Customer Prediction with ExtraaLearn

Elective Project: Practical Data Science

MIT-PE ADSP Sep23

11/24/23

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:**

  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 regularly, one of the issues faced by ExtraaLearn is to identify which of the leads are more likely to convert so that they can allocate resources accordingly.

- **Solution approach and methodology:**

  - Analyze and build an ML model to help identify which leads are more likely to convert to paid customers,
  - Find the factors driving the lead conversion process
  - Create a profile of the leads which are likely to convert

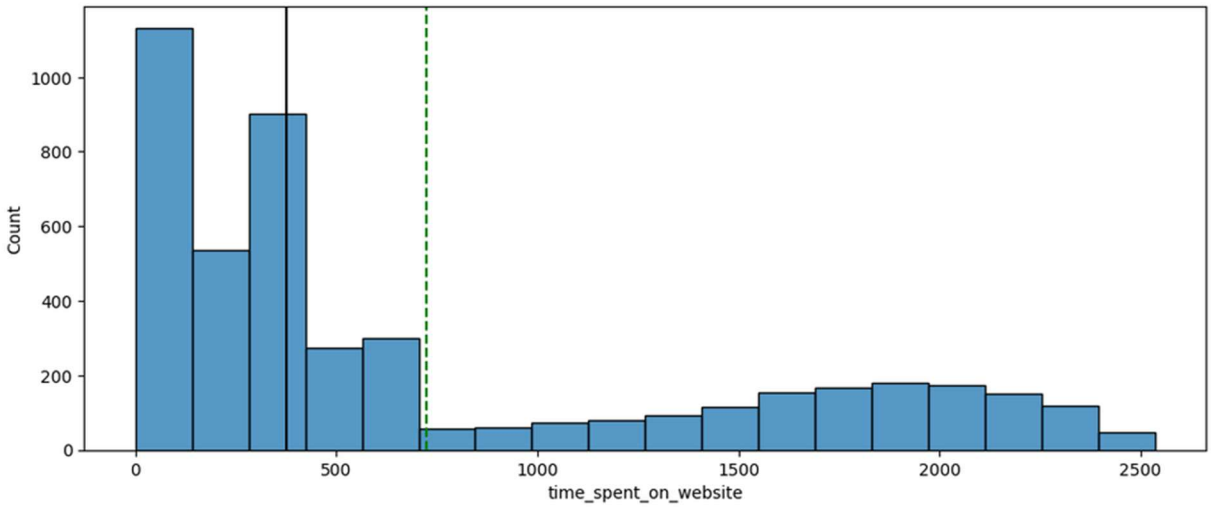
Data Overview

- Data consists of 4612 rows and 15 columns
- No duplicate and no null data
- Integers, objects and floats

