Lead Scoring Case Study

Prepared by:

Ritij Srivastava

Sanjay Tom Perayil

Problem Statement

- An education company named X Education sells online courses to industry professionals. It markets its courses on several websites and search engines like Google to generate leads.
- X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted which is typically 30%.
- Senior management of the company wants the lead conversion rate to be 80%.

Goal

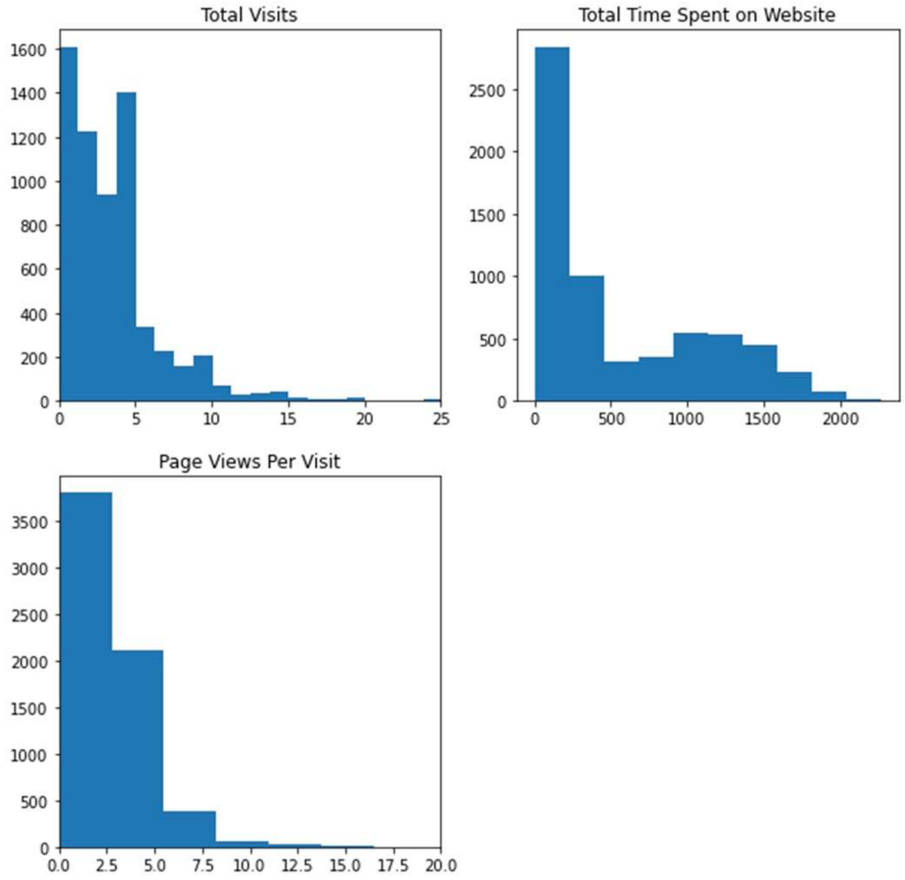
- To build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads.

- A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

Methodology

- Data Cleaning and Data manipulation.
  - Check and handle missing values from the dataset.
  - Drop columns ,if it contains large amount of missing values and not useful for the analysis.
- EDA
  - Univariate Analysis
  - Bi- Variate Analysis
- Dummy variables & Feature Scaling and splitting of the dataset.
- Classification technique: logistic regression is used for the modelling and prediction.
- Model Evaluation.
- Final Model presentation.
- Conclusion and recommendation.

