Lead Score Case Study

Problem Statement

X Education sells online courses to industry professionals. 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. To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Business Objective:

- X education wants to know most promising leads.
- For that they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use.

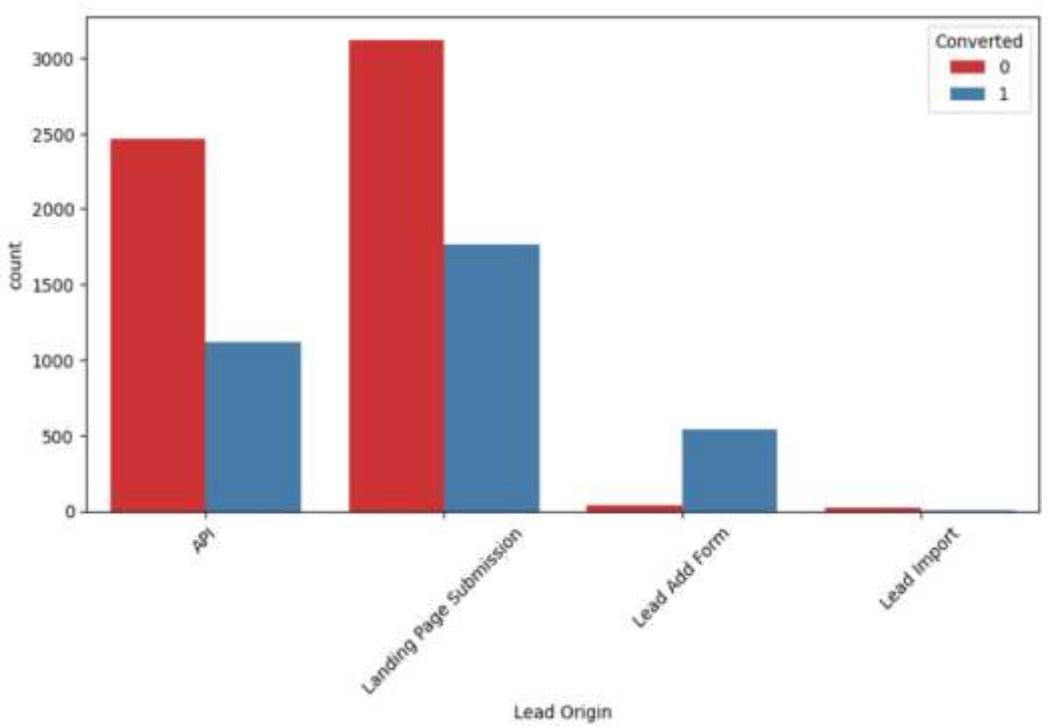
Data Cleaning and Data Manipulation

- 'Select' values in many columns may be present because the customer did not select any option from the list, hence it shows 'Select'. These values can be treated as nan.
- 37% values are missing from 'Specialization' column. New category as 'Other' in specialization and nan values can be included in it.
- Missing values are from tag column can be replaced by mode occurrence 'Will revert after reading the email' .
- 85% of occupation is 'Unemployed', nan values can be rplaced by 'Unemployed'.
- 58% of leads comes from Mumbai city.
- Columns 'What matters most to you in choosing a course' is highly skewed and can be removed from further analysis.

