Predicting Customer Conversion: Identifying High-Value Leads Through Data

Project 2 - Classification & Hypothesis Testing

Amanda Edwards
6/23/25

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

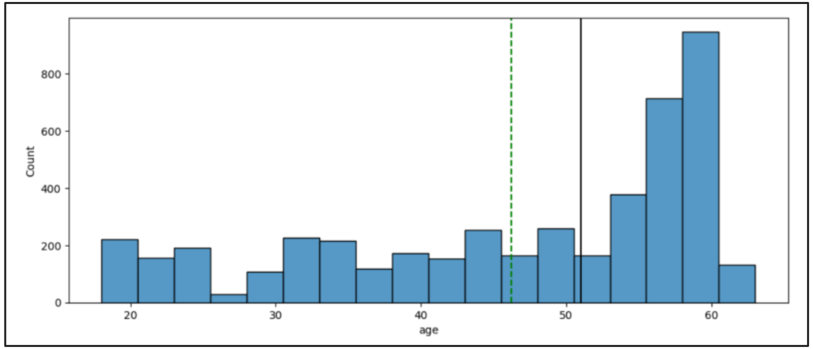
- ExtraaLearn is an online educational start-up company desiring to identify:
  1. Which leads are more likely to sign up and purchase the product (i.e. convert to paying customers)
  2. What factors are most important in driving the lead conversion process
  3. The profile of leads which are likely to convert to paying customers

- To approach this problem, the following steps will be performed:
  1. Become familiar with and explore the data
  2. Perform univariate and bivariate analyses to understand individual variables and their distributions as well as correlation between independent variables and the target variable
  3. Clean and prepare data, split data into training/testing sets
  4. Perform classification modeling
     a. Outcome variable is “status”, a binary variable indicating if customer converted or not
        i. “0” is a customer that DID NOT convert to a paying customer
        ii. “1” for those that DID convert to a paying customer

Data Overview

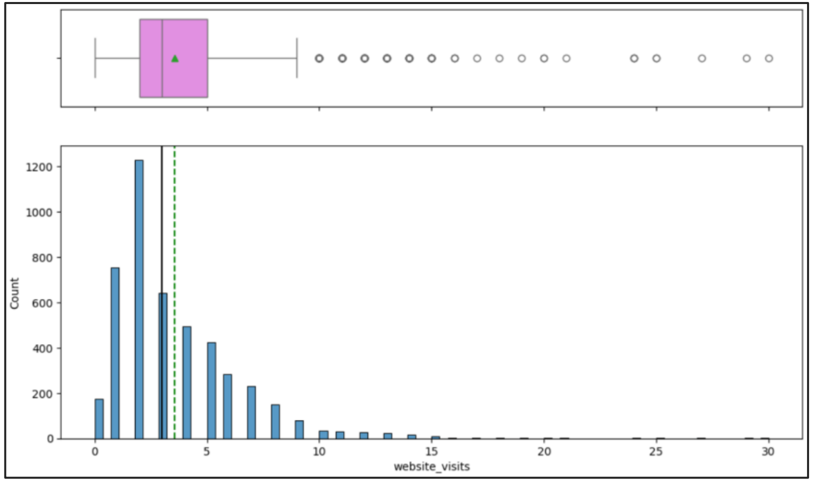
- Shape of the data set:
  - 4612 observations/rows
  - 15 columns/variables
- Variable types:
  - Object (10 variables): ID, current_occupation, first_interaction, profile_completed, last_activity, print_media_type1, print_media_type2, digital_media, educational_channels, referral
  - Integer (4 variables): age, website_visits, time_spent_on_website, status
  - Float (1 variable): page_views_per_visit
- There were no duplicate values or missing values (NA's) in this dataset
- Number of unique values for ID column: 4612 - this is the same number of observations in our data set, so each ID is unique

EDA Results - Numerical Variables



Age:

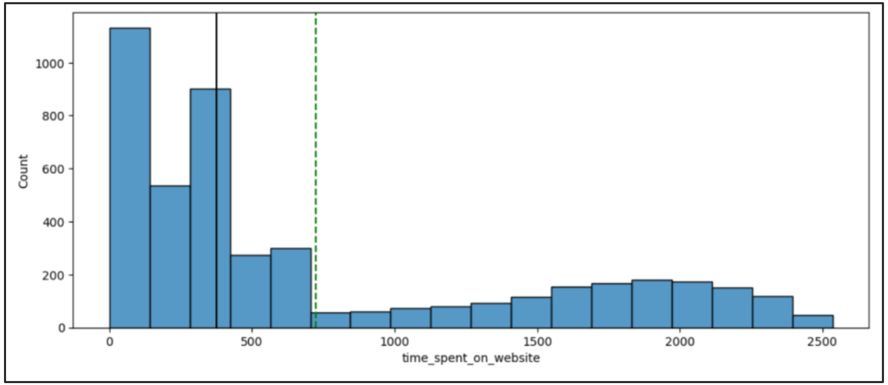
- Range: 20–62; Median ≈ 51, Mean = 45
- Left-skewed with more leads in 50s–60s
- No outliers



Website Visits:

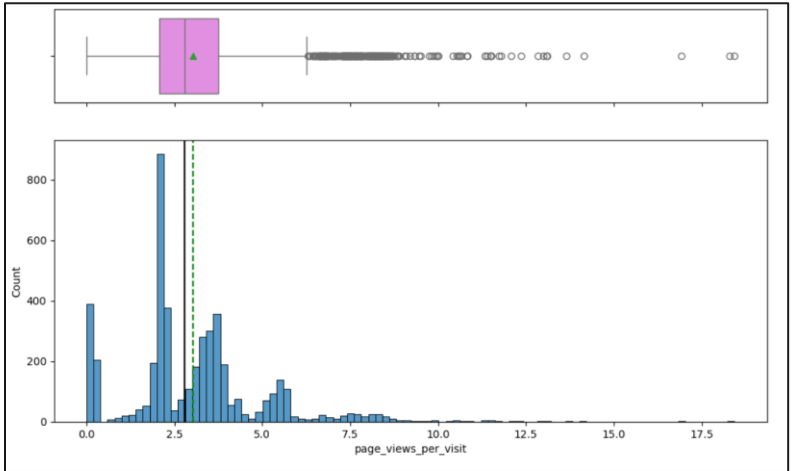
- Range: 0-30; Median ≈ 3, Mean ≈ 3.5
- Right-skewed with higher visit count of 2
- Outliers present from 10-30
- 174 leads have not visited the website

EDA Results - Numerical Variables



Time spent on website:

- Range: 0- 2500
- Mean(725) > median (400)
- Skewed right, no outliers observed



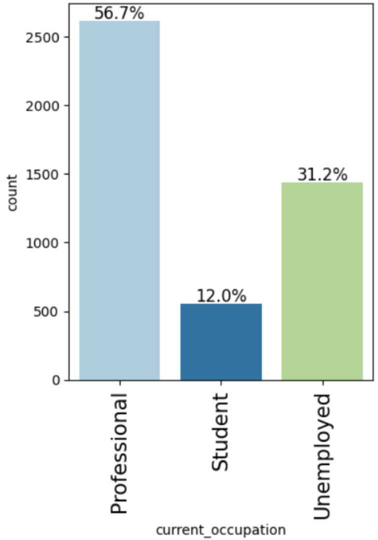
Page views per visit:

- Range: 0-18
- Mean/median similar ≈ 2.75-3
- Outliers extend from 6-18
- Majority of data points clustered below 7.5 with the distribution resembling a right skew with several spikes and dips

EDA Results - Categorical Variables

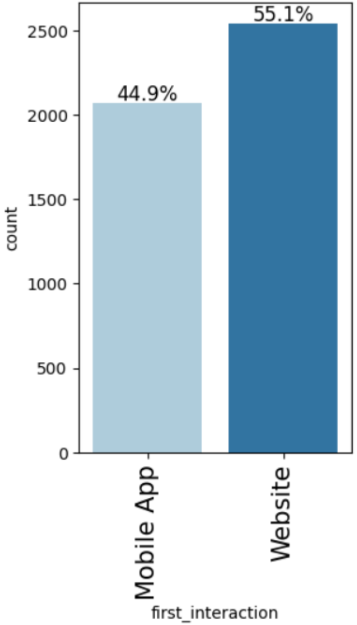
Current Occupation:

- The majority (56.7%) are professionals - learn new skills
- 31.2% unemployed - potential career changers



First Interaction:

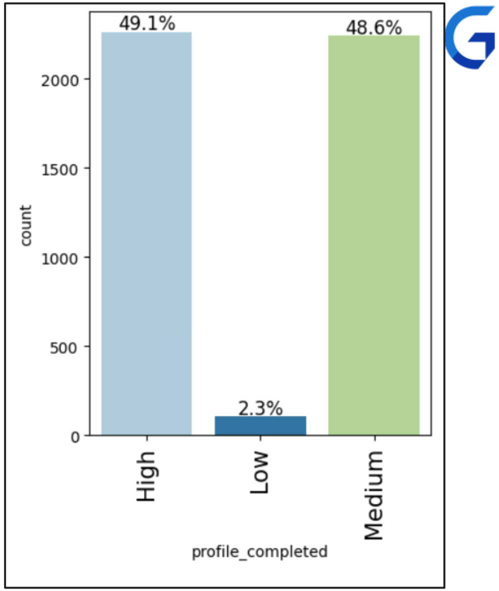
- 55.1% website
- 44.9% mobile app



EDA Results - Categorical Variables

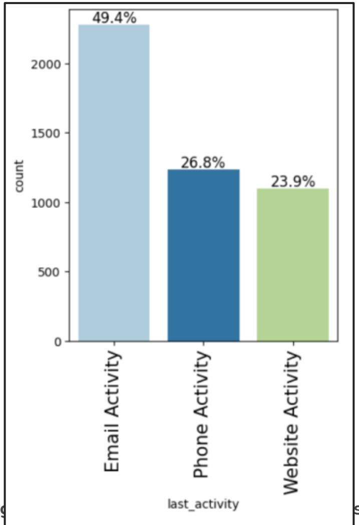
Profile Completed:

- High and medium completion comprises nearly all profiles, with only 2.3% in the low category



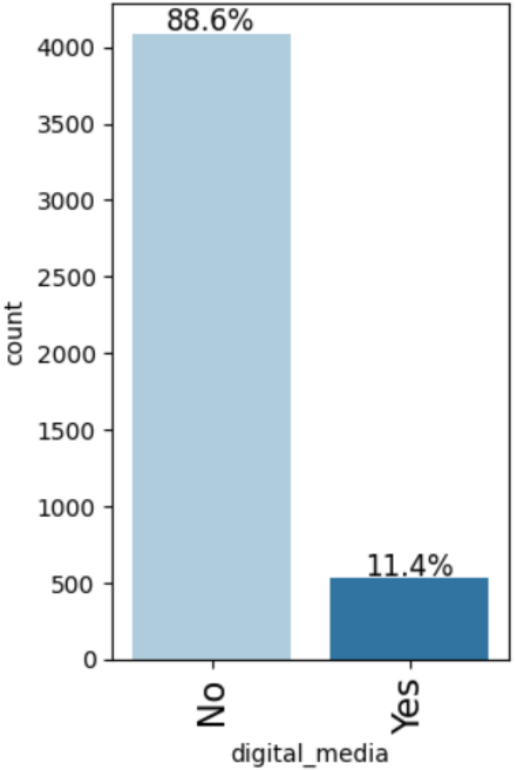
Last Activity:

- The highest percent of activity is email (49.4%) compared with roughly 25% for phone and website
- Since most people engage with email, the company needs to ensure this form is readily accessible and quick response is necessary



EDA Results - Categorical Variables

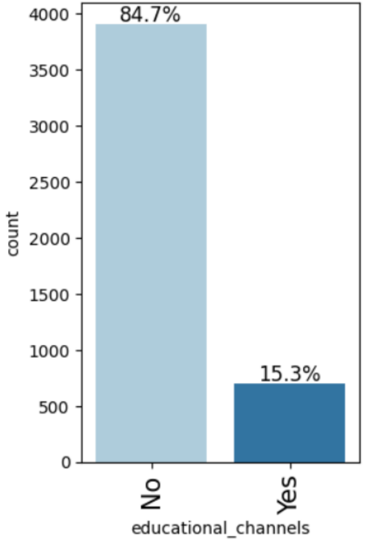
- Print media type 1 (newspaper):
  - No 89.2
  - Yes 10.8% - seeing newspaper ads
- Print media type 2(magazine):
  - No 94.9
  - Yes 5.1% seeing magazine ads
- Digital media (graph shown):
  - No 88.6%
  - Yes 11.4% seeing digital media ads
- More leads are seeing digital ads compared to other types of media, but most leads are not seeing ads.



