upGrad

X Education – Lead Scoring Case Study

Team Members: Sharatkumar Gaonkar, Nidhin John, Avinash

Background

- X Education sells online courses to industry professionals.
- The company markets its courses on various websites and search engines.
- Interested professionals visit the website and browse through the courses.
- Visitors may fill out a form, providing their email or phone number, in which case they are classified as a lead.
- The sales team contacts the acquired leads through phone calls, emails, etc.
- The conversion rate of leads to enrolled students is around 30%.

Problem Statement

X Education receives a large number of leads, but its lead conversion rate is currently low, at around 30%. To improve the efficiency of their lead conversion process, the company wants to identify the most promising leads, also known as "Hot Leads." The sales team would then focus their efforts on communicating with these potential leads, rather than making calls to everyone in the lead pool. By identifying and prioritizing the most likely converters, X Education hopes to increase its lead conversion rate and enroll more students in its courses.

Objective

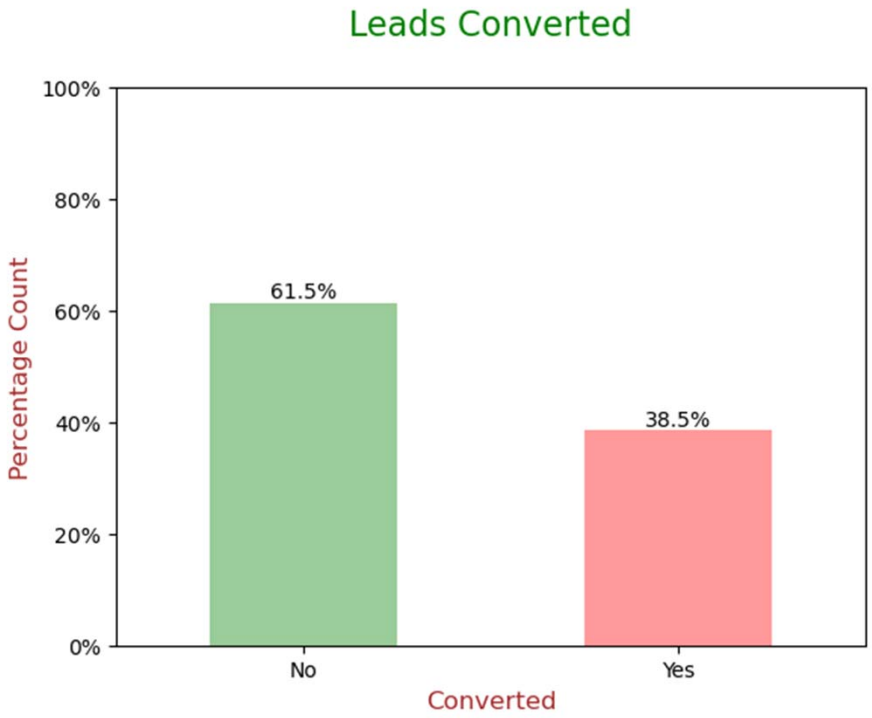
- X Education wants to improve its lead conversion rate by identifying the most promising leads, known as "Hot Leads."
- To achieve this, the company requires us to build a model that assigns a lead score to each lead.
- The lead score will help prioritize the leads based on their likelihood of converting to paying customers.
- Customers with a higher lead score are expected to have a higher conversion chance, while those with a lower score will have a lower chance of conversion.
- The CEO has set a target lead conversion rate of around 80%.