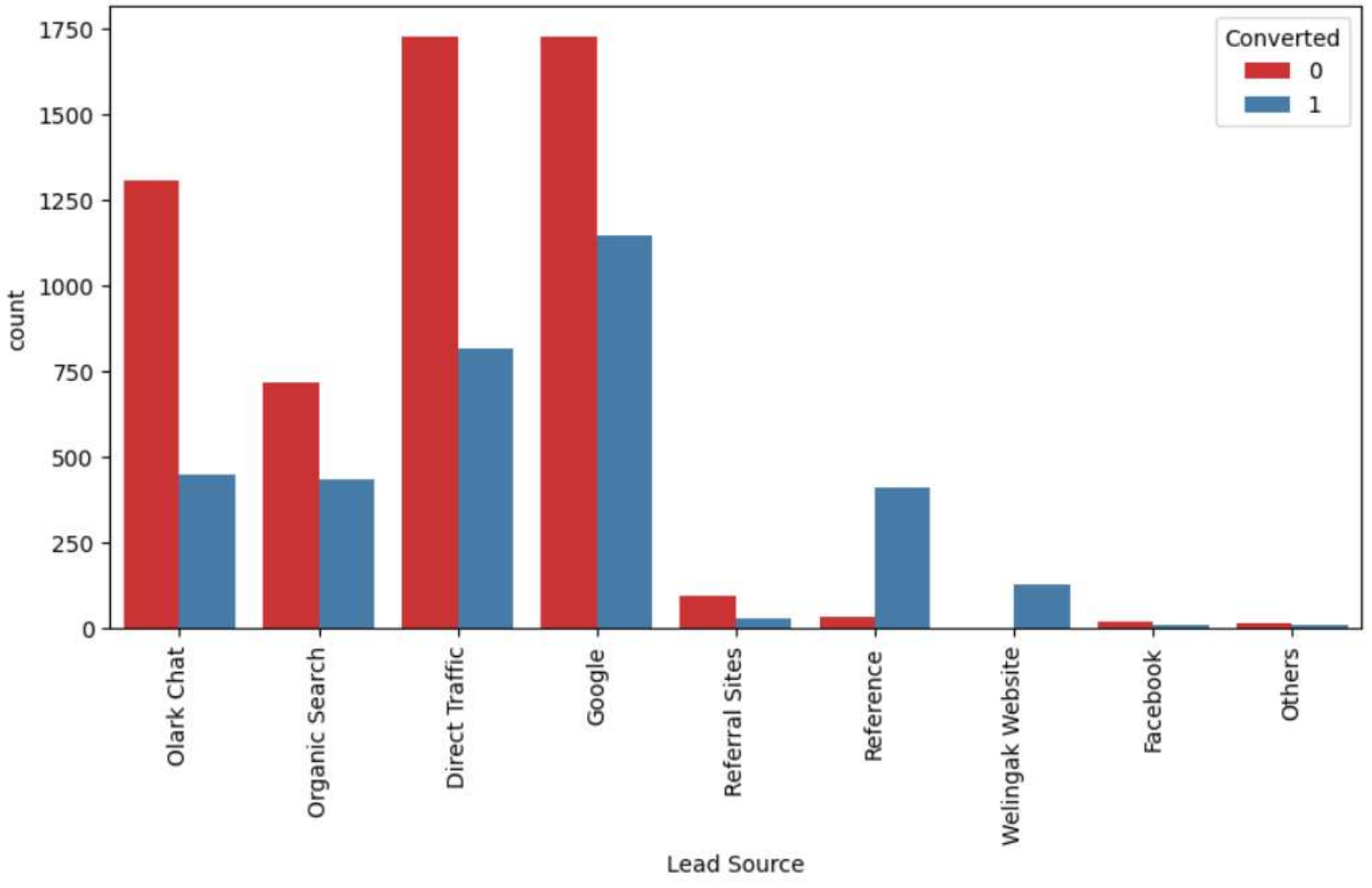
Univariate Analysis:

- The lead conversion rate is 38%.
- Lead Add Form has more than 90% conversion rate but count of lead are not very high.
- Leads spending more time on the weblise are more likely to be converted.
- 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%.
- Most leads are from mumbai with around 50% conversion rate.

