Lead Scoring Case Study

Table of Contents

Background of X Education

- An education company named X Education sells online courses to industry professionals.
- On any given day, many professionals who are interested in the courses land on their website and browse for courses.
- The company markets its courses on several websites and search engines like Google.
- Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos.
- When these people fill up a form providing their email address or phone number, they are classified to be a lead.
- Once these leads are acquired, employees from the sales team start making calls, writing emails, etc.
- Through this process, some of the leads get converted while most do not.
- The typical lead conversion rate at X education is around 30%.

Problem statement and objective

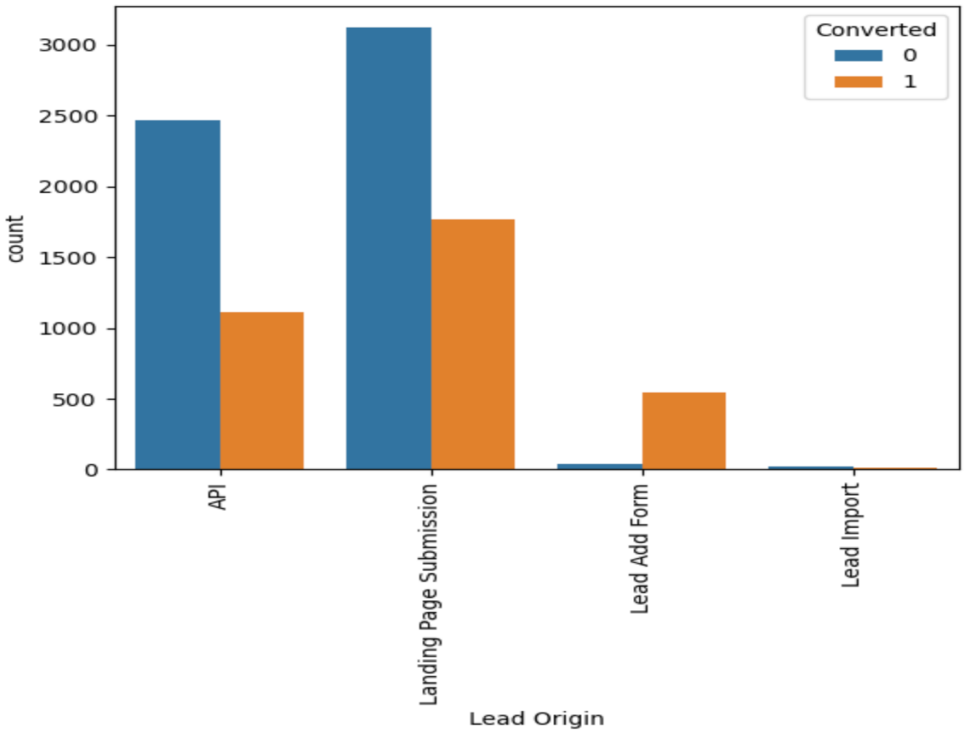
- Problem Statement:
  - X Education gets a lot of leads, its lead conversion rate is very poor at around 30%.
  - X Education wants to make lead conversion process more efficient by identifying the most potential leads, also known as Hot Leads.
  - Their sales team want to know these potential set of leads, which they will be focusing more on communicating rather than making calls to everyone.
- Objective of the Study:
  - To help X Education select the most promising leads, i.e., the leads that are most likely to convert into paying customers.
  - The company requires us to build a model wherein we need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.
  - The CEO has given a ballpark of the target lead conversion rate to be around 80%.