EDA Results

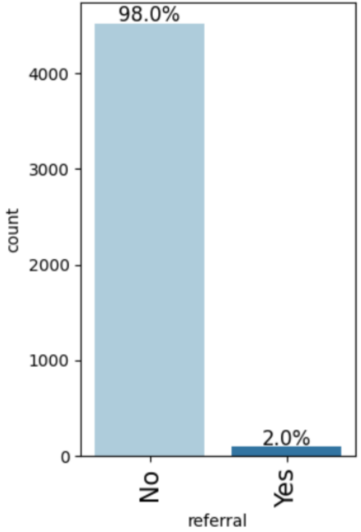
Educational channels:

- Yes 15.3%
- No 84.7%
- More leads are hearing about our service through educational channels than other media or referrals



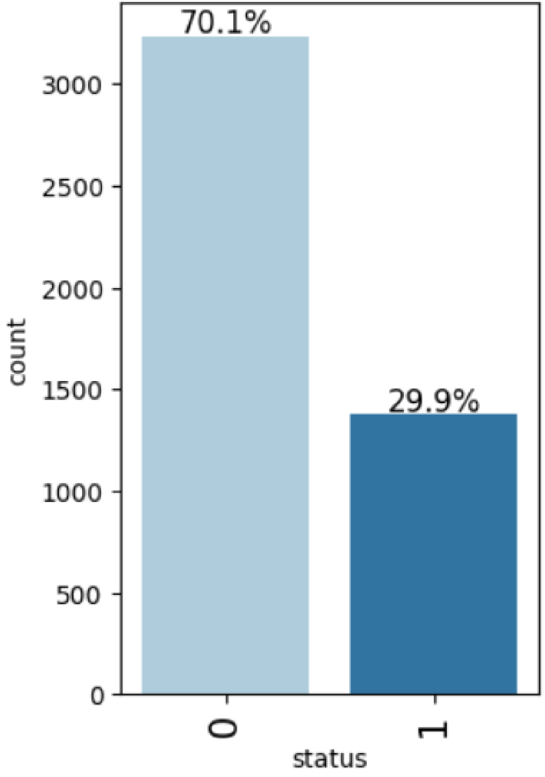
Referral:

- Yes 2%
- No 98%
- Very low number of referrals to our service, this would be helpful to increase



EDA Results - Dependent Variable “Status”

- 70.1% of leads did NOT convert into a paying customer
- 29.9% of leads DID convert into a customer
- There is much room to increase lead conversion to paying customers

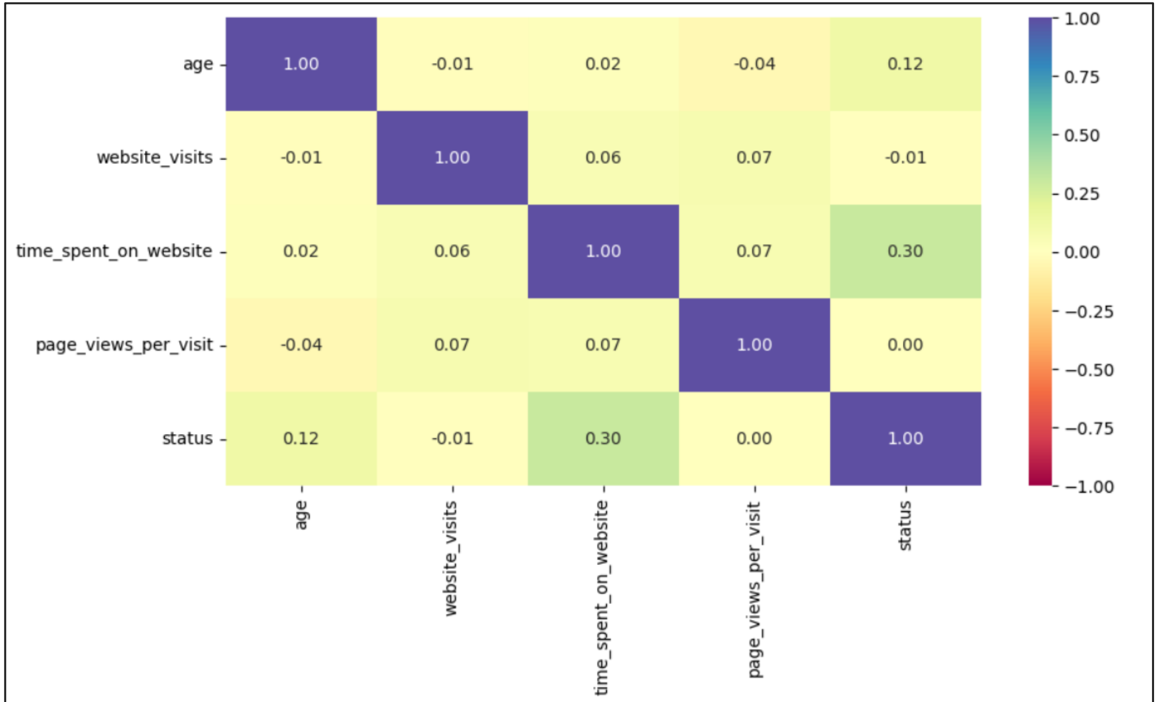


EDA Results - Univariate Key Observations

- Approximately 30% of individuals are converting to paid customers
- The majority of prospects are professionals followed by unemployed individuals
- The average age of the prospects is 45 years
- Only 2% of individuals are getting referred to the company
- Regarding seeing advertisements the highest percentage (15.3%) are seeing them on educational channels and the lowest (5.1%) seeing ads from magazines
- Approximately 50% of people reach out via email to inquire about the programs

EDA Results - Bivariate Analysis

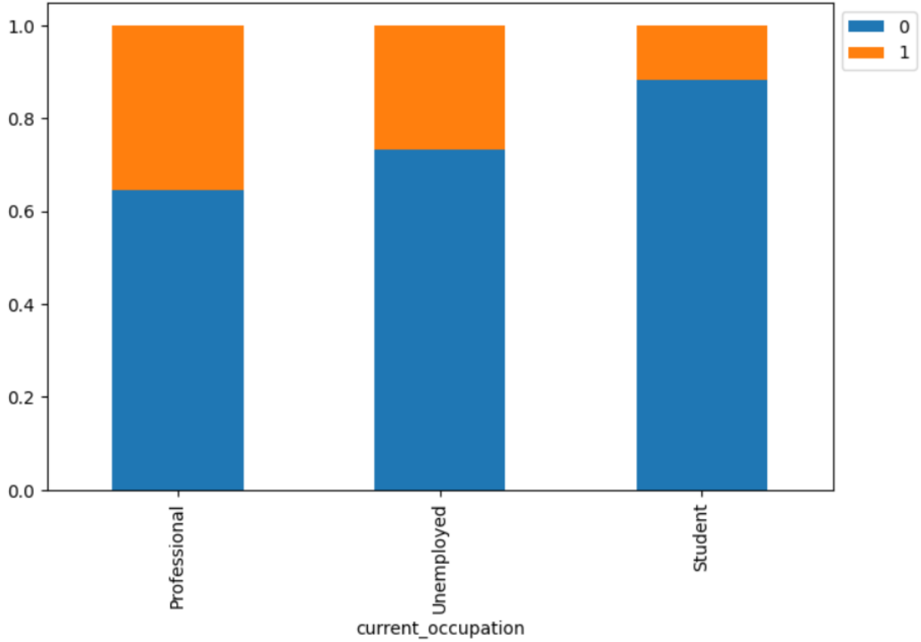
- Correlation heatmap shows minimal correlation between variables
  - The highest correlation of 0.30 was between status and time spent on website
  - Indicates low multicollinearity or correlation between independent variables



