ExtraaLearn Project

Classification and Hypothesis Testing Course

Nov 2024

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

- ExtraaLearn, an early-stage EdTech startup, is experiencing challenges in efficiently managing and converting the large number of leads they generate.
- They need to identify the leads that will most likely get converted into a customer and want to understand the drivers for customer conversion

Business Solution Approach

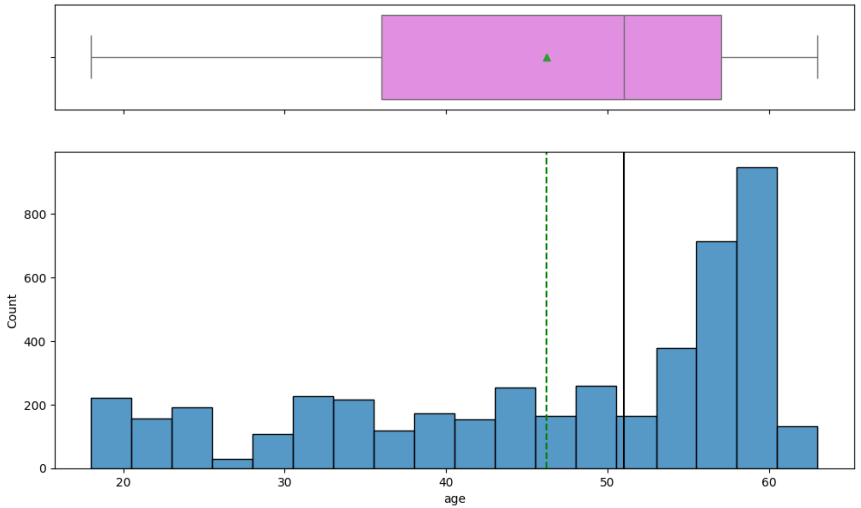
- The solution approach:
  - Understand the data using EDA
    - Univariate and Bivariate Analysis
    - Outlier Detection
    - Data Preprocessing
  - Model Building and Evaluation
    - Model selection: Decision Tree and Random Forest classifiers
    - Model Training and Performance evaluation
    - Hyperparameter tuning
    - Model Comparison
  - Feature Importance
  - Conclusion and Recommendation

Data Overview

- ID: Unique identifier for each lead.
- age: Age of the lead.
- current_occupation: Current occupation of the lead (e.g., 'Professional', 'Unemployed', 'Student').
- first_interaction: How the lead first interacted with ExtraaLearn (e.g., 'Website', 'Mobile App').
- profile_completed: Percentage of the profile completed by the lead (e.g., 'Low', 'Medium', 'High').
- website_visits: Number of times the lead visited the website.
- time_spent_on_website: Total time spent by the lead on the website.
- page_views_per_visit: Average number of pages viewed per website visit.
- last_activity: The lead's most recent interaction with ExtraaLearn (e.g., 'Email Activity', 'Phone Activity', 'Website Activity').
- print_media_type1: Whether the lead saw an ExtraaLearn ad in a newspaper (Yes/No).
- print_media_type2: Whether the lead saw an ExtraaLearn ad in a magazine (Yes/No).
- digital_media: Whether the lead saw an ExtraaLearn ad on digital platforms (Yes/No).
- educational_channels: Whether the lead heard about ExtraaLearn through educational channels (Yes/No).
- referral: Whether the lead heard about ExtraaLearn through a referral (Yes/No).
- status: Whether the lead converted into a paying customer (1 for Yes, 0 for No).