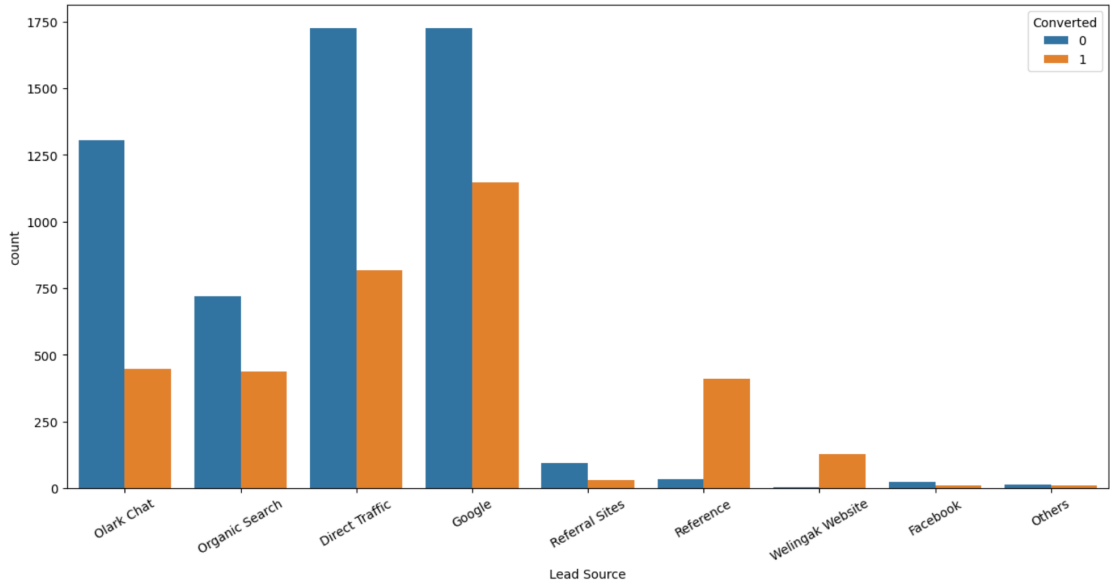
Data Cleaning

- "Select" level represents null values for some categorical variables, as customers did not choose any option from the list.
- Columns with over 40% null values were dropped.
- Missing values in categorical columns were handled based on value counts and certain considerations.
- Drop columns that don't add any insight or value to the study objective.
- Imputation was used for some categorical variables.
- Additional categories were created for some variables.
- Columns with no use for modelling or only one category of response were dropped.
- Numerical data was imputed with mode after checking distribution.