Analysis Approach

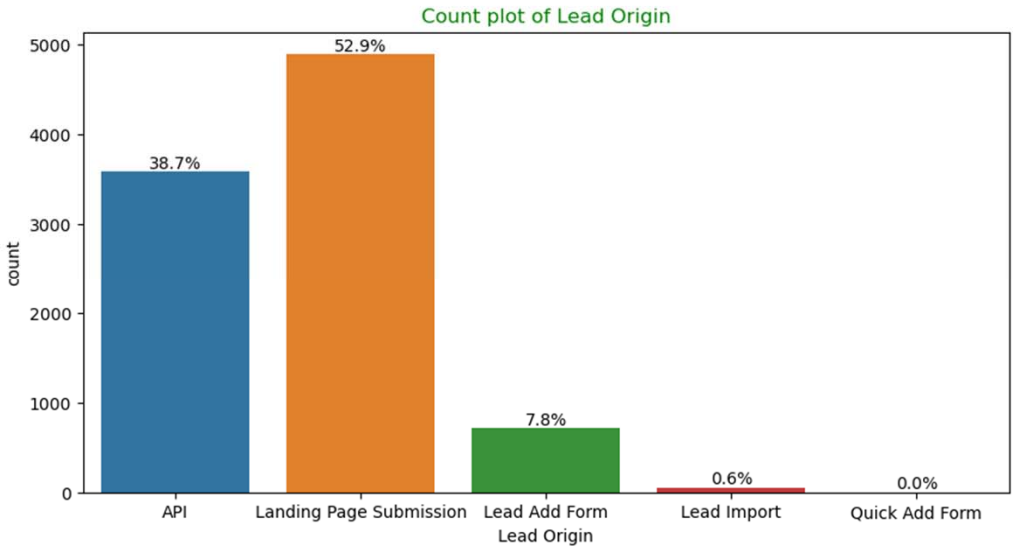
1. Read the Data
2. Data cleaning
3. Data preparation
4. Model Building
5. Model Evaluation
6. Prediction on Test Data
7. Recommendations

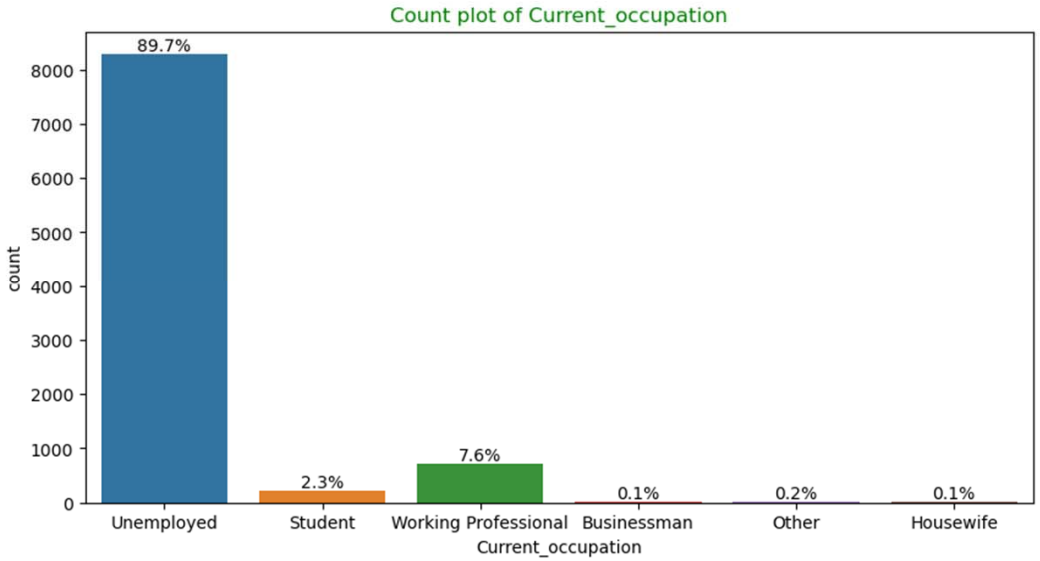
EDA – Data Imbalance

- The current lead conversion rate is 38.5%, indicating that only a minority of people who visit the X Education website convert to leads.
- In contrast, the majority of visitors, around 61.5%, do not convert to leads.