EDA Results - Bivariate Analysis

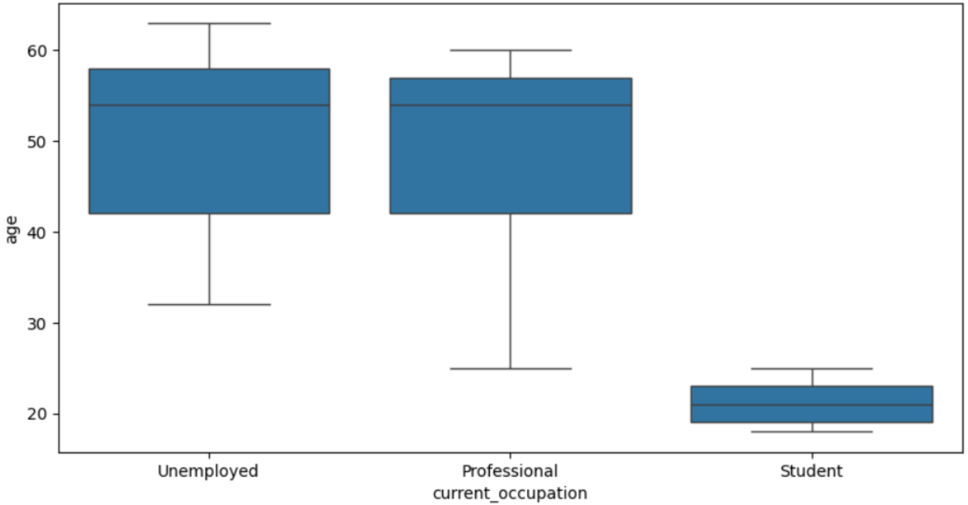
- Greater proportion of professional prospects convert to paying customers, followed by unemployed individuals.

- The lowest proportion of conversions are within the student group.



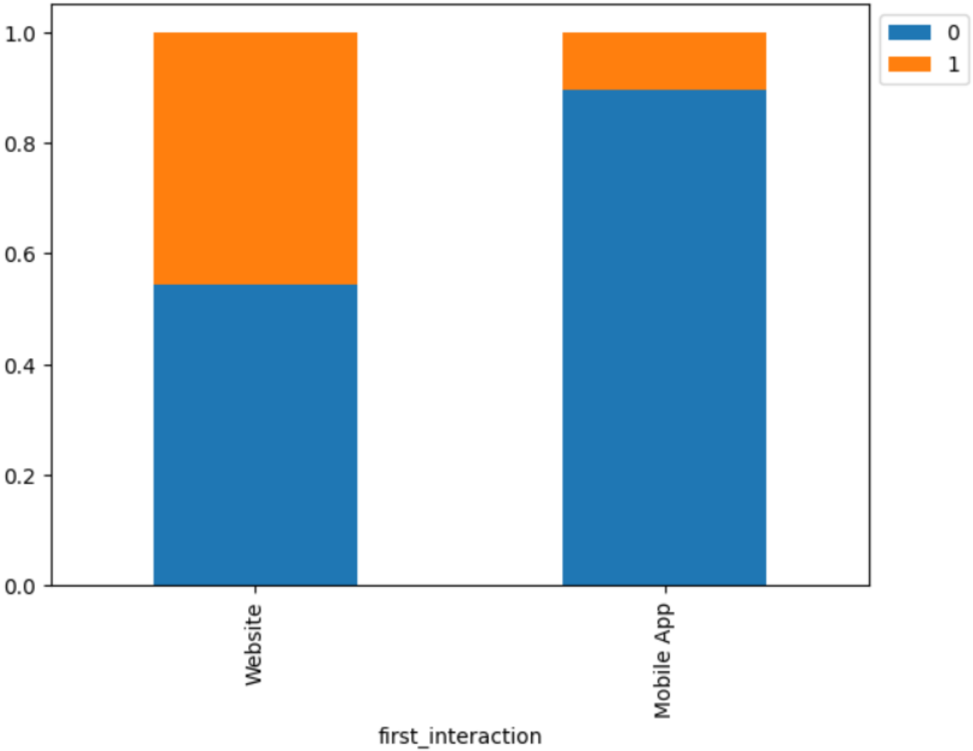
EDA Results - Bivariate Analysis

- The boxplots show distribution of age compared with occupational status
- Age of unemployed versus professional Prospects have similar profiles with median age approximately 54 years, though age range of professional is larger and younger.
- As one would expect, student age is much lower with median of 22 years