EDA

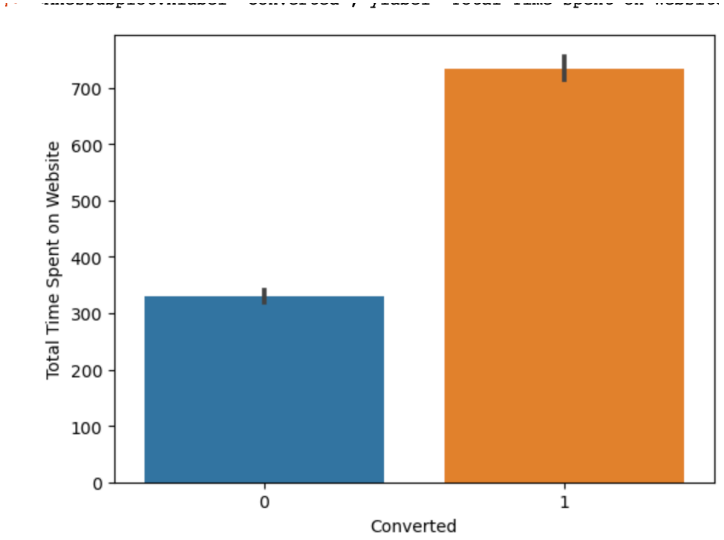
EDA

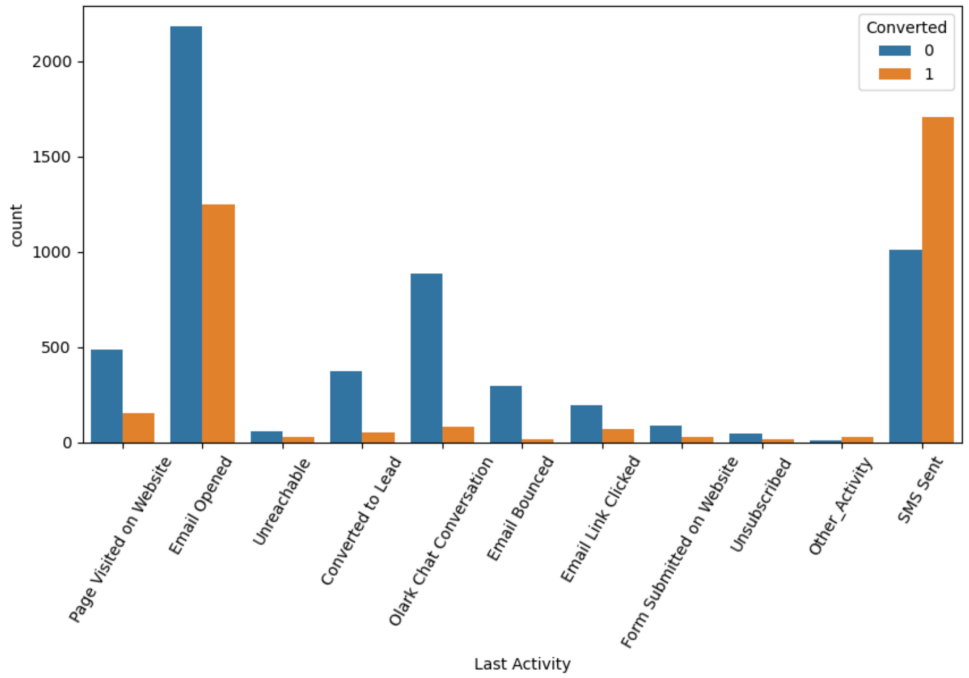




Google and Direct traffic generates maximum number of leads.
Conversion Rate of reference leads and leads through welingak website is high.
To improve overall lead conversion rate, focus should be on improving lead converion of olark chat, organic search, direct traffic, and google leads and generate more leads from reference and welingak website.