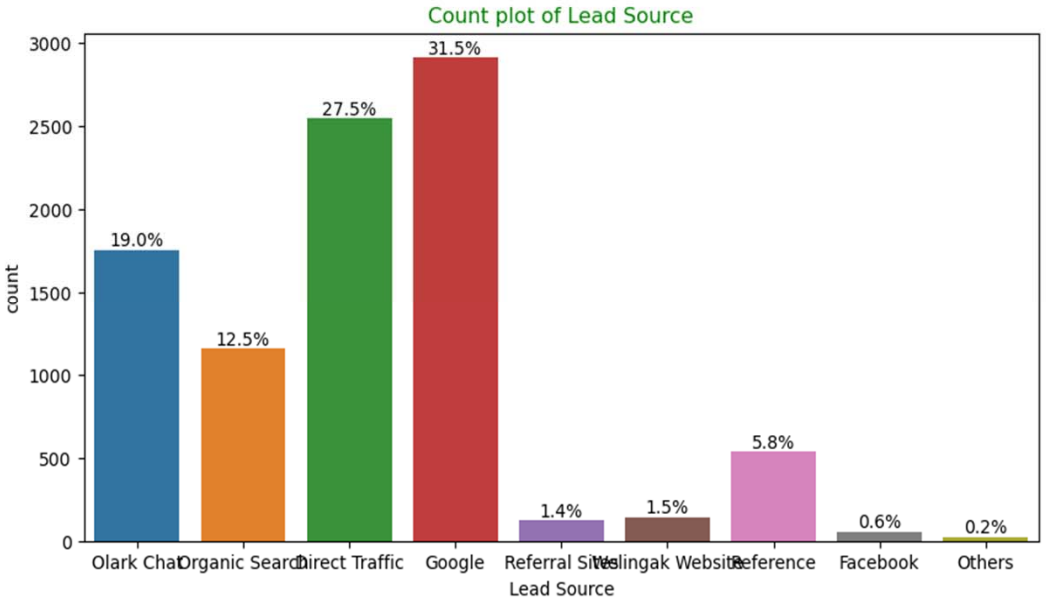
EDA – Univariate Analysis

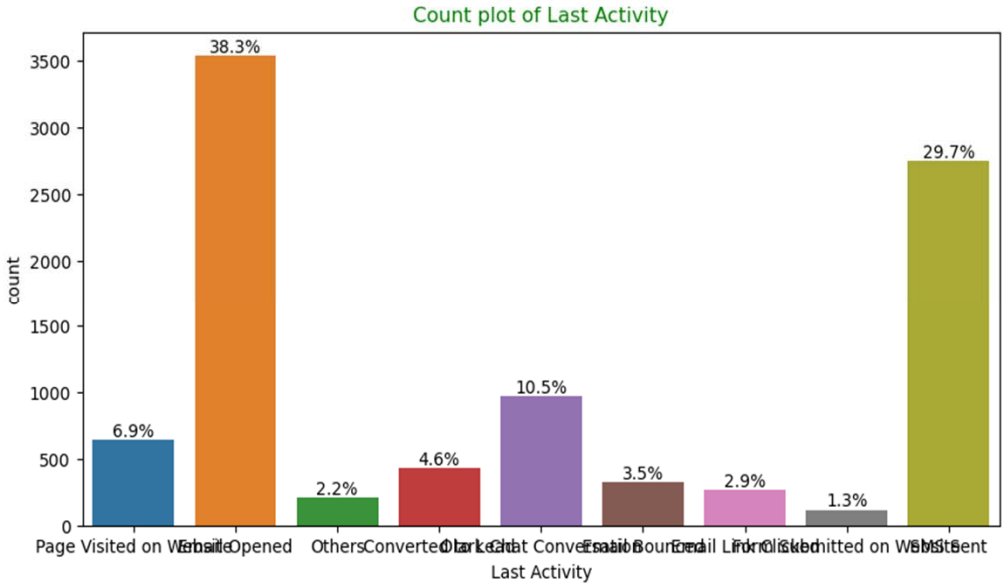




- Lead Origin: "Landing Page Submission" identified 53% of customers, "API" identified 39%.
- Current_occupation: It has 90% of the customers as Unemployed.