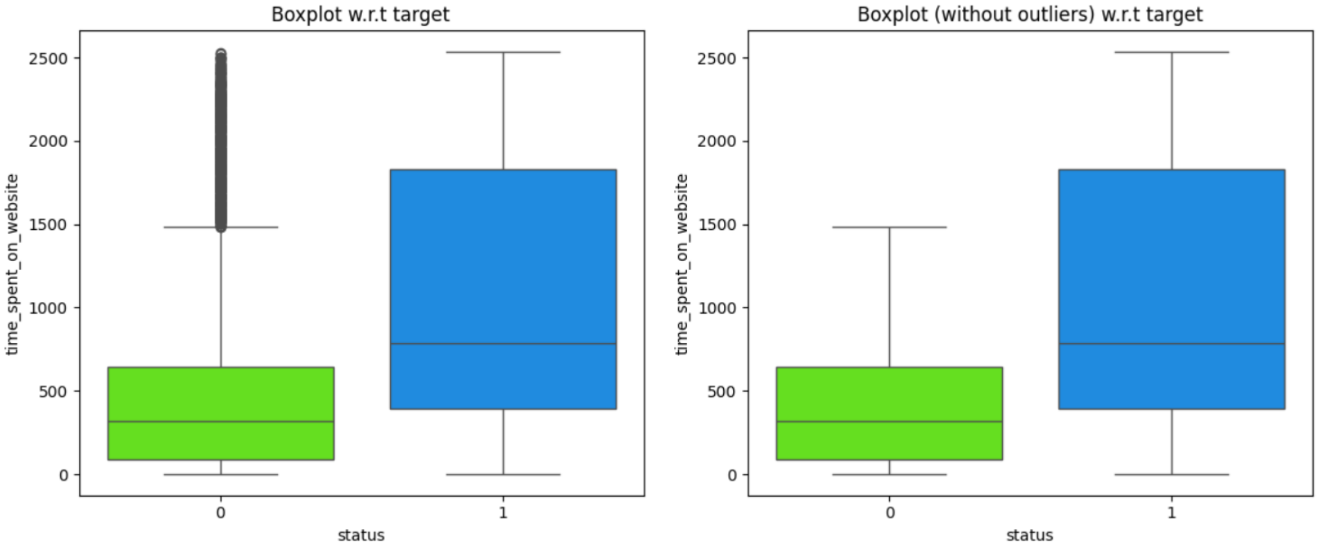
EDA Results - Bivariate Analysis

- This shows people who visit the website as their first interaction have a higher proportion of conversions to paying customers compared to if their first interaction is on the mobile app
  - 45% conversions with website as first interaction
  - 10% conversions with mobile app as first interaction



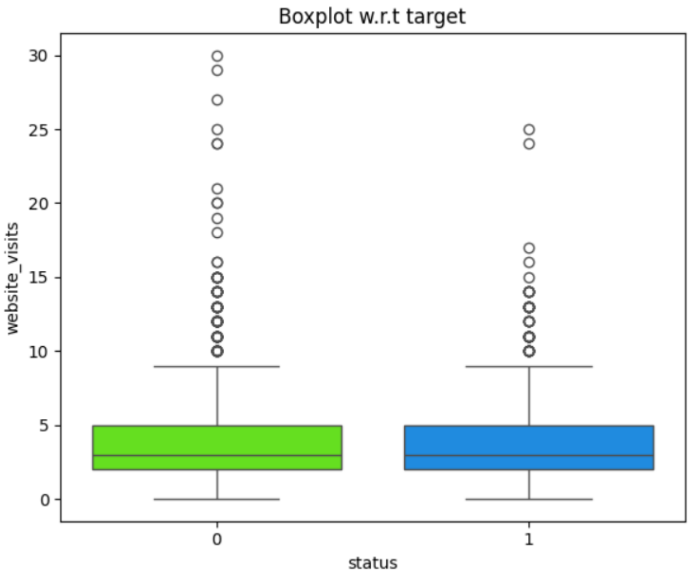
EDA Results - Bivariate Analysis

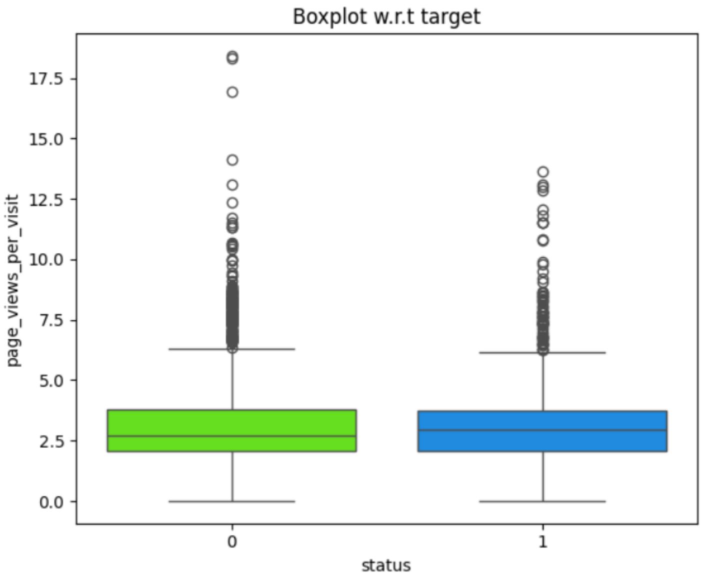
- Comparing time spent on the website between groups:
  - Converted prospective customers median time = 789
  - Not converted prospective customers median time = 317
    - Many outliers present
- This shows that people who became paying customers spent much more time on the website



EDA Results - Bivariate Analysis

- Conversely to amount of time spent on the website, plotting website visits and page views per visit yields similar results between groups
- This shows that both groups of customers (paying converted, non-paying) have similar characteristics in number of website visits and page views, but people who end up converting to paying customers spend much more time navigating the website



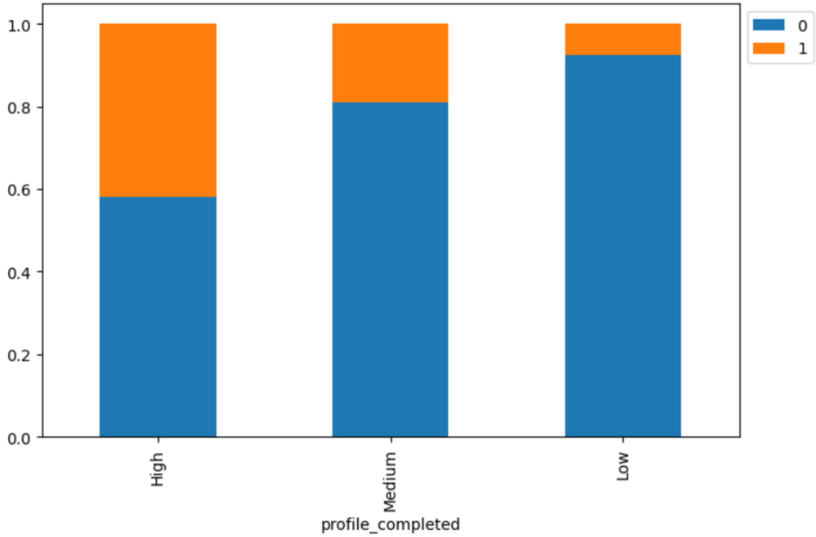


EDA Results - Bivariate Analysis

- The accompanying barplot shows level of profile completion has an impact on lead status:
  - Higher profile completion linked to higher proportion of paying customers (~42%)
  - Medium completion leads to approximately 20% of leads converting
  - Lower profile completion linked to smaller proportion of paying customers (<10%)

