X Education
Lead Scoring Case Study

About X Education Company

- 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 & Objective of the Study

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 and Preparation

- Columns with over 35% null values was dropped.
- Columns that don’t add any insight or value to the objective of the study was dropped (Ex: City, Country)
- Some categorical variables have the level ‘Select’, as customers did not choose any option from the list. ' How did you hear about X Education' and 'Lead Profile’ were dropped as it had lot of rows with the value ‘Select’.
- Columns with no use of modeling (Prospect ID , Lead Number) and only one category of response were dropped.
- Category columns which were skewed were checked and dropped to avoid bias in the logistic regression model.
- For columns which didn’t have significant number of null values, just the null rows for the column had been dropped.
- After cleaning the data ,imbalance is checked for the target variable and it has 48% leads conversion rate.

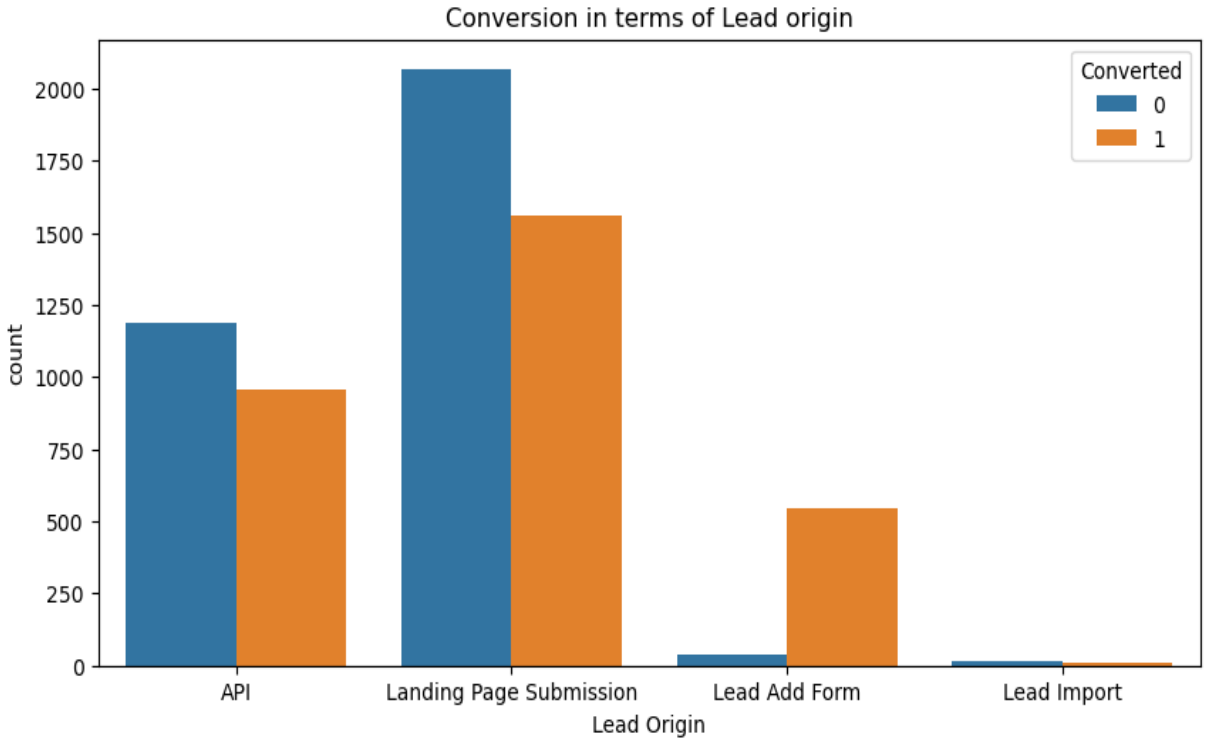
Data Preparation before Model building

- Dummy variables were created for categorical variables like Lead Origin, Lead Source etc.
- For the column Specialization dummy variables were created separately as it has the level ‘Select’ and the dummy variable for that level was dropped.
- Train and Test sets were split according to the ration 70:30.
- Normalization was used to scale the features.
- Correlations between the variables were checked. Identified the most correlated variables like ‘Last Notable Activity_Email Marked Spam’ and ‘Last Activity_Email Marked Spam’ are highly correlated to each other. Next to these were ‘Lead Source_Facebook’ and ‘Lead Origin_Lead Import’. And the third most correlated attributes were ‘Last Notable Activity_SMS Sent’ and ‘Last Activity_SMS Sent’

