SUMMARY – Lead Scoring Methodology

A leading online education company, X Education, sought to enhance its lead conversion rate. Despite a substantial volume of website visitors exploring their courses, only a fraction converted into paying customers. To address this challenge, X Education aimed to identify the most promising leads, dubbed "Hot Leads," prioritizing their sales team's efforts towards converting these high-potential prospects.

To achieve this objective, X Education implemented lead scoring, a methodology for ranking and prioritizing leads based on their conversion likelihood. The company developed a scoring model leveraging data such as website activity, form submissions, and referrals.

Data cleaning and Exploratory data analysis was conducted and on preliminary analysis of the data, a conversion rate of 37.85% was seen

Majority of the leads came from API and Landing Page submission while the conversion rate was only 30-35%. Small number of leads came from Lead Add Form however the conversion rate was significant with 90% of the leads converting. Almost no lead came or got converted from Lead import. A good strategy could be to look into getting more leads via Lead Add form to get a higher conversion.

Looking at the Lead sources it was discovered that majority of are generated by Google and Direct Traffic while the high conversion rate of leads came from Reference Leads and Welingak website

This suggested that it would be good to focus on lead conversion rate of Google and Direct Traffic while increase lead generation through Reference Leads and Welingak website as the conversion rate seem to be great there

Based on customer behavior spending more time on the website leads to higher conversion rate. People who have last activity as ‘Email opened’ or ‘sms sent’ seemed to have a higher conversion rate, with ‘sms sent’ having a significantly higher conversion as compared to no conversions.

Based on occupation working professionals are more likely to be converted as compared to unemployed. The unemployed have the second highest conversion rate

After the exploratory data analysis, the data was used to build a Logistic Regression Model. Feature selection was done by sieving data through RFE and VIF checks which concluded into a LRM model with 15 significant variables. On further analysis it was observed that **‘Lead Source_Welingak Website’**, **‘Lead Source_Reference’**, **‘What is your current occupation_Working Professional’** are the most contributing variables. This is also in concurrence with the Exploratory Data Analysis done previously hence building our confidence in the model being developed.

With an initial cutoff probability value of 0.5 labels were predicted for the train data. Confusion matrix was obtained and the sensitivity and specificity obtained were 70% and 88% resp. An area under ROC of 0.89 indicated good performance of the classification model. Further to accuracy, sensitivity and specificity were plotted for a range of cutoff probability to do the trade off and an optimal values of cutoff probability was identified as 0.34. This resulted in accuracy, sensitivity, and specificity of 80.84%, 82.62% and 79.72 on the training model and 80.05%, 81.39% and 72.29% on Test model.