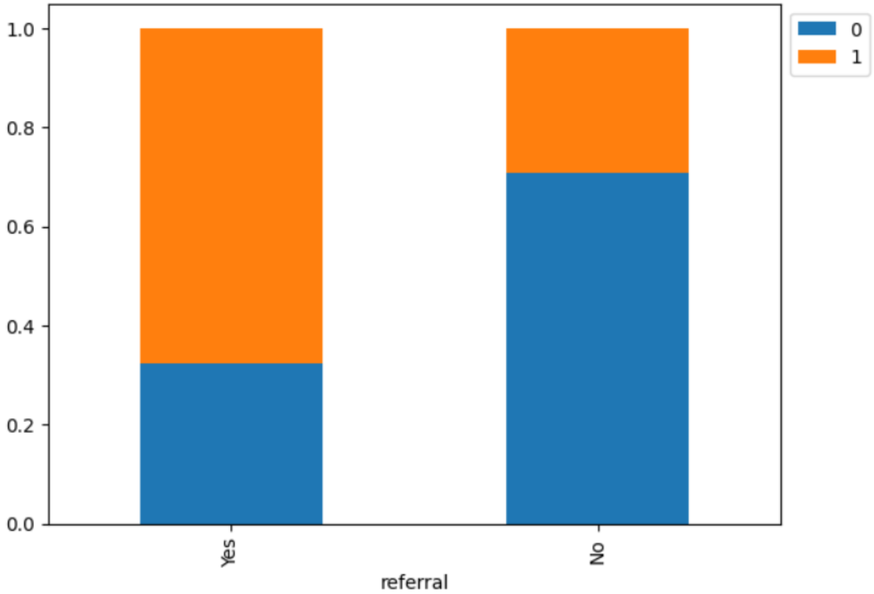
- Analysis between last activity and status:
  - Website activity has a higher conversion rate compared to email and phone

- Advertisement medium and status:
  - Similar between group of approximately 30% conversion across all groups of digital, educational, Newspaper, and magazine ads



EDA Results - Bivariate Analysis

- This graph shows proportion of people who were referred to ExtraaLearn or not and the impact on status
- Individuals who had a referral were more likely to convert to a paying customer
  - 65% converted if they had referral
  - 30% converted without a referral
- Recall that univariate analysis revealed that 2% of people were referred however the majority of them became paying customers - it may increase our paying customers if we can increase referrals



Data Preprocessing

- Outlier check of numeric variables:
  - No outliers in age or time spent on website
  - Outliers present in website visits and page views, however the values do not seem out of the ordinary or entered incorrectly and therefore were not be replaced or removed
- Created Y with the variable “status”
- Created X with the remaining variables and created dummies for the categorical variables
- Split the data into training/testing sets at 70/30 ratio
  - 3228 observations in training set
  - 1384 observations in testing set

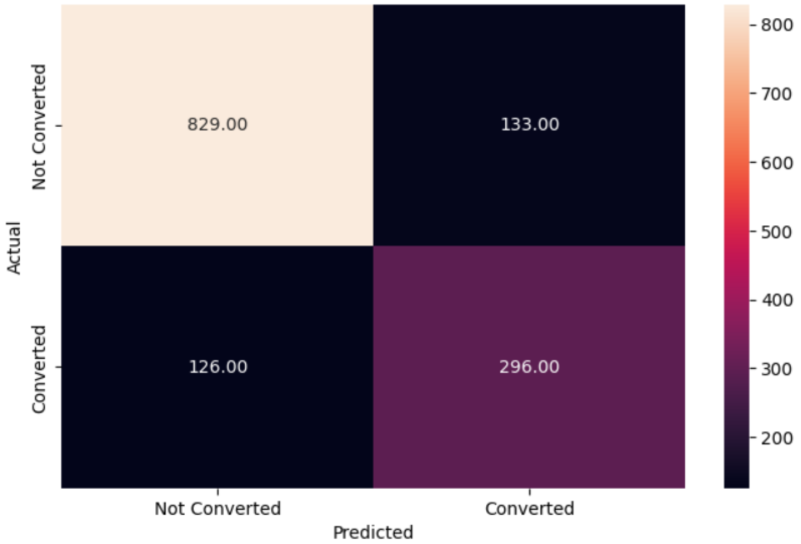
Model Building

- Built and ran the following models:
  - Decision Tree
  - Decision Tree with hyperparameter tuning
  - Random Forest
  - Random Forest with hyperparameter tuning

Decision Tree (DT) Results

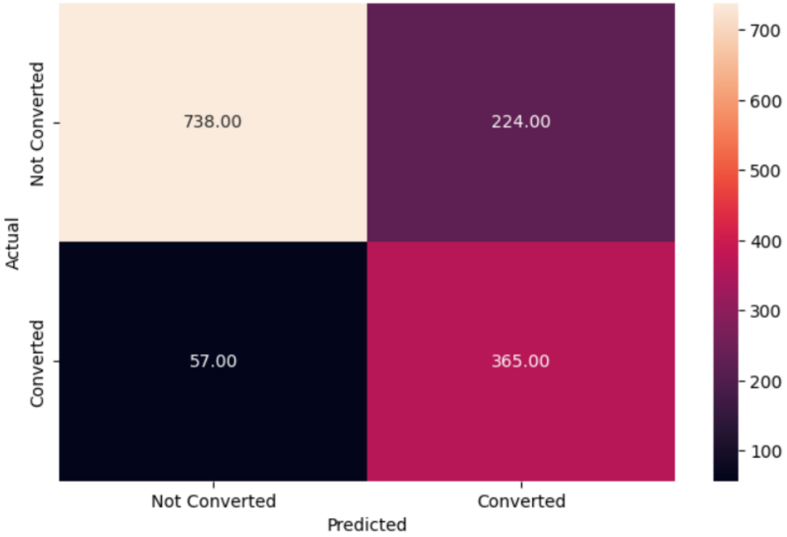
- Training data - Decision Tree confusion matrix:
  - Accuracy = 1.00
  - Precision = 1.00
  - Recall = 1.00
  - Decision tree is overfitting the training set and will not be useful in predicting testing data

- Testing data - Decision Tree confusion matrix (shown here):
  - Accuracy = 0.81
  - Precision
    - Group 0 = 0.87, Group 1 = 0.69
  - Recall
    - Group 0 = 0.86, Group 1 = 0.70
  - Model not fitting the testing data well - more accurate predicting group 0