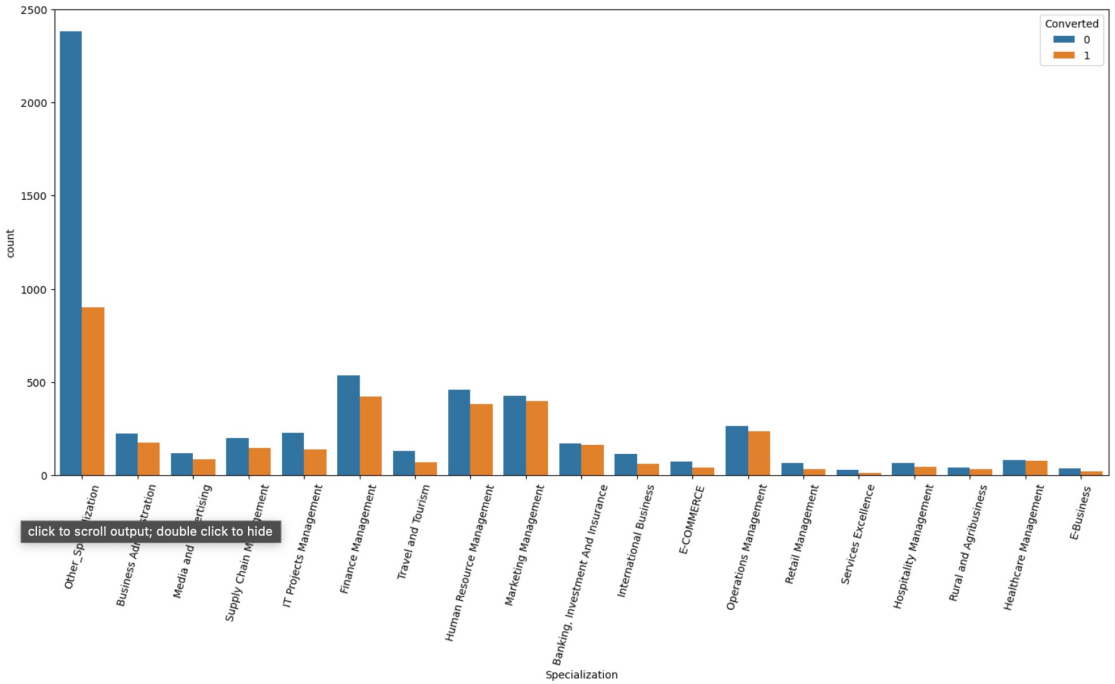
EDA

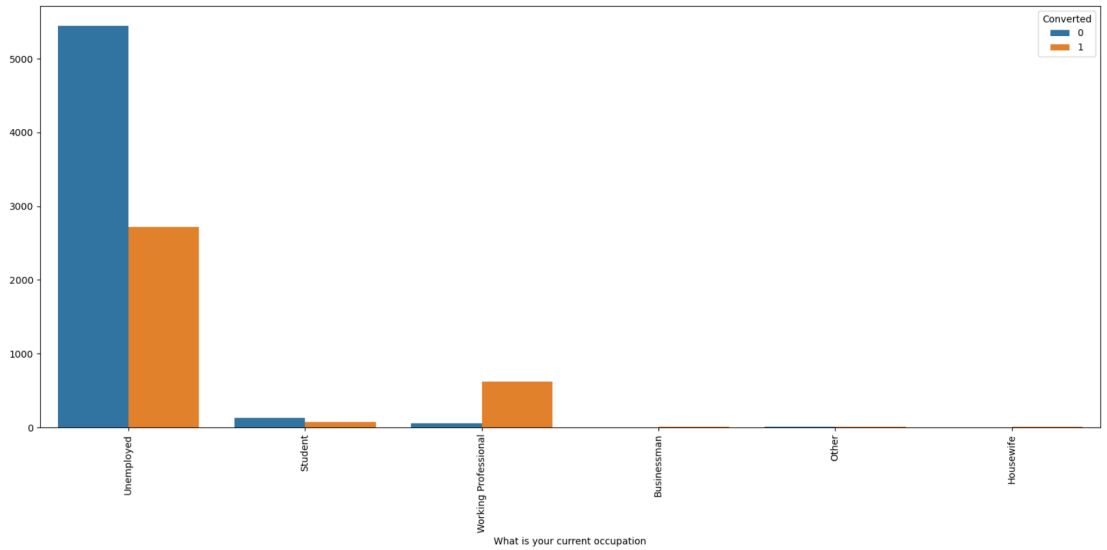




Leads spending more time on the weblise are more likely to be converted.
Website should be made more engaging to make leads spend more time.
Most of the lead have their Email opened as their last activity.
Conversion rate for leads with last activity as SMS Sent is almost 60%.

EDA





Focus should be more on the Specialization with high conversion rate.
We can focus here on Management related Professions.
Working Professionals going for the course have high chances of joining it.
Unemployed leads are the most in numbers but has around 30-35% conversion rate.

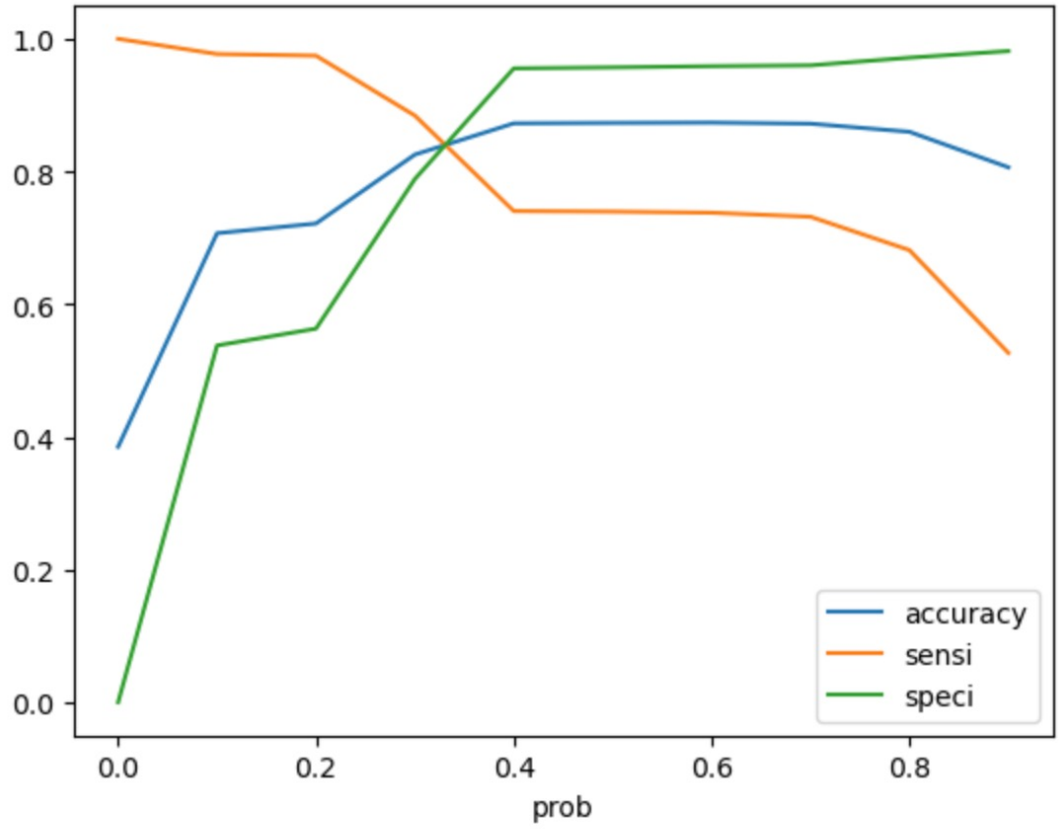
Data Preparation

- Binary level categorical columns were already mapped to 1 / 0 in previous steps.
- Created dummy features (one-hot encoded) for categorical variables – Lead Origin, Lead Source, Last Activity, Specialization, Current_occupation.
- Splitting Train & Test Sets ○ 70:30 % ratio was chosen for the split
- Feature scaling - Standardization method was used to scale the features.

