      0.81      0.81      1384
```

- After hypertuning to reduce overfitting, the model had a recall of 0.77 for unconverted leads and 0.86 for converted leads on the testing data. The model has the least number of false positives compared to the other categories.

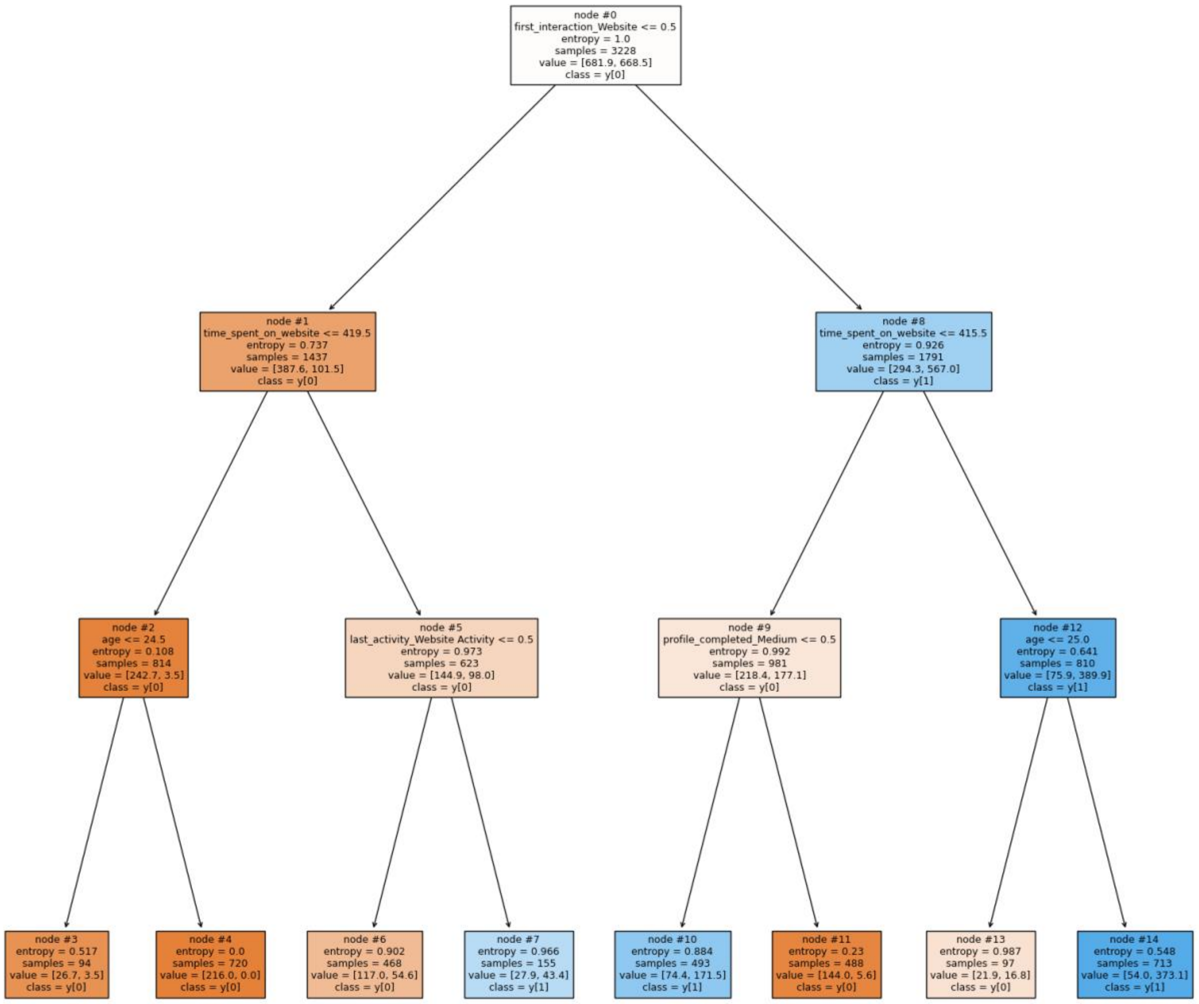


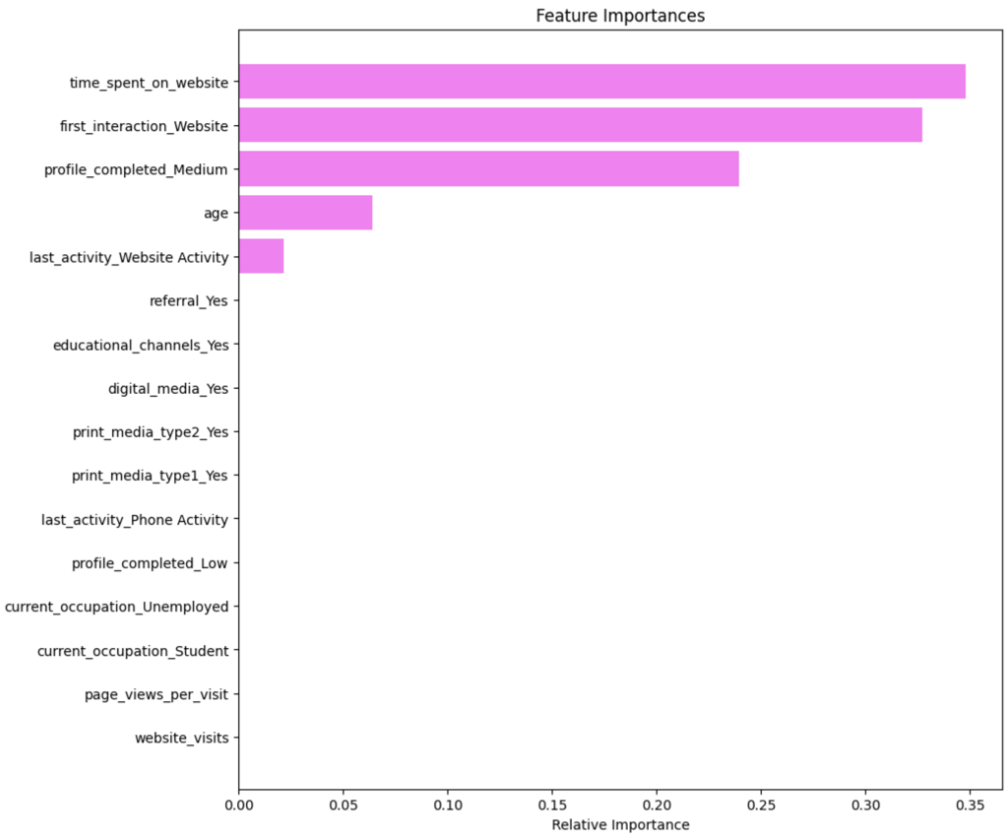
```
              precision    recall  f1-score   support

           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