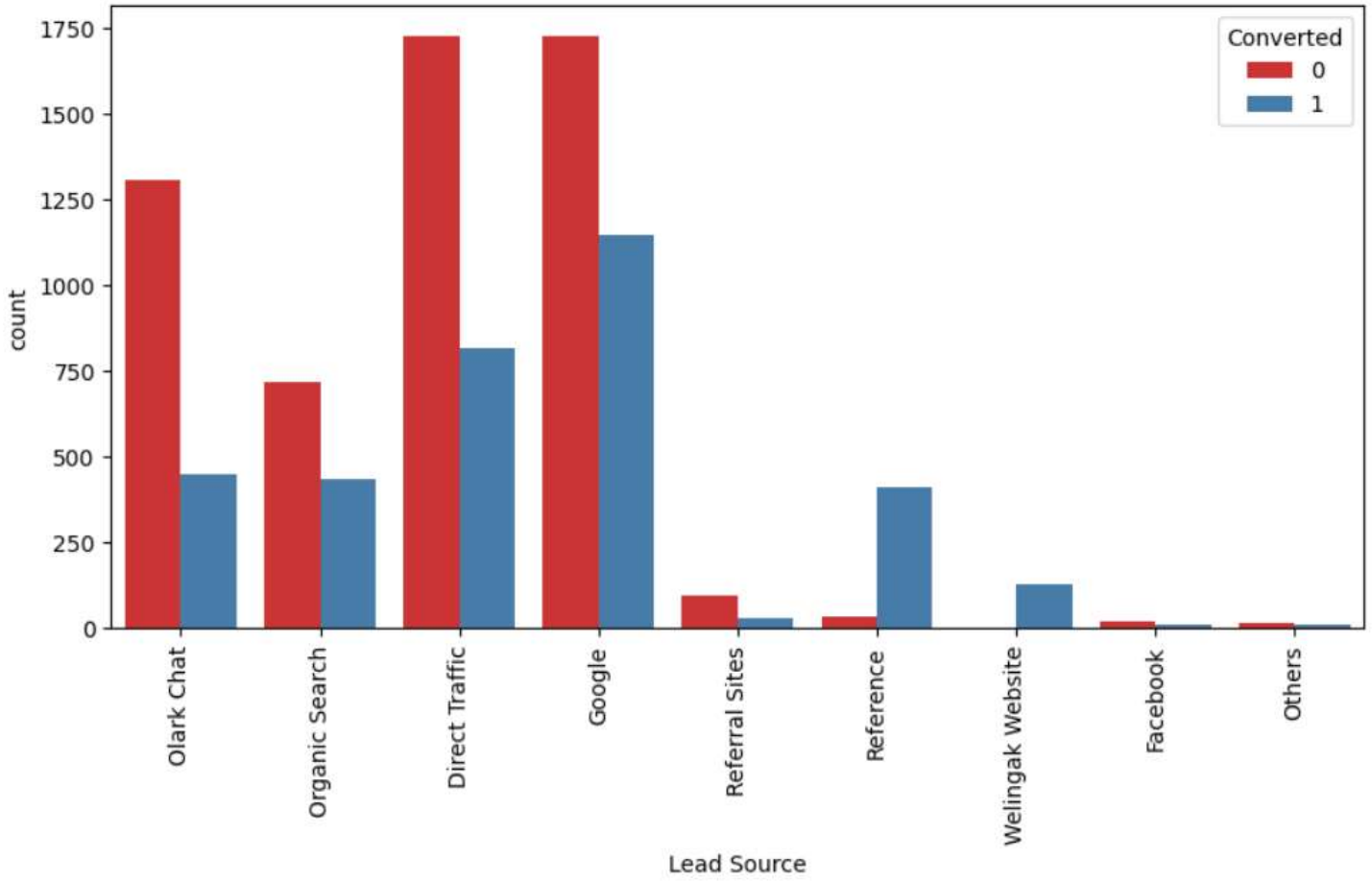
API and Landing Page Submission have 30-35% conversion rate but count of lead originated from them are considerable.



Google and Direct traffic generates maximum number of leads.



Conversion Rate of reference leads and leads through welingak website is high.



Data Preparation

- Dummy variables can be created for columns: 'Lead Origin', 'Lead Source', 'Last Activity', 'Specialization','What is your current occupation','City','Last Notable Activity’.
- We have almost 38% lead conversion rate.
- Following columns requires scaling - 'TotalVisits','Total Time Spent on Website','Page Views Per Visit']] = scaler.fit_transform(X_train[['TotalVisits','Total Time Spent on Website','Page Views Per Visit’.
- Top 20 columns is being selected by RFE and these features are further eliminated based on p and z value and VIF.