Model Building

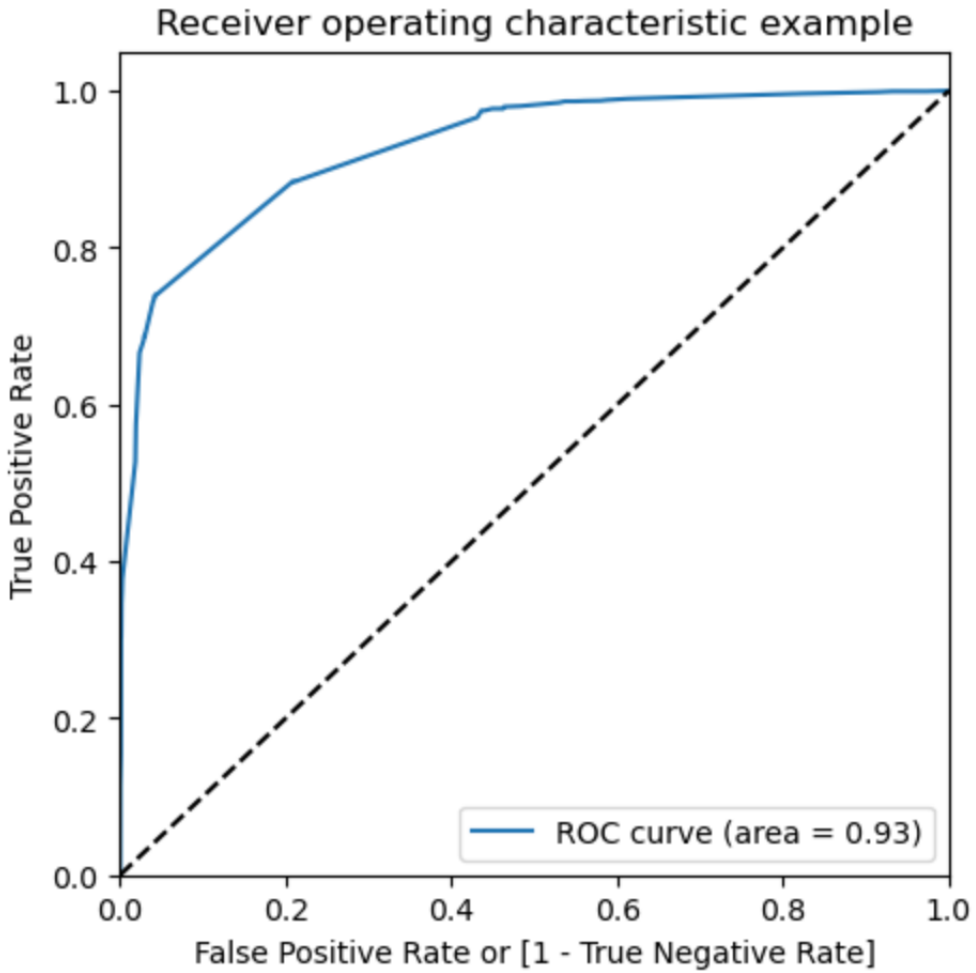
- The data set has lots of dimension and large number of features.
- This will reduce model performance and might take high computation time.
- Hence it is important to perform Recursive Feature Elimination (RFE) and to select only the important columns.
- RFE outcome ○ Pre RFE – 48 columns & Post RFE – 15 columns.
- Manual Feature Reduction process was used to build models by dropping variables with p – value greater than 0.05.

Model Evaluation



From the curve above, 0.25 is the optimum point to take it as a cutoff probability.

ROC Curve



Area under ROC curve is 0.88 out of 1 which indicates a good predictive model.
The curve is as close to the top left corner of the plot, which represents a model that has a high true positive rate and a low false positive rate at all threshold values.

Recommendation

To increase our Lead Conversion Rates :

- Focus on features with positive coefficients for targeted marketing strategies.
- Develop strategies to attract high-quality leads from top-performing lead sources.
- Optimize communication channels based on lead engagement impact.
- Engage working professionals with tailored messaging.
- More budget/spend can be done on Welingak Website in terms of advertising, etc.
- Incentives/discounts for providing reference that convert to lead, encourage providing more references.
- Working professionals to be aggressively targeted as they have high conversion rate and will have better financial situation to pay higher fees too.

To identify areas of improvement :

- Analyse negative coefficients in specialization offerings.
- Review landing page submission process for areas of improvement

Thank You