Summary Report

The Problem statement defines an online courses selling company called ‘X Education’ wishing to identify the most potential leads. We started the process by importing and inspecting the dataset, Then we treated the missing values by calculating the percentage and dropped the columns having more than 45 percent. The columns which were generated by the sales team were found and dropped. Many of the categorical variables have a level called 'Select' were treated as a null value. The outliers for the numerical column were treated by the soft capping. From the Bivariate Analysis we gained few insights.

Data preparation was done by converting some binary variables (Yes/No) to 0/1 and creating dummy variables for the remaining categorical variables and dropping the level with big names. Model building was started by splitting the dataset in 70:30 ratio then Scaling was done for the numerical columns such as ['TotalVisits', 'Total Time Spent on Website', 'Page Views Per Visit']. The Converted Rate was found to be 38 percent were we can say that the dataset is pity much balanced. From the correlation matrix we dropped highly correlated dummy variables. Coming to Feature Selection RFE method was used to reduce to 15 variables and Multicollinearity was checked by using VIF method.

Model Evaluation was done to derive the optimal cut-off probability. Here we chose the ROC method and derived the value as 0.35. Then created a confusion matrix from the test dataset. From the matrix we calculated the percentage of sensitivity, specificity, Positive predictive value, Negative predictive value. Thus found the predictive capability as 80.09 percent