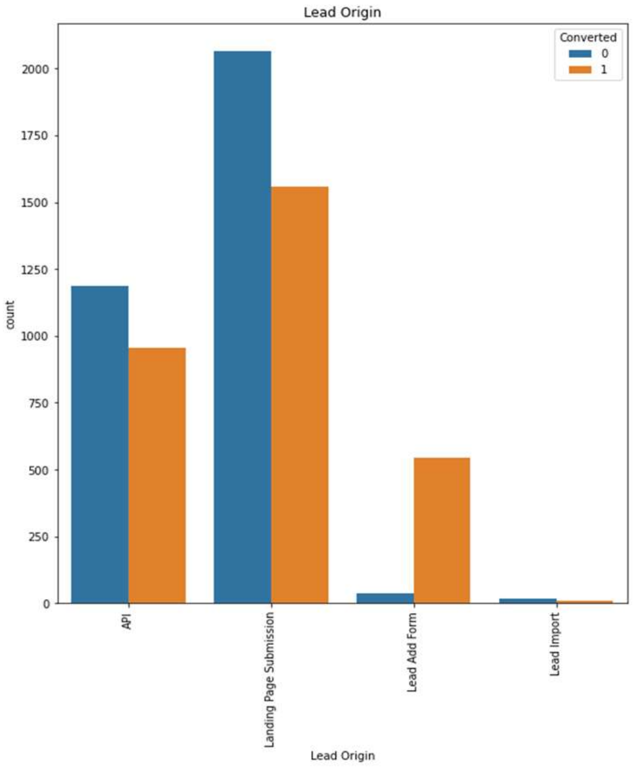
Conversion of Leads to Clients

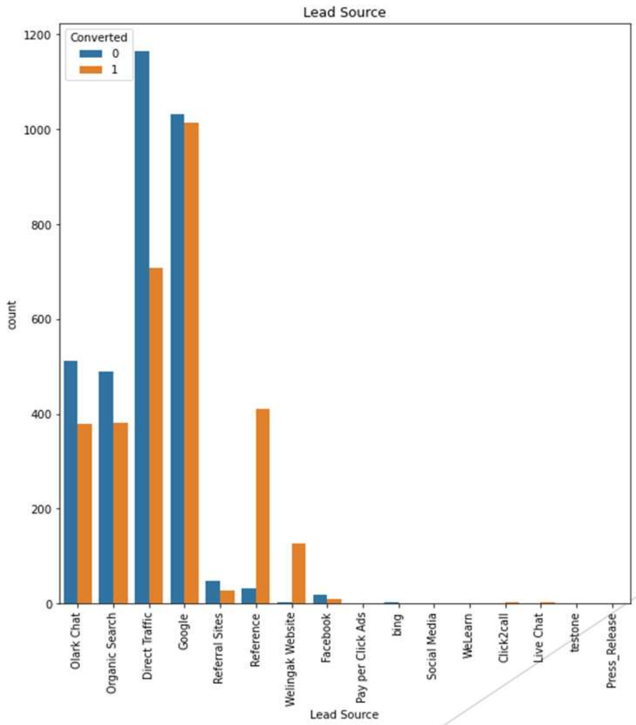


- Total Visit, Total time spent on website & Page Views per visit might impact in lead conversion. Hence, we keep these variables.