EDA Results

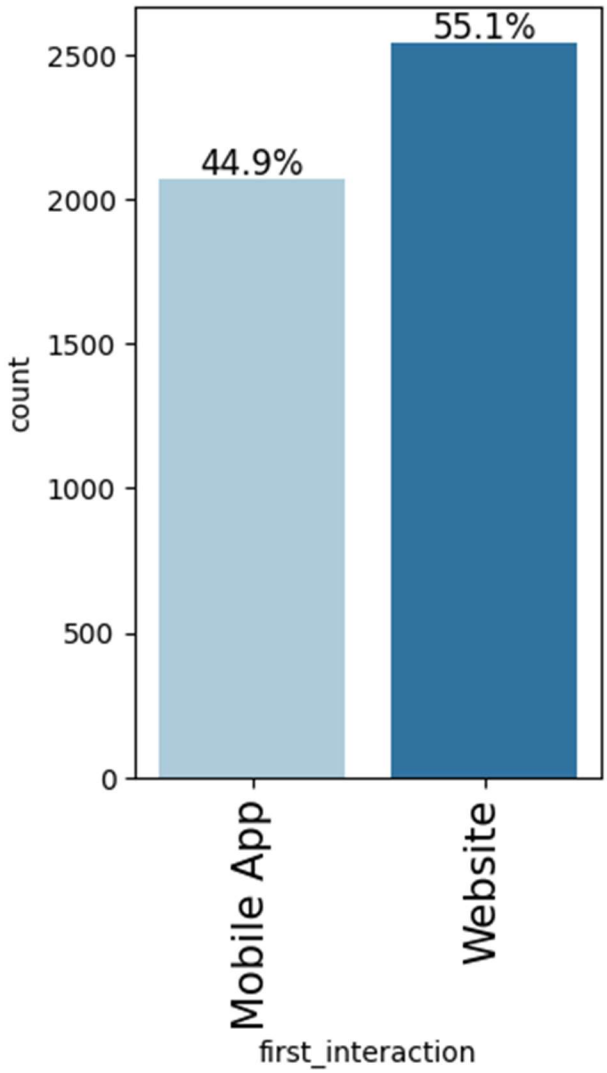
Exploratory Data Analysis

- A considerable number of leads are centered around 53-61 years old and the least number is around 26-31 years old.
- Majority of Leads usually spent their time on the website no more than 700, while the minority of them have spent either time 700 to 950 or 2355 to 2550.



- Average number of pages on the website viewed by leads during their visits is at around 2.5.
- The number of times a lead has visited the website are usually around 0-6 times with significant amount centered around 2 visits.
- Majority of leads are from professional segment at 57% with 12% students and 31% unemployed.
- ExtraaLearn ads generated 11% leads from digital platforms, 11% of leads came from Newspaper and Magazine generated 5% of leads. 73% comes from an unknown source.
- Only 15% of leads came from educational channels which leaves a lot of room for improvement

- Only 2% of leads came from referral. Referral is seen as a significant area of opportunity for the company.
- 30% out of 4612 leads resulted in a sale.
- First interaction mobile vs website is somewhat evenly split:



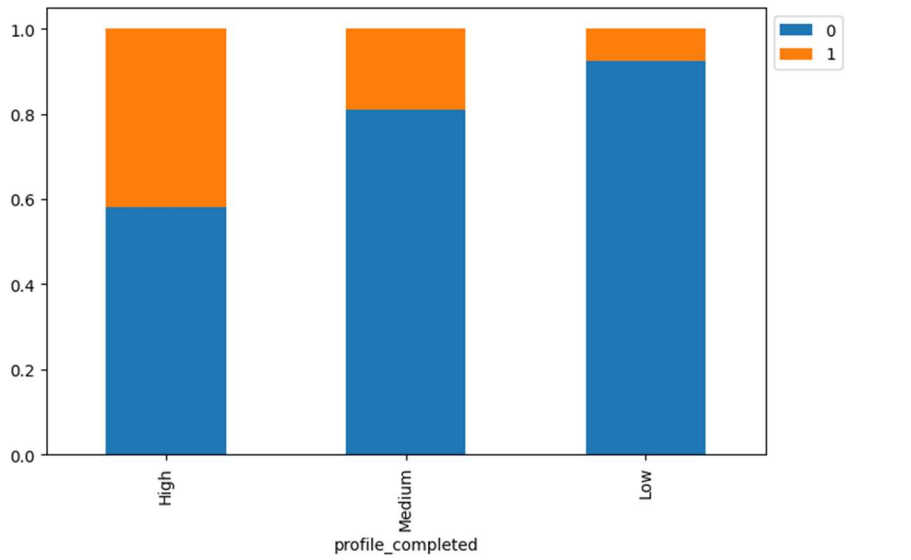
Bivariate Analysis

- Bivariate Analysis showed the strongest correlation between time spent on the website and status, and age and status. The worst is between age and page views per visit , and website visits and age.
- The highest success rate of converted leads are from professional occupation followed by unemployed and a small amount lead came

from students. Majority of students are not interested in what ExtraaLearn has to offer while a primary market seems to be Professionals in their 50s and are more likely to sign up for a course.