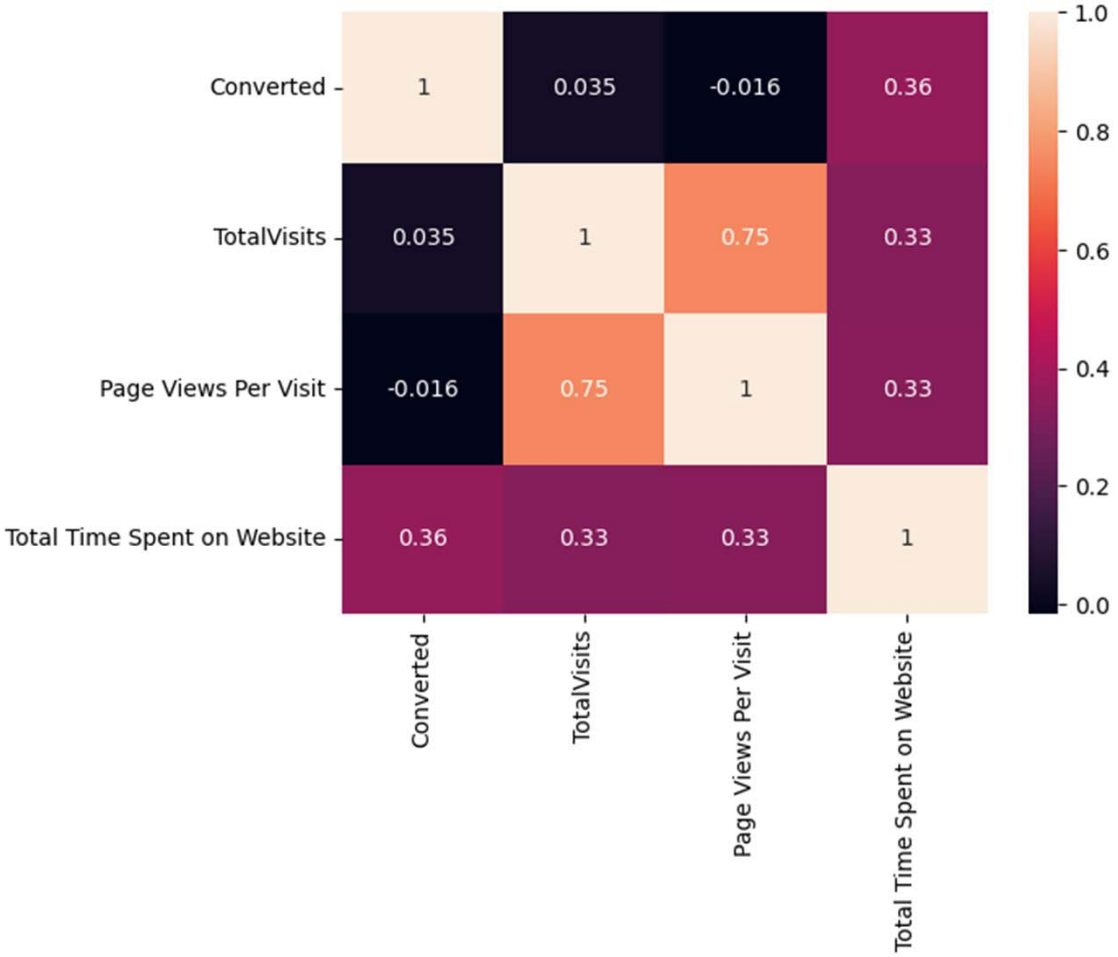
EDA – Univariate Analysis





- Lead Source: 58% Lead source is from Google & Direct Traffic combined.
- Last Activity: 68% of customers contribution in SMS Sent & Email Opened activities.

EDA – Heatmap


Converted
TotalVisits
Page Views Per Visit
Total Time Spent on Website
1
0.035
-0.016
0.36
0.035
1
0.75
0.33
-0.016
0.75
1
0.33
0.36
0.33
0.33
1
Converted
TotalVisits
Page Views Per Visit
Total Time Spent on Website
1.0
0.8
0.6
0.4
0.2
0.0

Data Preparation before Model building

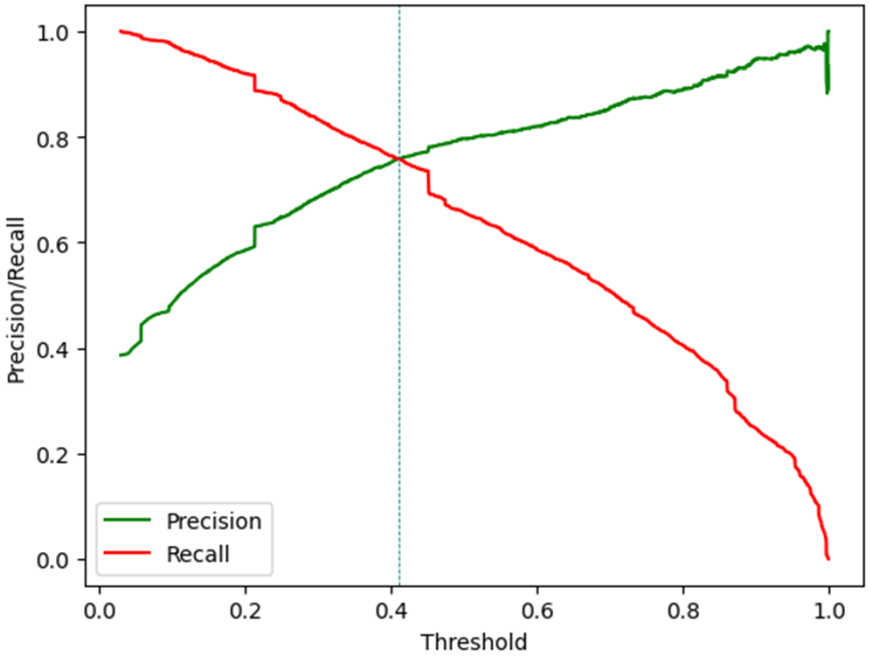
- Binary level categorical columns were mapped to 1 / 0 in previous steps to prepare them for modeling.
- Dummy features (one-hot encoded) were created for categorical variables such as Lead Origin, Lead Source, Last Activity, Specialization, and Current_occupation.
- The data was split into training and testing sets using a 70:30 ratio.
- Feature scaling was performed using standardization to ensure that all features were on the same scale.
- Correlations between predictor variables were checked, and highly correlated variables were dropped (Lead Origin_Lead Import and Lead Origin_Lead Add Form).

