

Summary

This analysis is done for X Education and to find ways to get more industry professionals to join their courses. The basic data provided gave us a lot of information about how the potential customers visit the site, the time they spend there, how they reached the site and the conversion rate. The following are the steps used:

Data Cleaning:

- '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 replaced 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.

EDA:

- 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 website 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 welinkak website is high.

Data Prepration & Model Building:

- 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.

Model Evaluation:

- 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%.
- Optimal probability 0.34 is chosen based on accuracy sensitivity and specificity.
- 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.