Bi-Variate Analysis of Categorical Variable

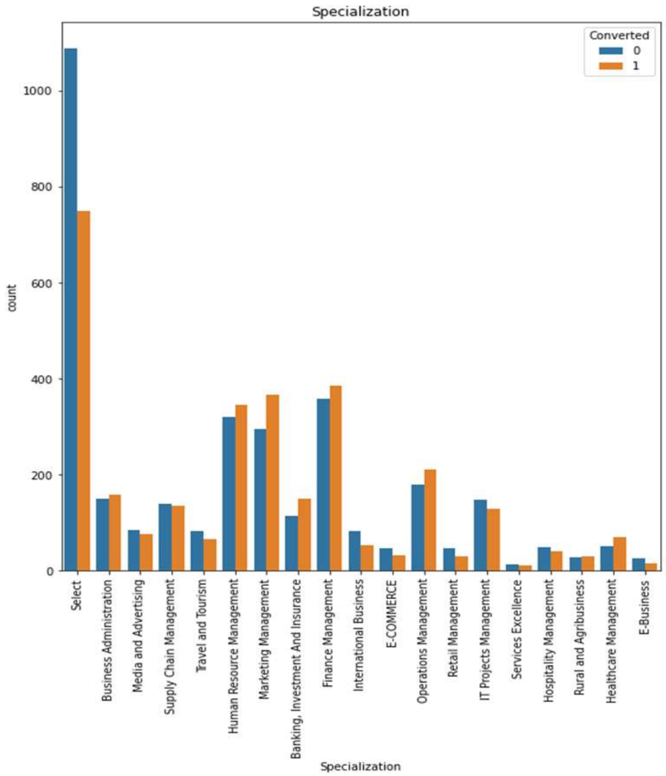
Lead Origin and Lead Source

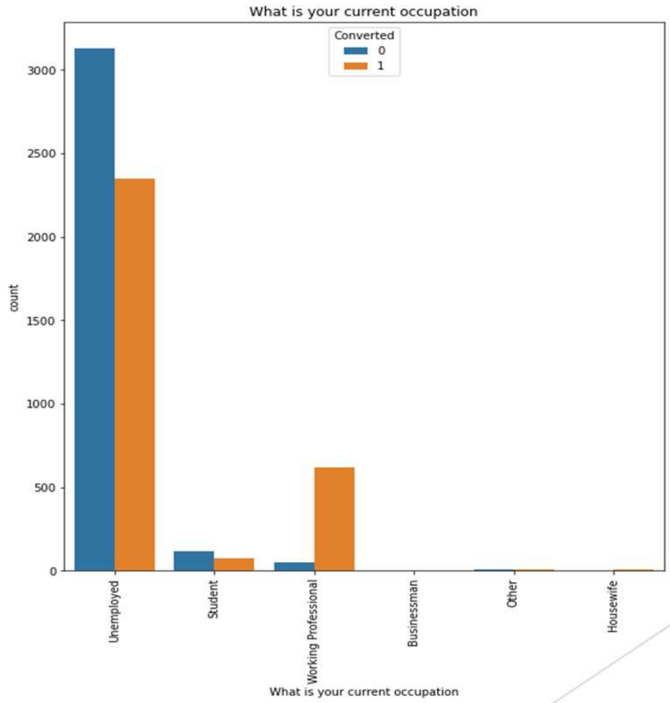




Categorical Variable Relation

Specialization and Current Occupation

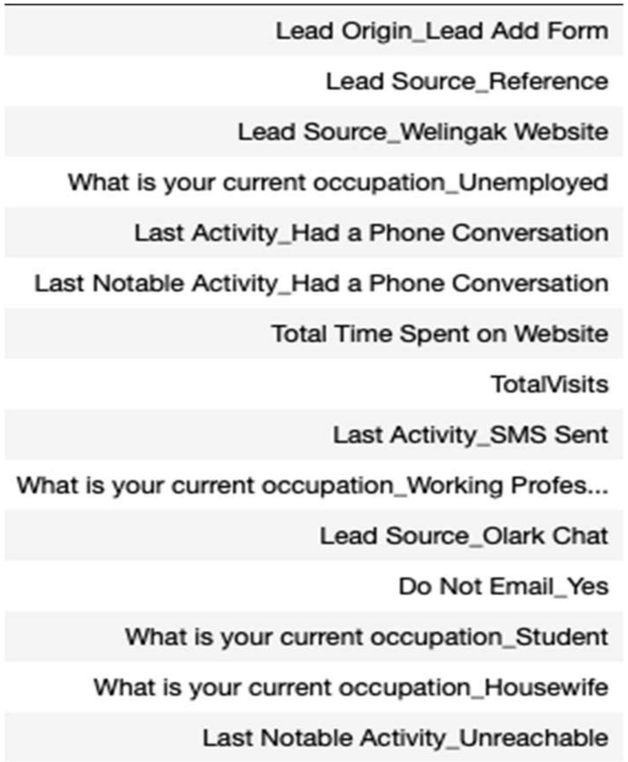




Model Building Steps

- Splitting the data into training and testing data sets.
- The primary step for regression is performing a train-test split with a ratio of 70:30.
- Use RFE for Feature selection.
- Running RFE with 15 variables as output.
- Building model by removing the variables whose p-value is greater than 0.05 and VIF value is greater than 5.
- Using above approach we have eliminated below variables:
  - Lead Source_Reference
  - Last Notable Activity_Had a Phone Conversation
  - What is your current occupation_Housewife
  - What is your current occupation_Working Professional