```

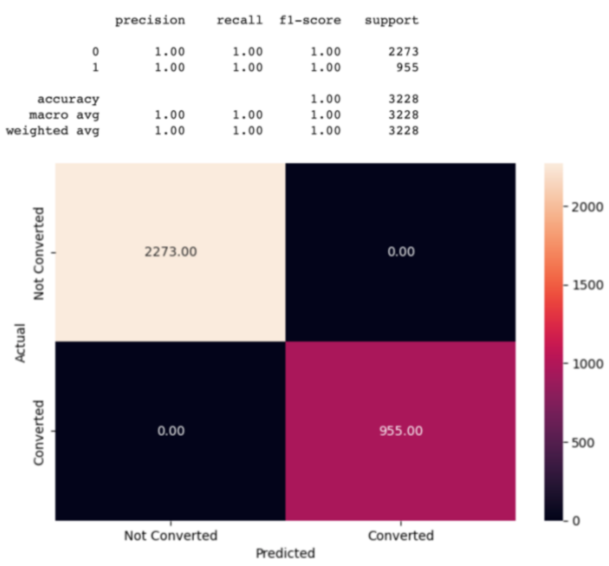
- The tuned decision tree is shown below. First interaction is the root node, while time spent on website is the next decision node. After that, the decision nodes vary on which variable they use to make a decision. The most successful conversions first interacted on the website, spent more time on the website, and were older than 25. This is also shown in the important features chart included after the decision tree.