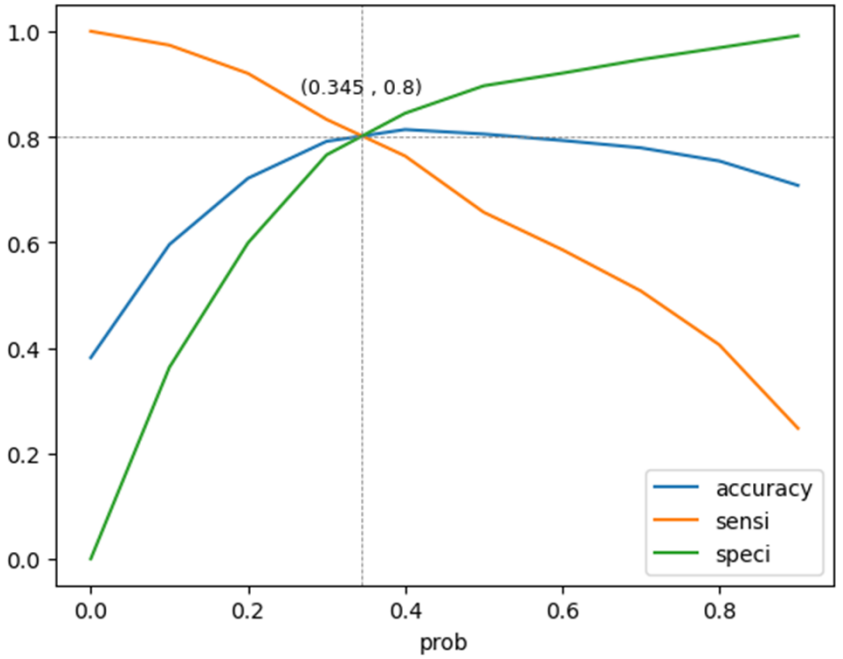
Model Building

- The dataset has a high number of dimensions and features, which can negatively impact model performance and increase computation time.
- Recursive Feature Elimination (RFE) is performed to select only the important columns.
- RFE results in a reduction of the number of columns from 48 to 15.
- Manually fine-tuning the model can then be performed on this reduced feature set.
- Manual Feature Reduction process was used to build models by dropping variables with p-value greater than 0.05.
- Model 4 was chosen as it looked stable after four iterations with significant p-values within the threshold (p-values < 0.05) and no sign of multicollinearity with VIFs less than 5.
- The final model is named logm4 (Model 4) and will be used for Model Evaluation and further predictions

Model Evaluation – Training Data set

- It was decided to go ahead with 0.345 as cutoff after checking evaluation metrics coming from both plots





Recommendations

- A regression model was developed to identify the most significant factors impacting lead conversion for X Education.
- The model highlighted the features that have the highest positive coefficients and could be given priority in marketing and sales efforts.
- The top positive features were Lead Source_Welingak Website, Lead Source_Reference, Current_occupation_Working Professional, Last Activity_SMS Sent, Last Activity_Others, Total Time Spent on Website, Last Activity_Email Opened, and Lead Source_Olark Chat.
- The model also identified features with negative coefficients that require attention for improvement, such as Specialization in Hospitality Management, Specialization in Others, and Lead Origin of Landing Page Submission.
- Prioritize the features with positive coefficients to develop targeted marketing strategies.
- Devise strategies to attract high-quality leads from top-performing lead sources such as Welingak Website and Reference.
- Optimize communication channels based on lead engagement impact, such as through SMS and email.
- Tailor messaging to engage working professionals and incentivize them to convert to leads.
- Allocate more budget to advertising on Welingak Website.
- Encourage existing leads to provide more references by offering incentives or discounts.
- Aggressively target working professionals as they have a higher conversion rate and may have better financial situations to pay higher fees.