EDA - Univariate and Bi-Variate Analysis

Lead Origin

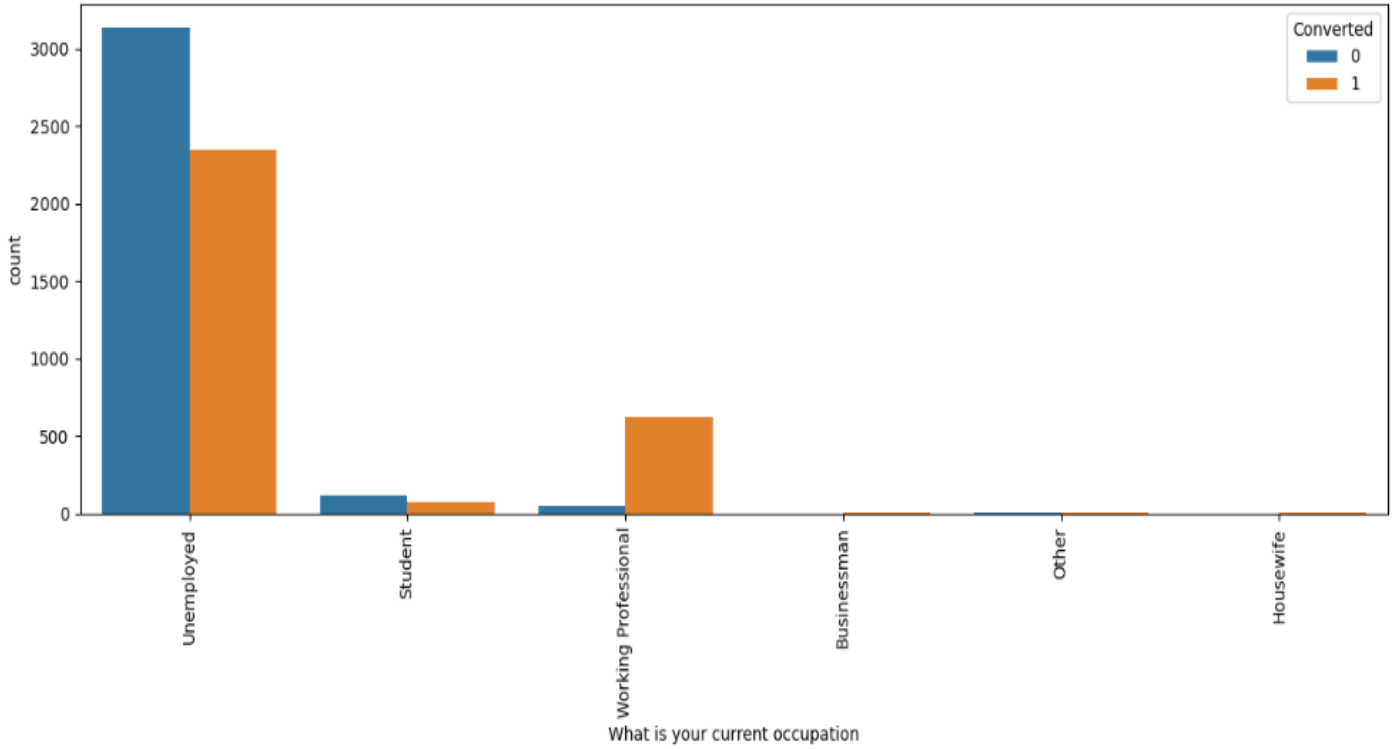
- Lead Add form has the highest conversion rate at 94%
- Although the conversion rate for API and Landing Page Submission is low, they generate maximum lead counts.



- So, we need to generate more leads from Lead Add form since they have high conversion rate.