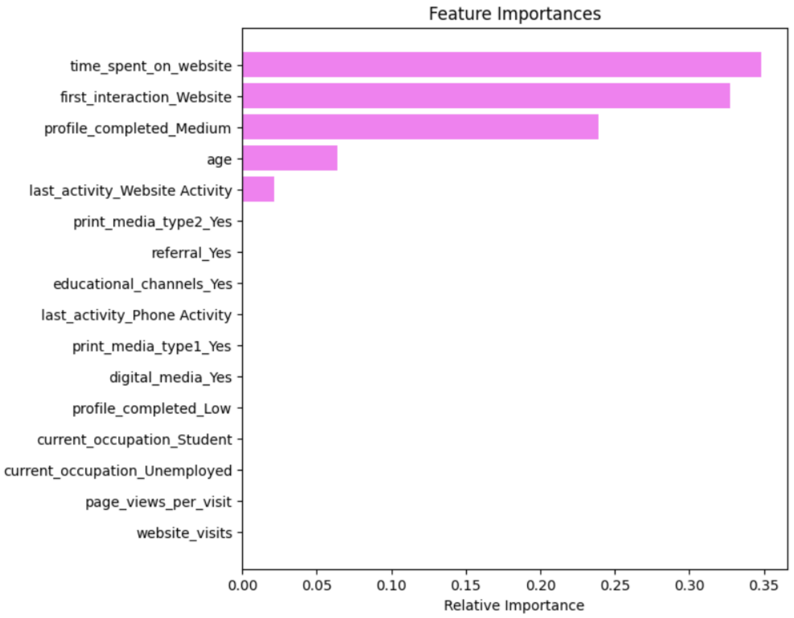
Decision Tree Results - After Hyperparameter Tuning

- Training data - DT confusion matrix:
  - Accuracy = 0.80
  - Precision:
    - Group 0 = 0.94, Group 1 =0.62
  - Recall
    - Group 0 = 0.77, Group 1 = 0.88
  - Much better than prior model - not overfitting
- Testing data - DT confusion matrix (shown):
  - Accuracy = 0.80
  - Precision:
    - Group 0 = 0.93, Group 1 = 0.62
  - Recall:
    - Group 0 = 0.77, Group 1 = 0.86
  - Tuning the model resulted in reduced overfitting
  - Good at predicting those who will not convert (93% of the time) - few false positives
  - Model correctly identified those who will convert 86% of the time



Decision Tree Tuned - Feature Importance

- Most important features of tuned decision tree (top 3):
  - Time spent on website - 34.8 % of total information gain - most important predictor in becoming a paid customer!
  - First interaction website - 32.7% of total information gain
  - Profile completed medium - 23.9% of total information gain



Random Forest (RF) & Random Forest Tuned Results

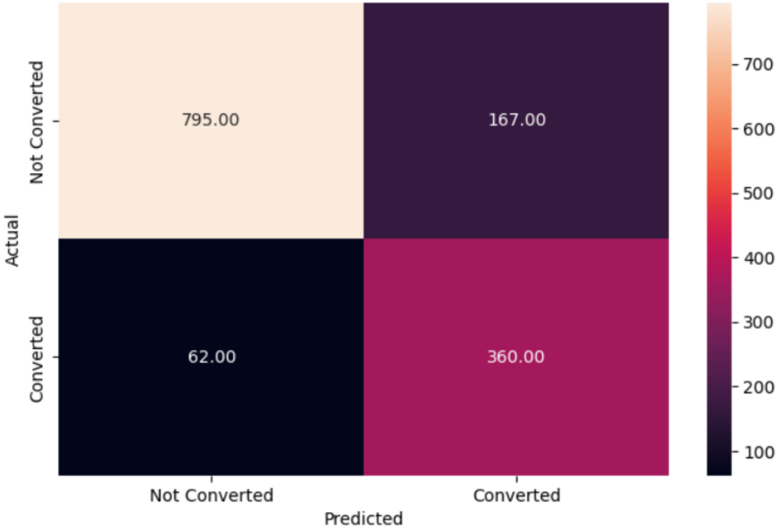
Random Forest:

- Overfitted model - not ideal for predicting test data

Random Forest Tuned (Confusion matrix shown):

- Training data:
  - Accuracy = 0.84
  - Precision: group 0 = 0.94, group 1= 0.68
  - Recall: group 0 = 0.83, group 1 = 0.87

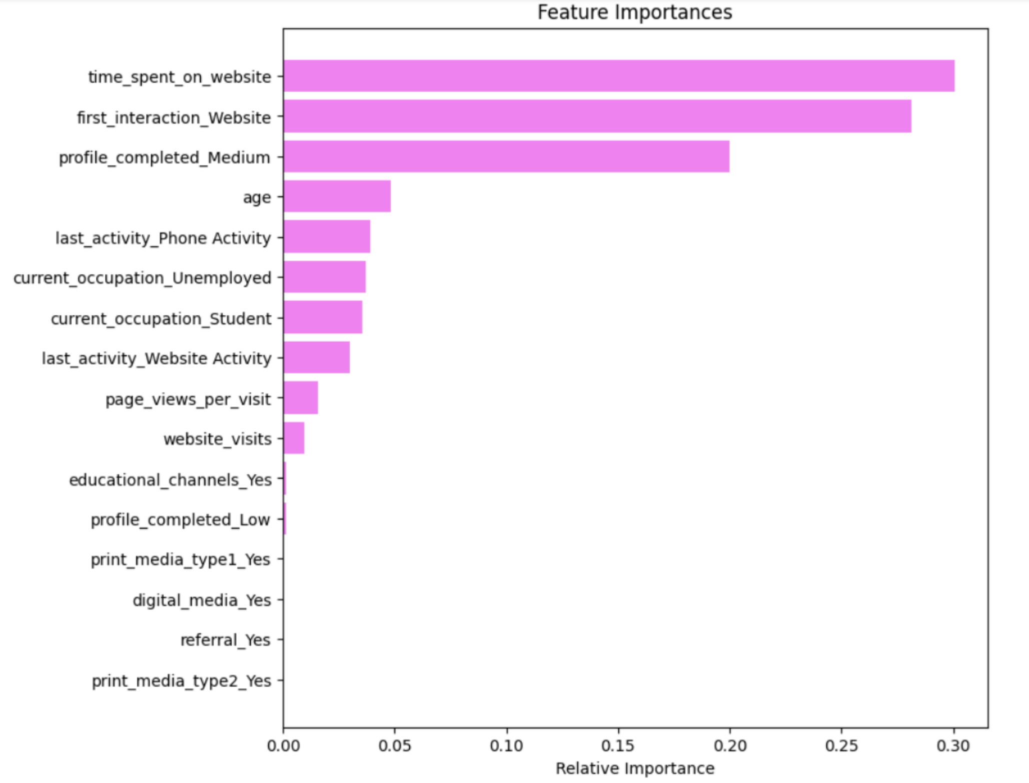
- Test data results:
  - Accuracy = 0.83
  - Precision: group 0 = 0.93, group 1= 0.68
  - Recall: group 0 = 0.83, group 1 = 0.85