node #0
first_interaction_Website <= 0.5
entropy = 1.0
samples = 3228
value = [681.9, 668.5]
class = y[0]
node #1
time_spent_on_website <= 419.5
entropy = 0.737
samples = 1437
value = [387.6, 101.5]
class = y[0]
node #8
time_spent_on_website <= 415.5
entropy = 0.926
samples = 1791
value = [294.3, 567.0]
class = y[1]
node #2
age <= 24.5
entropy = 0.108
samples = 814
value = [242.7, 3.5]
class = y[0]
node #5
last_activity_Website Activity <= 0.5
entropy = 0.973
samples = 623
value = [144.9, 98.0]
class = y[0]
node #9
profile_completed_Medium <= 0.5
entropy = 0.992
samples = 981
value = [218.4, 177.1]
class = y[0]
node #12
age <= 25.0
entropy = 0.641
samples = 810
value = [75.9, 389.9]
class = y[1]
node #3
entropy = 0.517
samples = 94
value = [26.7, 3.5]
class = y[0]
node #4
entropy = 0.0
samples = 720
value = [216.0, 0.0]
class = y[0]
node #6
entropy = 0.902
samples = 468
value = [117.0, 54.6]
class = y[0]
node #7
entropy = 0.966
samples = 155
value = [27.9, 43.4]
class = y[1]
node #10
entropy = 0.884
samples = 493
value = [74.4, 171.5]
class = y[1]
node #11
entropy = 0.23
samples = 488
value = [144.0, 5.6]
class = y[0]
node #13
entropy = 0.987
samples = 97
value = [21.9, 16.8]
class = y[0]
node #14
entropy = 0.548
samples = 713
value = [54.0, 373.1]
class = y[1]



Random Forest Model

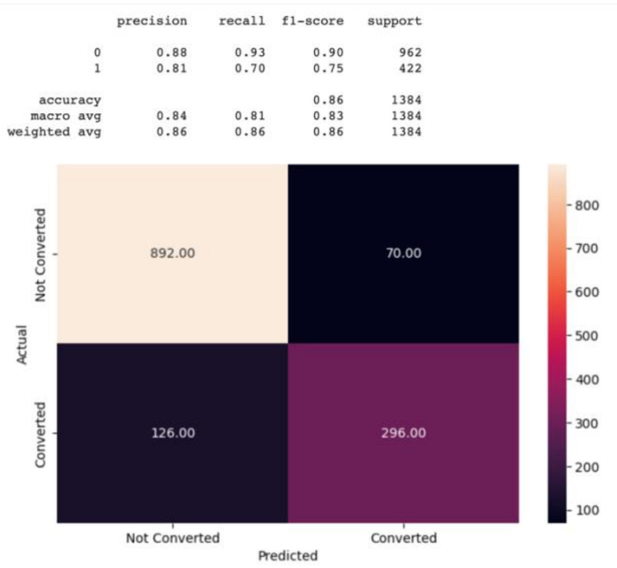
- **Before hypertuning the model, the random forest had a recall of 1 on the training data for both outcomes. On the testing data, the decision tree had a recall of 0.93 for unsuccessful lead conversions and a recall of 0.70 for successful lead conversions. The model also has more false negatives than false positives, but is still pretty accurate.**



```
              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
```

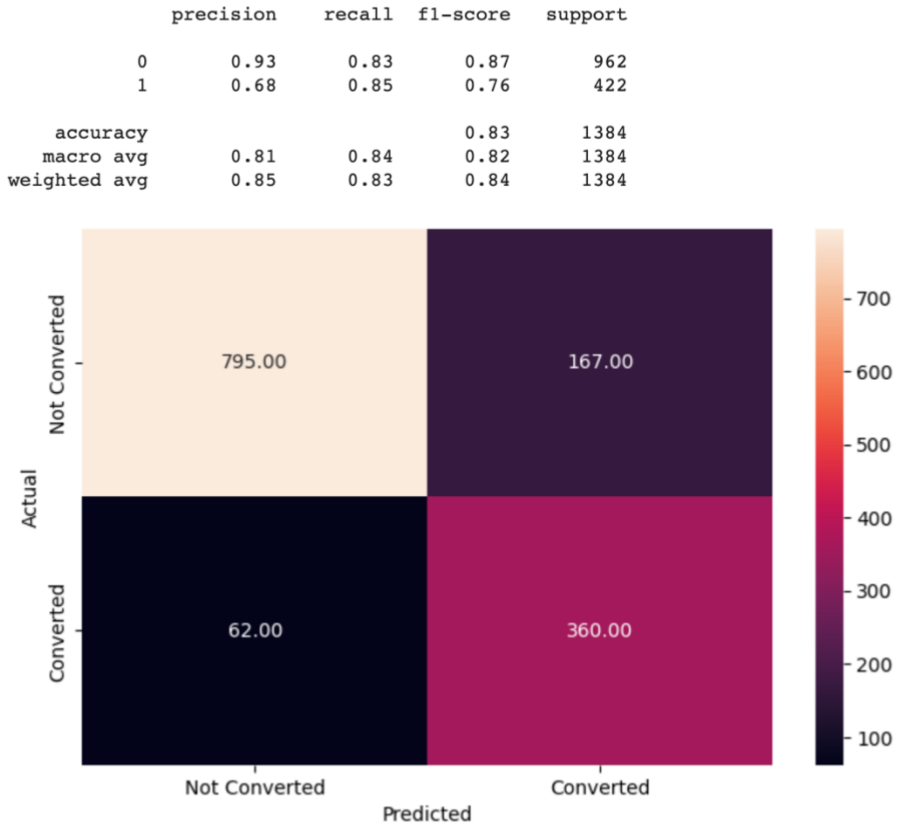


```