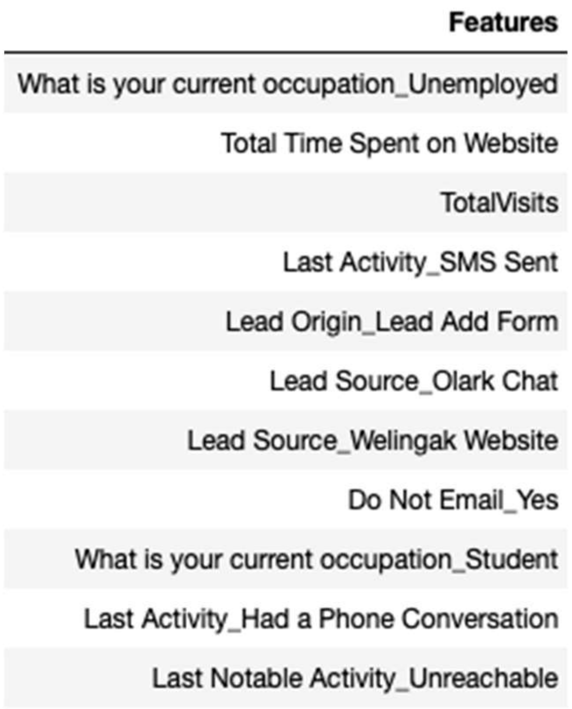
Feature Selection Using RFE



Lead Origin_Lead Add Form
Lead Source_Reference
Lead Source_Welingak Website
What is your current occupation_Unemployed
Last Activity_Had a Phone Conversation
Last Notable Activity_Had a Phone Conversation
Total Time Spent on Website
TotalVisits
Last Activity_SMS Sent
What is your current occupation_Working Profes...
Lead Source_Olark Chat
Do Not Email_Yes
What is your current occupation_Student
What is your current occupation_Housewife
Last Notable Activity_Unreachable

- Initially we started building the model with 15 variable selected through RFE method.

Factors/Features affecting the lead conversion



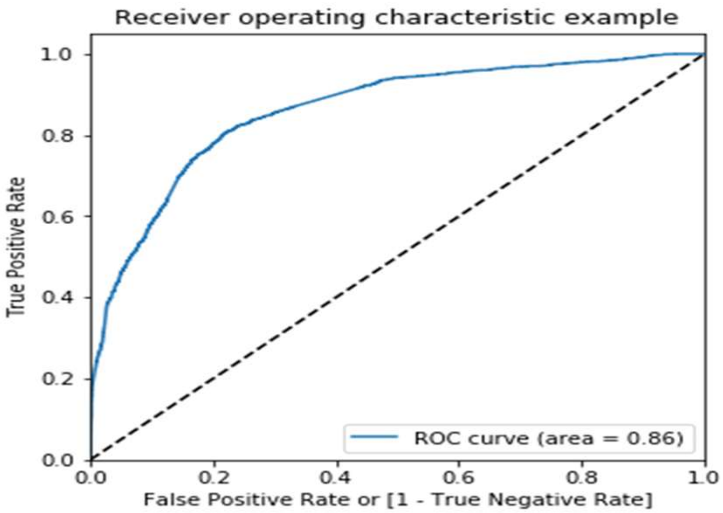
Features
What is your current occupation_Unemployed
Total Time Spent on Website
TotalVisits
Last Activity_SMS Sent
Lead Origin_Lead Add Form
Lead Source_Olark Chat
Lead Source_Welingak Website
Do Not Email_Yes
What is your current occupation_Student
Last Activity_Had a Phone Conversation
Last Notable Activity_Unreachable