EDA - Univariate and Bi-Variate Analysis

- Working Professionals and Unemployed people generate maximum leads.
- Conversion rate for Working Professional is high at around 92% and that of Unemployed is low at around 33%.



- We need to focus on improving lead conversion of unemployed to improve overall lead conversion rate. Also , generate more leads from Working Professionals.

Model Building

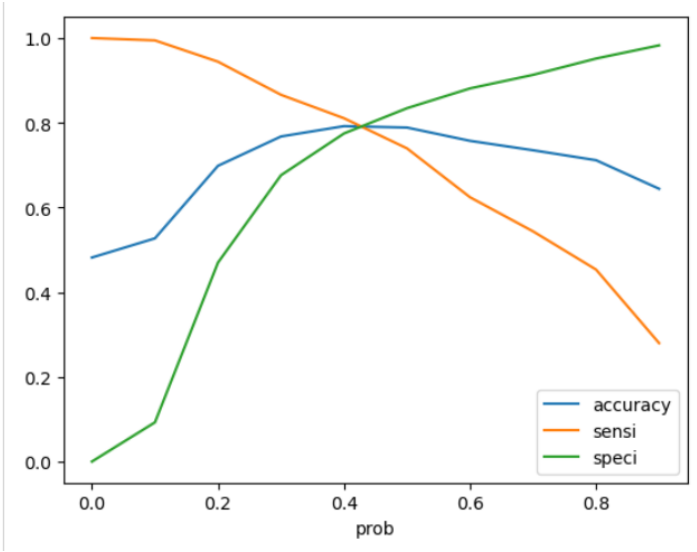
- The dataset has large number of features.
- This will reduce performance and might result in high computation time.
- Hence Recursive Feature Elimination (RFE) was used to select only the important columns.
- This was done to manually tune the model
- After feature elimination using RFE from 75 variables to only 15 variables were selected.

Model Building

- Statsmodel was used to assess the model.
- Feature variables with p-value greater than 0.05 was dropped.
- VIF values were also checked.
- We got a stable model after four iterations.
- The p-values were within the threshold (p-values < 0.05) and no sign of multicollinearity was observed as the VIFs were less than 5.
- The final model will we used for model evaluation and prediction.

Model Evaluation

- Conversion probability of 0.5 was chosen but the results were not satisfactory.
- ROC was plotted and we got an area under the curve as 0.86 which indicates good predictive probability.
- To find the optimal cutoff, sensitivity and specificity tradeoff was checked.
- The optimal value of the three metrics accuracy, sensitivity and specificity came at around 0.42.



Model Evaluation

- After running the model on the Train Dataset these are the figures that we obtain:

  Accuracy : 79.08%
  Sensitivity : 79.33%
  Specificity : 78.84%

- After running the model on the Test Dataset these are the figures that we obtain:

  Accuracy : 78.45%
  Sensitivity : 77.94%
  Specificity : 78.91%

Model Evaluation

- Using a cutoff value of 0.42, the model achieved a sensitivity of 79.33% in the train set and 77.94% in the test set.
- Sensitivity indicates the number of leads the model identified correctly out of all the potential leads which are converting.
- The CEO of X education had set a target of around 80%.
- The model also achieved an accuracy of around 80% which is in line with the study’s objectives.

Recommendations

- As per the final model, increasing the lead conversion is very important for growth and success of X education. For this we have developed a regression model that helps us identify significant factors that impact lead conversion.
- For marketing and sales efforts to increase lead conversion, we have determined the following features with highest positive coefficitents.

  TotalVisits: 11.14

  Total Time Spent on Website: 4.42

  Lead Origin_Lead Add Form: 4.20

  Last Notable Activity_Unreachable: 2.78

  Last Activity_Had a Phone Conversation: 2.75

  Lead Source_Welingak Website: 2.15

  Lead Source_Olark Chat: 1.45

  Last Activity_SMS Sent: 1.18

Recommendations

- We have negative coefficients too which may indicate potential areas of improvement.

  What is your current occupation_Unemployed: -2.54

  What is your current occupation_Student: -2.35

  Do Not Email_Yes: -1.50