

Summary

Objective -

The 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. X Education needs help to select the most promising leads, i.e. the leads that are most likely to convert into paying customers. We are required 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 target lead conversion rate required is to be around 80%.

Process Followed -

1. **Data Cleaning** - The data had some columns having more than 35 percent of null values. So, removed those columns and the ones which had fewer null values were imputed using the measures of central tendencies. Columns such as Country had value India for 96 percent of the rows. And in order to avoid biases, drop that as well.
2. **EDA** - An EDA was performed for both categorical and numerical variables wherein univariate and bivariate analysis was performed. And Columns such as TotalVisits and Page Views Per Visit had outliers present in them so treated those values. 'TotalVisits' and 'Page Views per Visit' are highly correlated with correlation of 0.72
3. **Dummy Variables** - We created dummy variables for some of the categorical columns such as Lead Origin', 'Lead Source', 'Last Activity', 'Specialization', 'What is your current occupation' etc.
4. **Train Test Split** - The split was done at 70% and 30% for train and test data respectively.
5. **Model Building** - We used RFE to get the top 15 relevant variables. Later the rest of the variables were removed manually depending on the VIF values and p-value.
6. **Model Evaluation** - A confusion matrix was made. Later on the optimum cut off value (using ROC curve) was used to find the sensitivity and specificity which came to be around 80% each.
7. **Precision – Recall** - This method was also used to recheck Precision around 72% and recall around 79% on the test data set.
8. **Prediction** - Prediction was done on the test data with a cut off value of 0.35 with sensitivity and specificity of 80%.

Model Built -

```
Converted = 0.261843 + 3.17 * Lead Origin_Lead Add Form + 2.80
* Lead Source_Welingak Website + 2.54 * What is your current
occupation_Working Professional + 1.42 * Last Activity_Unsubscribed +
1.49 * Last Activity_Had a Phone Conversation + 1.29 * Last
Activity_SMS Sent + 1.10 * Lead Source_Olark Chat + 1.09 * Total Time
Spent on Website - 0.24 * Lead Origin_Landing Page Submission - 0.84
* Specialization_Hospitality Management - 1.26 * What is your current
occupation_Other - 1.31 * Last Activity_Olark Chat Conversation -
1.64 * Do Not Email
```

Recommendations -

To improve the potential lead conversion rate X-Education will have to -

1. **Lead Origin_Lead Add Form** - Leads through 'Lead Add Form' having higher conversion rate so maintaining focus on it to get more number of leads cause they have a higher chances of getting converted.
2. **Lead Source_Welingak Website** - Conversion rate is higher for the leads from 'Welingak Website',so company can focus on this website to get more number of potential leads.
3. **What is your current occupation_Working Professional** - The lead whose occupation is 'Working Professional' having higher lead conversion rate ,company should focus on working professionals and try to get more number of leads.
4. **Last Activity_Had a Phone Conversation, Last Activity_SMS Sent, Lead Source_Olark Chat** - These 3 features suggest that there is good chances of leads getting converted when they had been somehow in touch with company be it messages or chat or through a phone call. Hence, the work can be done upon these factors as well.
5. **Total Time Spent on Website** - Leads spending more time on website can be our potential converted leads. Hence, a more interactive user experience can be integrated in the website to attract a good amount of leads.
6. **Lead Origin_Landing Page Submission** - is affecting the leads not getting converted hence the improvements are needed on this part from the company as the same was evident in one of the above graphs.
7. **Specialization_Hospitality Management** - Leads have low conversion rate hence the company can check how it can cater to those kinds of professionals.