              precision    recall  f1-score   support

           0       0.88      0.93      0.90       962
           1       0.81      0.70      0.75       422

    accuracy                           0.86      1384
   macro avg       0.84      0.81      0.83      1384
weighted avg       0.86      0.86      0.86      1384
```

- **After hypertuning to reduce overfitting, the model had a recall of 0.83 for unconverted leads and 0.85 for converted leads on the testing data. The model also has the least number of false positives compared to the other categories.**

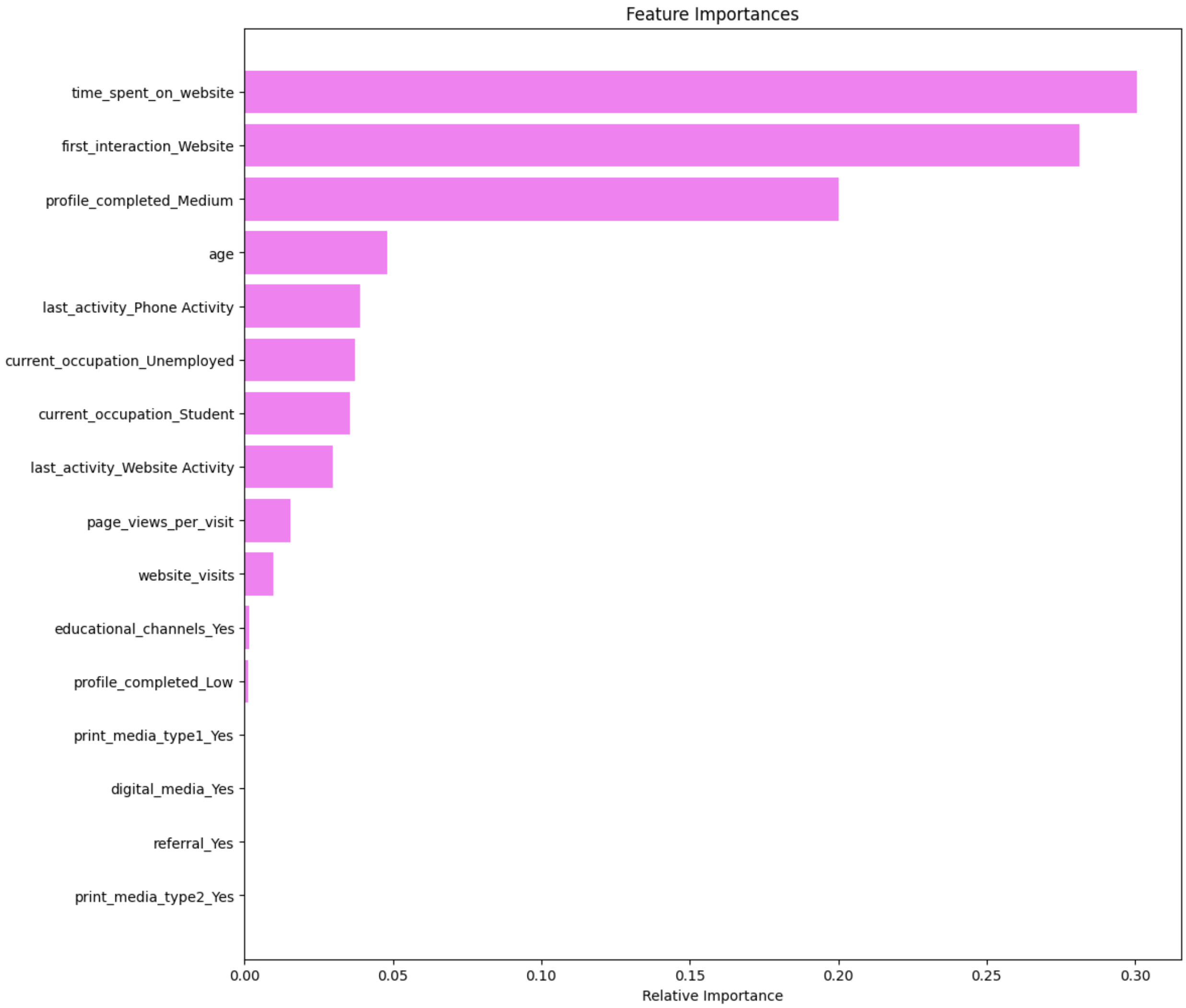


```
              precision    recall  f1-score   support

           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
```

- **The most important features in the random forest model are time spent on the website, the first interaction, and the amount of completion per profile followed by age.**



Model Performance Summary

- In the case of ExtraaLearn, losing a potential customer is a greater loss than it is to spend resources on a lead that does not get converted. Because of this, the recall score should be maximized to minimize the number of false negatives in the model.
- The random forest model has a better recall on the hypertuned data than the decision tree model. Random forest has an average recall of 0.84 and the decision tree has an average recall of 0.82. For this reason, the random forest model is better for this situation because it is more costly for ExtraaLearn to have false negatives than false positives.
- The random forest model indicates that the time spent on the website, the type of first interaction (specifically being on the website), and profile completion (50-75%) are all the best predictors for determining if a lead will be converted or not.

Conclusion

Based on this analysis, there are a few insights and recommendations to consider.

Insights:

First, the website is a key component in converting leads to customers. Both models highlight the time spent on the website and the first interaction on the website as important indicators on successful lead conversions. In addition to this, advertising is not a large factor in the success of converting leads. Neither model uses media types or other advertising as important indicators. Finally, most successful leads are older, and are not likely teenagers or young adults. That means that specific groups of people are more likely to become customers based on their age. Based on these insights, this is what I recommend.

Recommendations:

- Focus on improving the website and app interface to create a more enjoyable first experience for customers
    - The website is the most successful converter so far, and putting more effort into improving the app could also create more conversions if the experience of using the app is similar to the website
    - Continue to adjust the website to customer feedback and allow leads to experience what ExtraaLearn has to offer. Offering a demo lesson for leads would keep them on the website for longer and see if ExtraaLearn has what they are looking for before paying for an entire course
- Target middle-age or retired populations in advertising strategies and stop print-based advertising

- The current advertising strategies are not successful in converting leads so it is wasting money that could be used for other things
- A targeted marketing strategy towards older people would be more likely to convert leads because older people have more money to spend. Retired people also have more time to spend on extracurricular learning, which could be why they are more likely to be customers.