Final Model

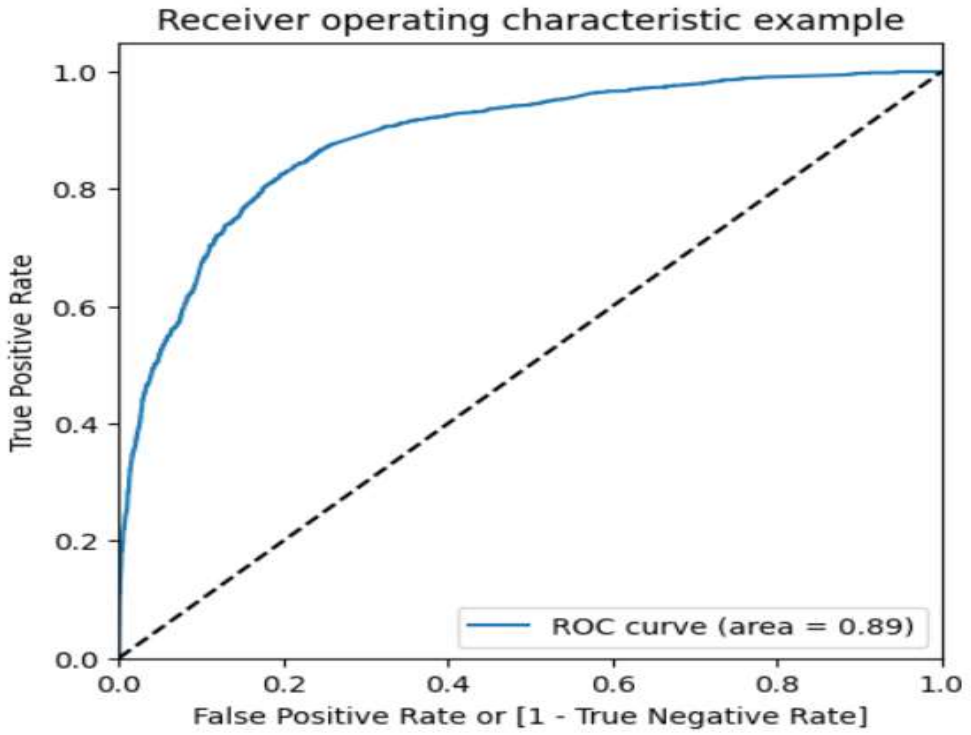
	Features	VIF
9	Specialization_Others	2.16
3	Lead Source_Olark Chat	2.03
11	Last Notable Activity_Modified	1.78
2	Lead Origin_Landing Page Submission	1.69
6	Last Activity_Olark Chat Conversation	1.59
8	Last Activity_SMS Sent	1.56
1	Total Time Spent on Website	1.29
4	Lead Source_Reference	1.24
10	What is your current occupation_Working Profes...	1.18
0	Do Not Email	1.13
5	Lead Source_Welingak Website	1.09
7	Last Activity_Other_Activity	1.01

9 Models were created. Few feature was eliminated based on statistical and VIF factor. Left fig shows the features included in final model.