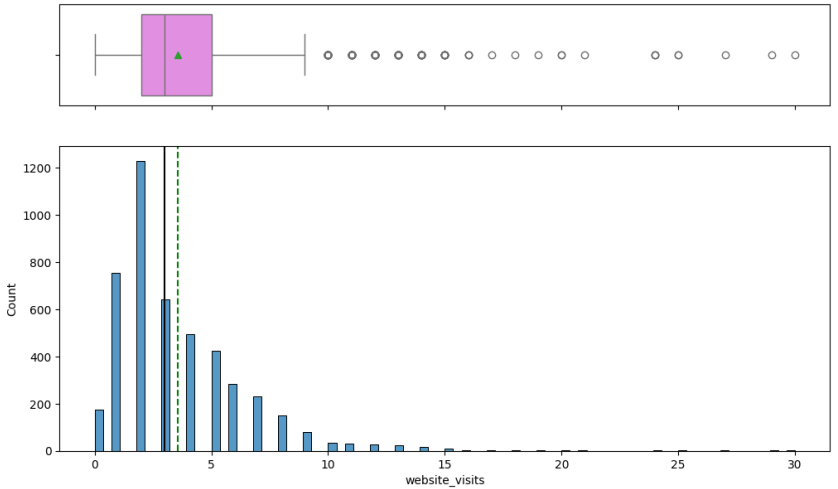
EDA Results: Age

- We can observe that they data is skewed to the left with the majority leads being of a higher age.
- Median Age is just over 51 and Mean is 46



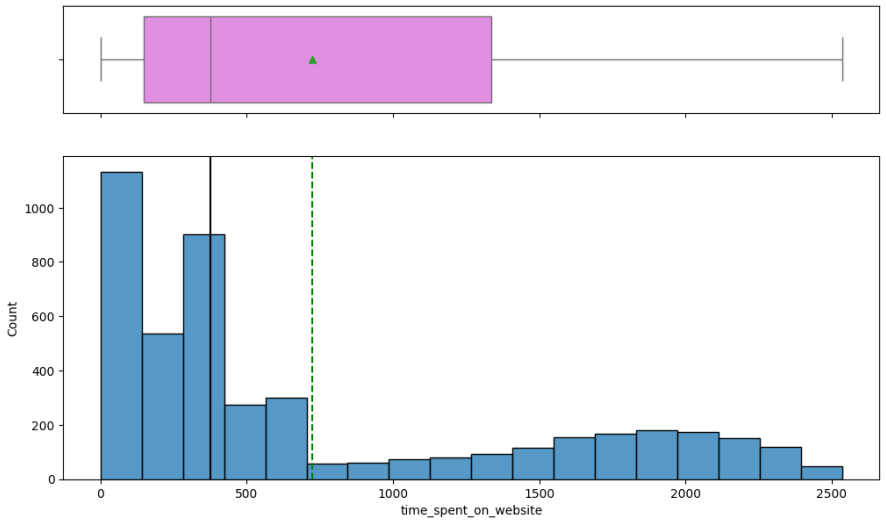
EDA Results: Website Visits

- We can observe that they data is skewed to the right with the majority leads visiting the website 2 times. There are a lot of outliers in the data as well.
- Median website visits is just over 3 and Mean is 3.5



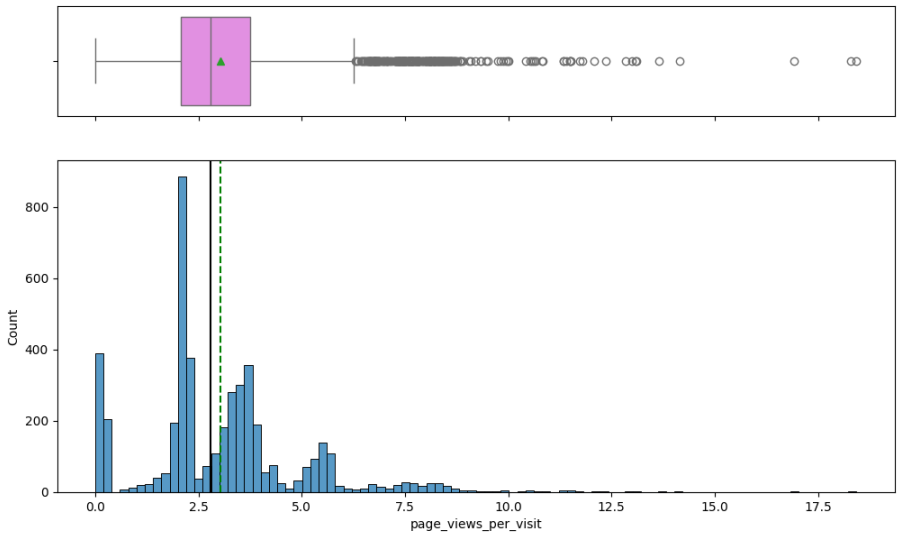
EDA Results : Time spent on Website

- We can observe that they data is skewed to the right with the majority leads spend less time on the website. This may be due to the population have lower conversion rate
- Median time on the website is just over 376 seconds and Mean is 724 seconds.



