**Addressing conversion rate for an Education company X**

Education company X sells online courses to industry profiles. The company collects information of customers either by referrals or customers who have filled out the forms while consuming content online.

The company gets a lot of leads and has a very poor conversion rate, which basically means the Sales team effort is channeled to all the leads and finds it not to be very efficient.

Its implication is Company doesn’t focus on target customers who are highly likely to buy the course. The sales effort is directed toward the wrong customers and has no real impact on buying.

If it were focused on targeting potential customers, the Sales team could prioritize and spend further time explaining the course contents to convert customers.

**Goals of the Case Study:**

Build a model to assign a lead score between 0 and 100 to each of the leads so that salespersons could use it to target higher score leads. A higher score would mean that the lead is hot, i.e., Customers with a higher lead score have a higher conversion chance and customers with a lower lead score have a lower conversion chance.

Approach Followed:

To build a lead scoring model for Education company X, we followed a data-driven approach that utilizes customer lead data. Here's a step-by-step process followed:

1. Refer data dictionary to understand the data and the column details.

Conversion event: When a lead purchases a course defined as Converted in the dataset. Other variables like Total time spent on the Website, current occupation, Source of lead, recent activity interest in specialization, and their choice to be contacted by email, or call are captured in the column details for each prospect id.

1. Start with Data Preparation: Clean and preprocess the collected data.

Handle missing values- as a general approach, we removed columns having more than 20% null values.

Remove duplicates.

Format the data for analysis:

Convert Binary variables to 0/1

For categorical variables with multiple levels, create dummy features (one-hot encoded), and drop the original column.

Check for Outliers: There is an outlier on Total visits as we could see the max value and 99 percentile values are miles apart. We visualized the same for a better understanding. Remove Outliers by capping values at 99 percentiles.

1. Split Test and Train Data and assign Converted to response variable and other variables except for Prospect ID is assigned to Predictor variables.
2. Feature Scaling was required for 'TotalVisits','Total Time Spent on Website' & 'Page Views Per Visit' as these could dominate the contribution to the model. We used Standard Scaling to scale the numerical variables.
3. Perform Exploratory Data Analysis. We plotted the Correlation matrix for all variables to find multicollinearity between variables, as there was a huge corpus of data after the creation of dummy variables, we decided to handle the same with VIF.
4. Model Building: Build a baseline model with all columns we achieved 64% accuracy at this point

Feature selection: As we had around 71 columns, we proceed with feature reduction using RFE for coarse tuning of data to 15 columns.

Build Model with these 15 features compute all performance measures, the accuracy measure at this point was 79.6%. We decided to use the ROC curve at the end to validate the best model.

Fine-tuned model by removing high p value columns at first and running Model again

Accuracy measure remained the same at 79%

We used the ROC curve to select the optimal cutoff and found to have the best cutoff at 0.43 which maximizes AUC.

And built a final model with this cutoff and the accuracy measure at this point was 80% Recall 76% and Precision of 82%. For this case, Higher recall is desired as we don’t want to lose out on any potential lead. We chose recall to be an effective measure.

Make predictions on test and obtained a recall of 80%

**Final Outcome: Variables and their impacts are as below.**

Business recommendations can be derived from these data and is included on the presentation

'Do Not Email' negatively impacts the likelihood of conversion.

'Total Time Spent on Website ‘has a positive impact.

'Lead Origin Landing Page Submission' negatively impacts the likelihood of conversion

'Lead Origin Lead Add Form ‘has a positive impact

'Last Activity\_Converted to Lead'negatively impacts the likelihood of conversion

'Last Activity\_Email Bounced'negatively impacts the likelihood of conversion.

'Last Activity\_SMS Sent'has a positive impact

'What is your current occupation\_Working Professional',

'Last Notable Activity\_Email Bounced'has a positive impact

'Last Notable Activity\_Unreachable' has a positive impact.