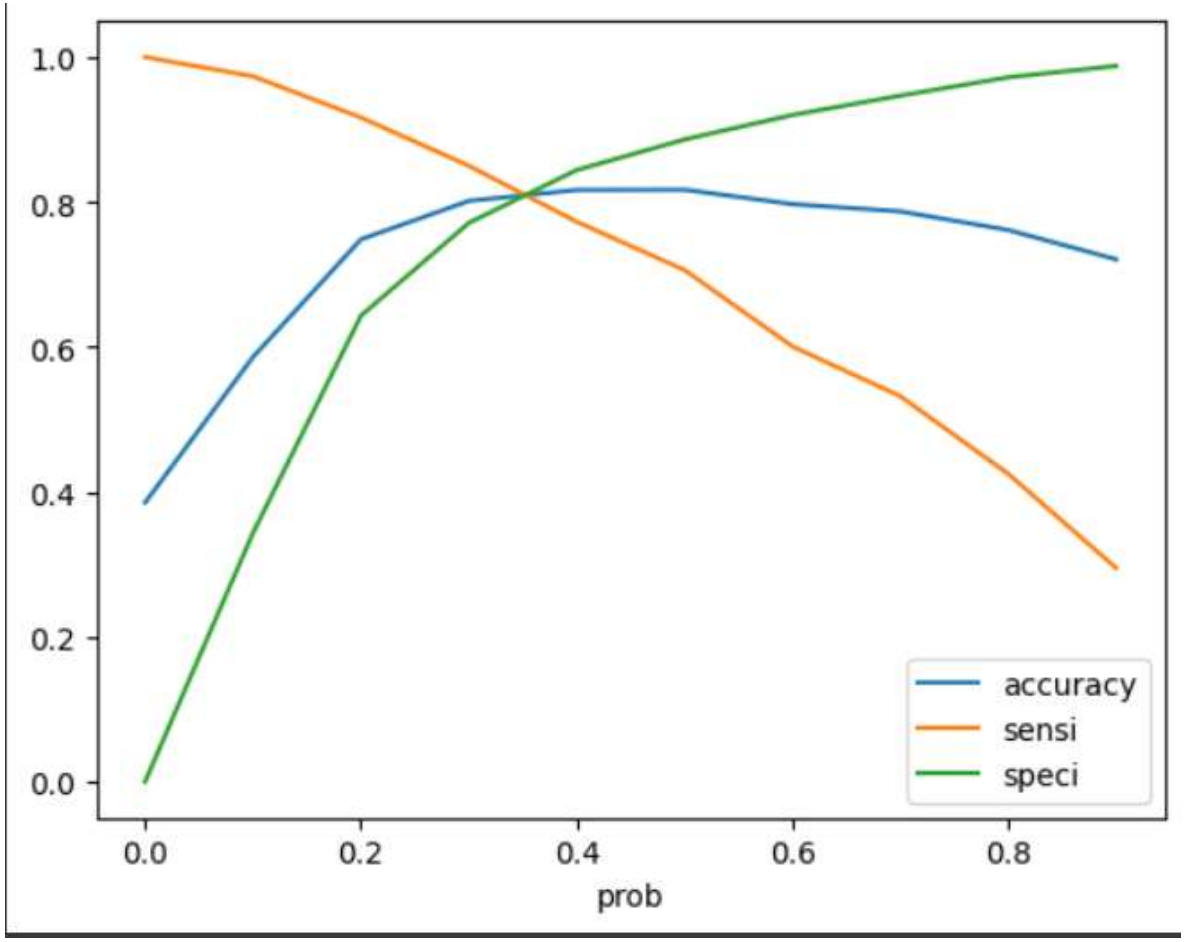
Choosing cut-off probability

- While choosing cut-off probability as 0.5 we observed that We found out that our specificity was good (~88%) but our sensitivity was only 70%.
- We have got sensitivity of 70% and this was mainly because of the cut-off point of 0.5 that we had arbitrarily chosen. Now, this cut-off point had to be optimised in order to get a decent value of sensitivity and for this we will use the ROC curve.

Since we have higher (0.89) area under the ROC curve , therefore our model is a good one.



Optimum Point



From the curve above, 0.34 is the optimum point to take it as a cutoff probability.

Results:

1) Comparing the values obtained for Train & Test:

Train Data:

- Accuracy : 81.0 %
- Sensitivity : 81.7 %
- Specificity : 80.6 %

- Test Data:
  - Accuracy : 80.4 %
  - Sensitivity : 80.4 %
  - Specificity : 80.5 %
- Thus we have achieved our goal of getting a ballpark of the target lead conversion rate to be around 80% . The Model seems to predict the Conversion Rate very well and we should be able to give the CEO confidence in making good calls based on this model to get a higher lead conversion rate of 80%.

Recommendation

- The company should make calls to the leads coming from the lead sources "Welingak Websites" and "Reference" as these are more likely to get converted.
- The company should make calls to the leads who are the "working professionals" as they are more likely to get converted.
- The company should make calls to the leads who spent "more time on the websites" as these are more likely to get converted.
- The company should make calls to the leads coming from the lead sources "Olark Chat" as these are more likely to get converted.
- The company should make calls to the leads whose last activity was SMS Sent as they are more likely to get converted.
- The company should not make calls to the leads whose last activity was "Olark Chat Conversation" as they are not likely to get converted.

- The company should not make calls to the leads whose lead origin is "Landing Page Submission" as they are not likely to get converted.