EDA Results : Page views per Visit

- We can observe that they data is skewed to the right with the majority leads view 2-3 pages per visit. There are a lot of outliers in the data as well.
- Median website visits is just over 2.79 and Mean is 3



EDA Results

- Based on the univariate analysis, the majority of leads are professional or are unemployed individuals who are looking for upskill or reskill themselves.
- Leads first interact with ExtraaLearn through Website more often than Mobile App.
- Leads generally have complete or almost complete profiles.
- As there are a lot of professional, we see a higher number of Leads interacting via the Email.
- Across Digital, Print, Educational channels and Referrals, we see that leads had more impact via education channels like online forums, discussion threads, educational websites, etc.
- Lead to Customer conversion rate is 30%.

Model Building

- Decision Tree
  - **Training Performance:** The model achieved perfect accuracy and recall on the training data. This indicates overfitting, where the model has memorized the training data and may not generalize well to unseen data.
  - **Testing Performance:** The model's performance dropped significantly on the testing data, with lower accuracy and recall scores. This confirms the overfitting issue.
  - **Recall:** The recall score on the test data was around 0.70, meaning the model correctly identified 70% of the leads who actually converted.

Model Building

- Decision Tree – Hyper tuning
  - **Tuning**: The model was tuned using GridSearchCV to find optimal hyperparameters, primarily focusing on improving recall. This involved adjusting parameters like max_depth, criterion, and min_samples_leaf.
  - **Training Performance**: The tuned model still performed well on the training data, but with slightly lower accuracy(0.8) and recall(0.88) compared to the untuned model, which is expected.
  - **Testing Performance**: The tuned model showed significant improvement on the testing data compared to the untuned model. The recall score increased to around 0.86 which is a substantial boost.
  - **Accuracy**: The accuracy also improved slightly, indicating better overall performance.

Model Building

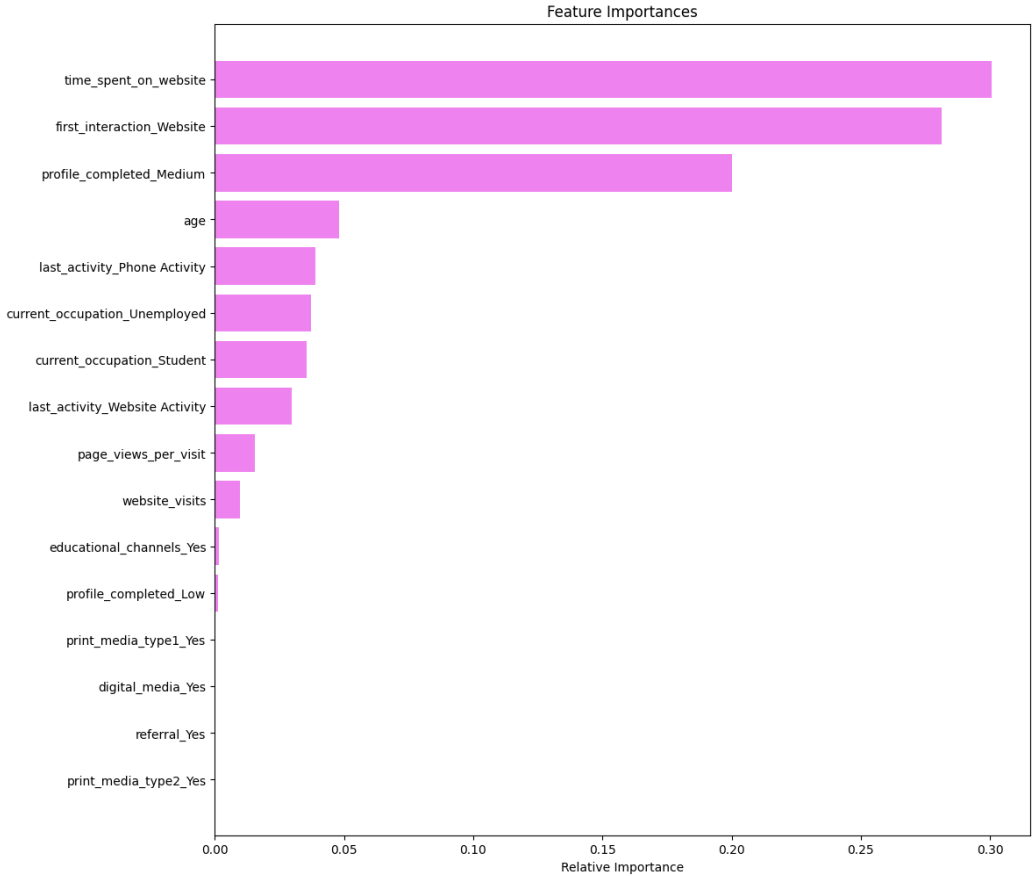
- Random Forest
  - **Training Performance:** Similar to the initial Decision Tree, the untuned Random Forest achieved perfect accuracy and recall on the training data, indicating overfitting.
  - **Testing Performance:** While it showed good accuracy(0.85) on the testing data, the recall score (0.69) was relatively lower compared to the tuned Decision Tree.