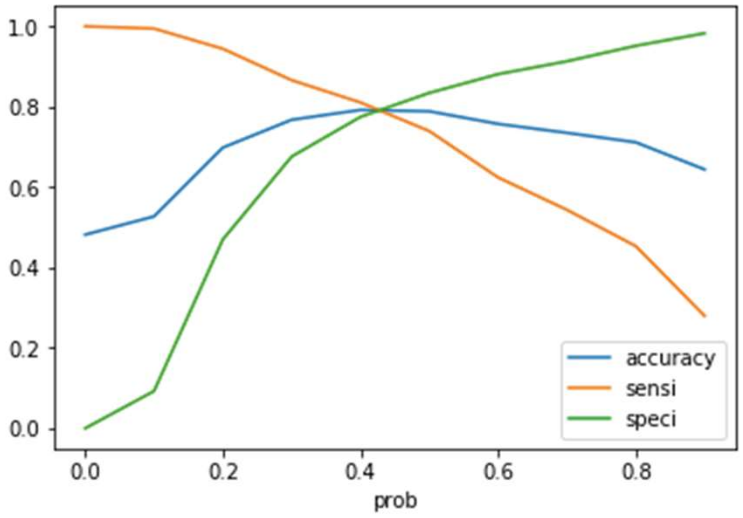
These are the feature which is having direct impact on the lead conversion.

Sensitivity & Specificity

- Checking Sensitivity & Specificity is import to evaluate our model.
- In our model:
  - Sensitivity is 0.7394
  - Specificity is 0.8343
- Sensitivity measure is used to determine the proportion of actual positive cases, which got predicted correctly.
- Specificity measure is used to determine the proportion of actual negative cases, which got predicted correctly.

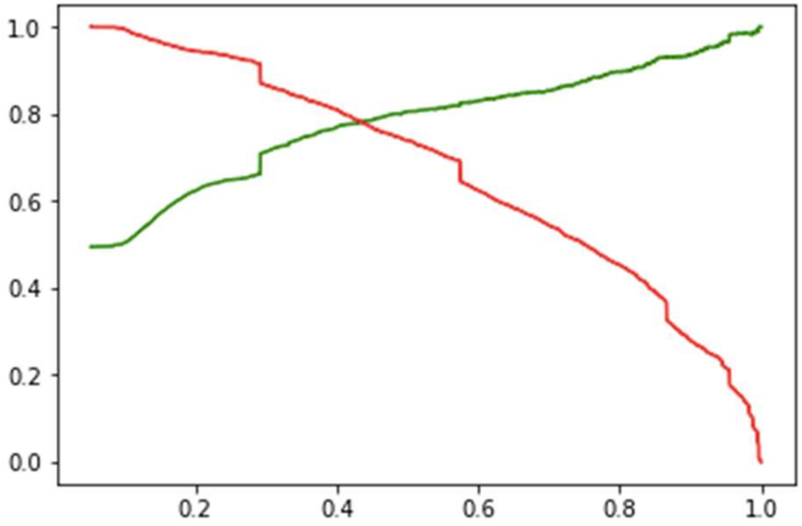
ROC Curve





- Plotting the ROC curve by randomly choosing the 0.5 as cut-off
- After plotting Accuracy, Sensitivity & Specificity we found that optimal cut-off value will be around 0.42.
- Sensitivity & Specificity is 0.7933 & 0.7884 resp. which quite significant

Precision - Recall Curve



- As we can see from above plot that our model is working find on test set and can predict the values with 78.95% accuracy with Precision value of 0.7840 & Recall of 0.7771

Conclusion

- We can observe from our model feature which matter most in lead conversion are mentioned below (In descending order) :
  - Total time spend on the Website
  - Total number of visits
  - When lead was sourced from:
    - Direct traffic on website
    - Olark Chat
    - Welingak website
  - Basis upon the last activity:
    - SMS
    - Had phone Conversation
  - When the lead is generated from the Landing Page Submission.
  - Students are also likely to enrol for the course but we should focus on specific segment of student.
- In order to reach the targeted conversion ratio which is 80%. X education must focus on these features & segment for lead generation. As lead generated from these channel are most likely to get converted.

Recommendation

- We would recommend X Education to focus on below datapoints:
  - What is your current occupation_Unemployed
  - Total Time Spent on Website
  - TotalVisits
  - Last Activity_SMS Sent
  - Lead Origin_Lead Add Form
  - Lead Source_Olark Chat
  - Lead Source_Welingak Website
  - Do Not Email_Yes
  - What is your current occupation_Student
  - Last Activity_Had a Phone Conversation
  - Last Notable Activity_Unreachable

Thank You