- This shows that the currently offered courses are more oriented toward working professionals or unemployed individuals. The courses offered might be suitable for the working professionals who might want to transition to a new role or take up more responsibility in their current role. Extraalearn may not be as appealing to students as to professionals or unemployed due to lack of degree designation or lack of work/on the job experience to apply new certifications earned thru ExtraaLearn programs.
- Website shows a better pull thru rate than Mobile app.
- Bivariate analysis also shows higher success rate in getting a lead to sign up for a course when a lead has a high profile completion level.

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



- Out of all of those contacted, those that were contacted via website showed the most success in bringing leads on board at 38% signing

for the course, with phone activity showing the lowest pull thru or success rate of 21%.

- Out of all of those leads that were referred, 68% ended up signing up for a course.

Outlier Check

- There is a multiple number of outliers for website visits from 10 to 30.
- There is also outliers for page views per visit from 7.5 to 17.5. Age and time spent on the website data has more relevance to how likely leads will turn into customers.

Model Building and Performance Summary

Decision Tree:

- Model performance on a training set shows 94% overall accuracy. 96% precision for non-conversion class and 90% precision for conversion. Recall for non-conversion is 96% and 90% for conversion which means that the model correctly identifies 90% of the actual instances of conversion into a paid subscriber of ExtraaLearn program. F1-scores are .96 for non-conversion and .90 for conversion. Model seems to perform well, with high precision, recall, and F1-score for both classes.
- Model performance on a test data set shows 81% accuracy. Model predicts non-conversion better than conversion at 86% vs 68% for all precision, recall, f1-score.
- The Decision Tree works well on the training data but not so well on the test data as the recall is 68% in comparison to 90% for the training dataset for conversion rate, i.e., the Decision Tree is overfitting the training data.

Decision Tree – Hyperparameter Tuning

- Hyperparameter tuning with flipped weights for training set produced 100% accuracy in both classes and a 100% precision, recall, f1-score for both classes.
- Test data set didn't fit the model as perfect as the training set and produced 81% accuracy with precision, recall, f1-score figures matching test data results of the decision tree completed without hyperparameter tuning.

Visualizing the decision tree

- Leads who first interacted on website, spent their time on website less than 415.5 and age under 25 are likely to be converted to paid customers.
- Leads who spent time on website less than 419.5 were usually not converted to paid members.
- Another interesting point is even though the leads interacted on website in their first time, but they spent their time less than 415.5 and did not fully complete their profile (50%-75%), they had a greater entropy to convert to a paid subscriber of Extraalearn.
- In conclusion, interaction on website in the first time and duration of time spent on website are likely to be crucial factors to convert leads to paid customers.

Feature Importance - Decision Tree (*see appendix for a chart*)

The feature importance plot for the base model and tuned model are quite similar. The model seems to suggest that time spent on the web, first interaction thru the website, and profile completed at medium level are the most important features. The rest of the variables have no impact on deciding whether a lead will be converted or not.

Random Forest Classifier

- The performance of a training data set on a random forest model showed 100% accuracy, precision, recall, f1-score.
- Performance of a test set on a random forest model showed 85% accuracy, 80% precision, 69% recall, and 74% f1-score for a conversion class. Random forest classifier is definitely overfitting the training data set.

Random Forest Classifier – Hyperparameter Tuning

- The performance of a training data set on a random forest classifier model with hyperparameter tuning showed 85% accuracy, 70% precision, 86% recall, 77% f1-score for conversion.
- The performance of a test data set on a random forest classifier model with hyperparameter tuning showed 85% accuracy, 71% precision, 85% recall, 77% f1-score for conversion.
- Random Forest classifier with hyperparameter tuning shows lower predictability then previous models however training set and test set produced same precision, recall, f1-score and accuracy which means the model is fitting the training data too closely capturing noise or outliers that don't generalize well to new, unseen data.