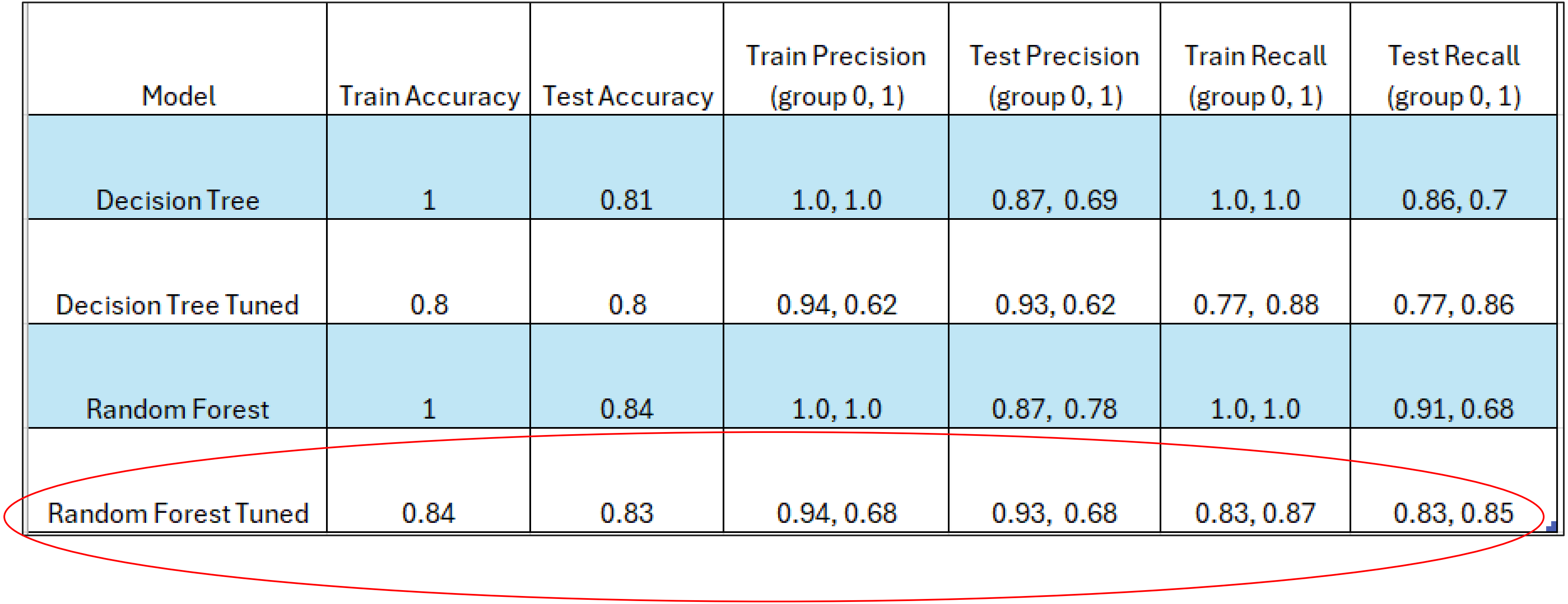
Random Forest Tuned - Important Features

- Top 5 important features:
    - Time spent on website (30%)
    - First interaction website (28%)
    - Profile completed medium (21%)
    - Age (5%)
    - Last activity phone activity (4%)

- Similar to the important featured identified with the decision tree model. Time spent on website was the most important factor followed closely by first interaction website. These variables provide the most information in predicting the status outcome.



Model Performance Summary Table



Model	Train Accuracy	Test Accuracy	Train Precision (group 0, 1)	Test Precision (group 0, 1)	Train Recall (group 0, 1)	Test Recall (group 0, 1)
Decision Tree	1	0.81	1.0, 1.0	0.87, 0.69	1.0, 1.0	0.86, 0.7
Decision Tree Tuned	0.8	0.8	0.94, 0.62	0.93, 0.62	0.77, 0.88	0.77, 0.86
Random Forest	1	0.84	1.0, 1.0	0.87, 0.78	1.0, 1.0	0.91, 0.68
Random Forest Tuned	0.84	0.83	0.94, 0.68	0.93, 0.68	0.83, 0.87	0.83, 0.85

Which Model Performed Best?

- Best model = Random Forest Tuned model
- Reasoning:
  - Decision Tree and Random Forest models both overfit the training data and therefore were not ideal to predict testing data
  - Our goal is to maximize recall (i.e. identifying false negatives, or a lead that *would have converted*, but the model predicted *they would not*). The RF tuned model had the best overall recall for both training and testing data in determining placement into group 0 or group 1
  - Though accuracy is not the best measure due to imbalanced data in group 0 versus group 1, accuracy is higher in the Random Forest tuned model compared to Decision Tree tuned model

Conclusions & Recommendations

- Profile of a lead more likely to convert to a paying customer:
    - Leads who visit the website as their first interaction and spend significant time navigating the website
    - Professionals or unemployed individuals, in their 40’s and 50’s
    - Leads who have medium to high profile completion

Conclusions & Recommendations

1. The most important factor in determining whether a lead will convert is the amount of time spent on the website, followed by first interaction being via website. This exemplifies the importance of:
    a. Ensuring the website is user friendly, up-to-date, engaging, and informative
    b. Evaluating the mobile app to improve the ease of use or appeal of our programs. This may help to increase the proportion of people accessing our website via mobile app to convert to paid customers
    c. Allocating resources to identify these individuals early
    d. Potentially provide additional incentive to sign up such as a free one-week trial
2. Profile completion of ‘medium’, age, and last activity phone activity were also important predictors of converting to paying customers
    a. Target leads age 40’s and 50’s as they are more likely to become customers
    b. Encourage leads with partial profile completions to complete their profiles, as it may tip them into converting
    c. Identify if there are consistent challenges to completing profiles
3. The variable ‘referral’ did not appear as an important factor potentially because of the small number of leads who were referred, however of the 2% of people that had a referral, 65% converted to a paying customer. This may be an area to gain leads and conversions. Offering incentives or additional programs for referring others to our company.

APPENDIX

Great Learning

Happy Learning !