Model Building

- Random Forest – Hyper tuning
  - **Training Performance:** After tuning, the model's performance on the training data decreased slightly, indicating a reduction in overfitting.
  - **Testing Performance:** The tuned Random Forest showed significant improvement in recall (0.87) and accuracy(0.84) on the testing data, outperforming the tuned Decision Tree and the untuned Random Forest.
  - **Recall and Accuracy:** The model achieved a higher recall score, meaning it was able to identify a greater proportion of the leads who actually converted. The accuracy also improved, indicating better overall predictive ability.

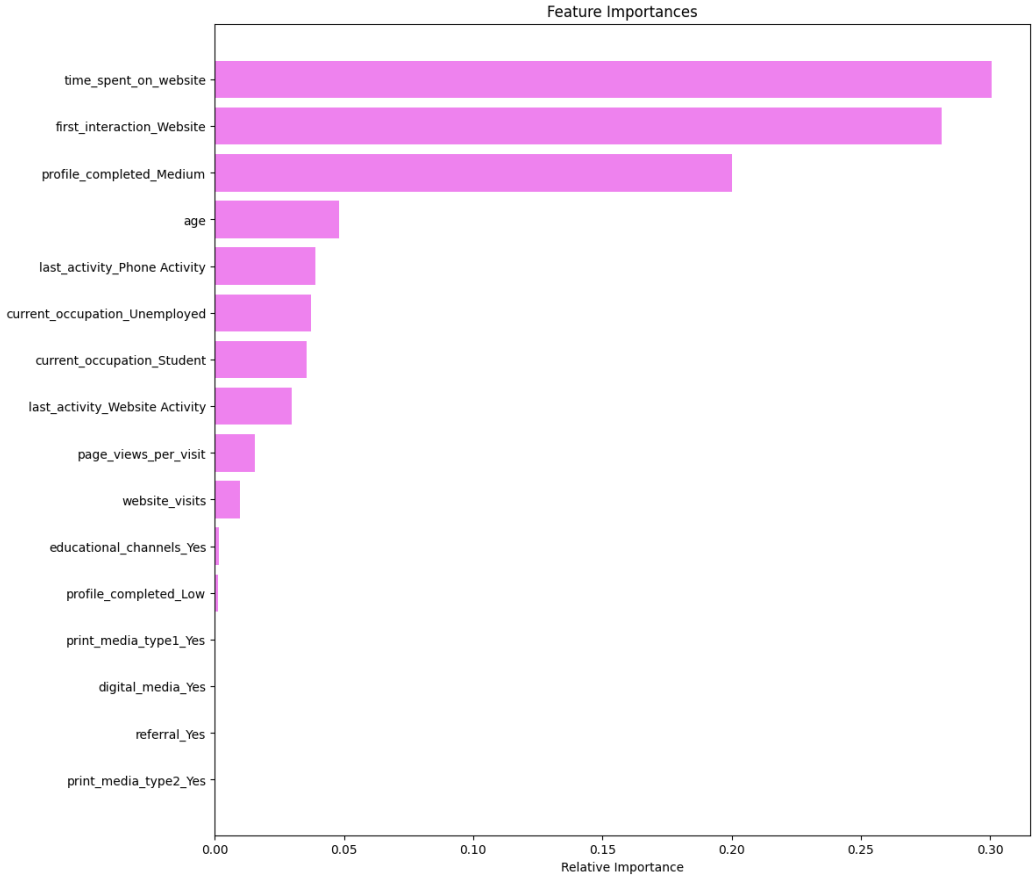
Feature Importance

- Feature Importance of Tuned Decision Tree
    - **Time spent on the website** and **first_interaction_website** are the most important features followed by profile_completed, age, and last_activity



Feature Importance

- Feature Importance of Tuned Random Forest
  - **time spent on website, first_interaction_website, profile_completed,** and **age** are the top four features that help distinguish between not converted and converted leads.
  - Unlike the decision tree, the random forest gives some importance to other variables like occupation, page_views_per_visit, as well. This implies that the random forest is giving importance to more factors in comparison to the decision tree.



Model Performance Summary

- The model achieved a higher recall score, meaning it was able to identify a greater proportion of the leads who actually converted. The accuracy also improved, indicating better overall predictive ability.
- Hyperparameter tuning significantly improves the performance of both models on the test data.
- Recall is the most important metric for this problem, as it measures the model's ability to correctly identify leads who will convert.
- The tuned Random Forest model emerged as the best-performing model. Its ability to achieve high accuracy(0.85) and a higher recall score(0.86) makes it valuable for ExtraaLearn's lead conversion prediction task.

Model	Accuracy (Train)	Recall (Train)	Accuracy (Test)	Recall (Test)
Decision Tree (Untuned)	1.00	1.00	0.79	0.70
Decision Tree (Tuned)	0.88	0.88	0.83	0.86
Random Forest (Untuned)	1.00	1.00	0.83	0.74
Random Forest (Tuned)	0.88	0.86	0.85	0.88

Key Findings and Insights

- Key Drivers of Lead Conversion: The analysis revealed that the most important factors driving lead conversion are:
  - Time spent on the website: Leads who spend more time on the website are more likely to convert.