Feature Importance - Random Forest (*see appendix for a chart***)**

- Similar to the decision tree model, time spent on website and first interaction website are the top two features that help distinguish between not converted and converted leads.
- Unlike the decision tree, the random forest gives greater degree of importance to profile completed High then Medium. Random Forest also gives greater importance to first interaction Mobile App, current occupation professional, last activity phone activity, age last activity website activity, current occupation student. This implies that the random forest gives importance to more factors in comparison to the decision tree model.

Conclusion

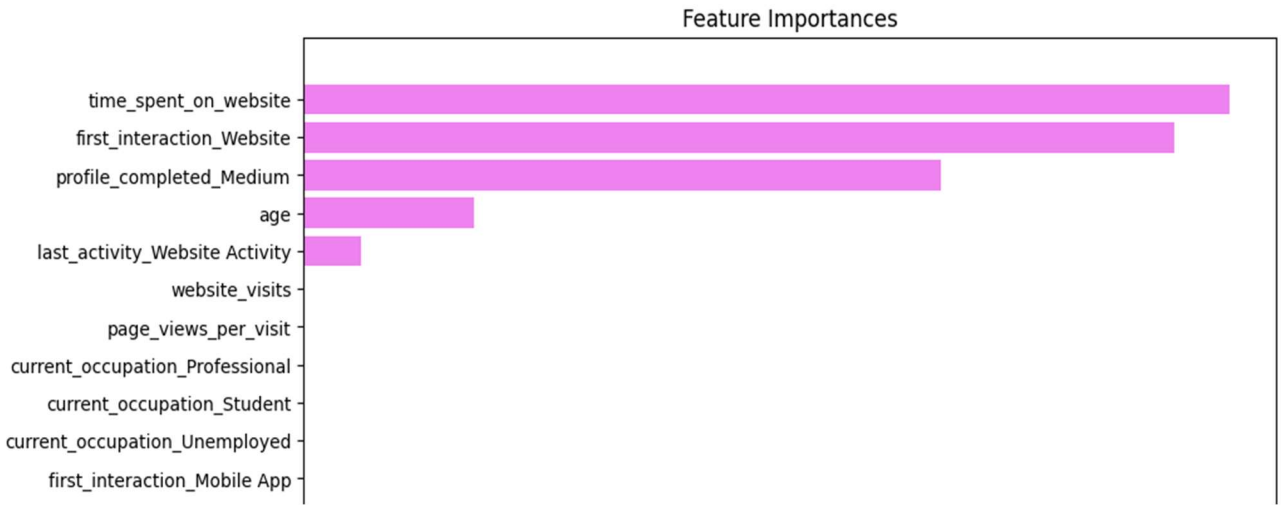
- Interaction on ExtraaLearn website for the first time and duration of time spent on website are likely to be crucial factors to convert leads to paid customers.
- Both models used in learning models showed insignificance to use of ads whether it be newspaper, magazine, or digital media.
- The agreement between the two models on the two parameters that are most important to conversion of leads into paid customers suggest some improvements to the business.

Business Improvement Recommendations:

- ExtraaLearn has to focus its attention and resources on improving it's website in order to encourage leads to interact on website. First time website impression seems to have a great impact of the leads. More updated website will attract student population and help leads stay and interact with the website more efficiently, hence generate more revenue.
- ExtraaLearn's focus on improving incentive system that would promote profile completion, would improve company's chances of bringing more leads into paid subscribers space.
- Spending money on ads doesn't seem to bring any additional value in bringing paid customers on board. That can be due to outdated nature of Extraalearn website and mobile app. Ads remain important part of the business however, as the models showed, can be inefficient without interactive website and mobile app.

Appendix

- **Feature Importance - Decision Tree**



- **Feature Importance - Random Forest**