  - First interaction through the website: Leads who first interact with ExtraaLearn through the website have a higher conversion rate compared to those who interact through the mobile app.
  - Profile completion: Leads with higher profile completion levels are more likely to convert.
  - Age: Younger leads tend to have a slightly higher conversion rate.
  - Last activity: Leads whose last activity was related to email or website interactions are more likely to convert compared to those whose last activity was phone-related.
  - Model Performance: Both Decision Tree and Random Forest models achieved good performance in predicting lead conversion. The tuned models showed improved generalization ability compared to the initial untuned models.
  - Random Forest Advantages: The Random Forest model generally performed better than the Decision Tree, demonstrating higher recall and better handling of potential overfitting.
  - Insights for Resource Allocation: The models and analysis provide valuable insights for ExtraaLearn to allocate resources effectively. Focusing on leads with higher engagement on the website, younger demographics, and those who complete their profiles can improve conversion rates.
  - Marketing Strategy Optimization: The findings suggest that ExtraaLearn should prioritize website traffic and engagement, optimize the website experience, and personalize outreach based on lead characteristics to enhance lead nurturing and conversion efforts.
  - Referrals and Educational Channels: Referrals have a positive impact on lead conversion, while educational channels show comparable conversion rates. Strategies to encourage referrals and optimize educational channel content can further improve lead generation and conversion.

Business Recommendations

1. **Enhance Website Engagement:**
    - Make the Website engaging, informative, and relevant to the target audience.
2. **Digital Marketing Campaigns:**
    - Run targeted digital marketing campaigns, such as search engine marketing (SEM), social media advertising, and email marketing, to drive traffic to the website and generate leads.
3. **Partnerships:**
    - Explore partnerships with educational institutions or organizations that cater to younger demographics to expand reach and generate leads.
4. **Referral Program:**
    - Implement a structured referral program that incentivizes existing customers to refer new leads. Offer rewards, discounts, or exclusive benefits to both the referrer and the referred lead.

Great Learning

Happy Learning !