Q. A brief summary report in 500 words explaining how you proceeded with the assignment and the learnings that you gathered

Answer

The problem statement provided was around us trying to highlight the **Hot Leads** for X Education by assigning lead scores to the data provided. Looking at the data given, we initially found several missing values in several columns that needed to be treated. The technique used for this was:

- Removing rows and columns with over 30% of null values
- Imputation was done for 'Lead Source' column based on 'Lead Origin' column
- Dropping categorical columns with massive imbalance in the levels

This was then followed with exploratory data analysis to visualize the distribution of various features in the data using scatter plot and box plots. Also, bivariate and segmented univariate analysis was done to identify the levels within the categorical variables that contributed to most of the conversions in the data.

Post EDA, we performed outlier analysis on the numerical variables and removed the outlier data points that didn't contribute to conversions. Encoding was done for the categorical variables with 2 levels, and those with more than 2 variables (n), n-1 dummy variables were created.

Post this, the data was split into test and train (30-70 split) and scaling (min-max) was performed for the numerical variables to bring them to the same scale. We tried identifying any multicollinearity present in the data by plotting a correlation heat map.

After completing with the data preparation, we went ahead with model building on the train set. Logistic Regression was first applied on all the train data variables as we tried to understand what a non-optimized model would look like. Post that, we performed logistic regression with RFE to identify the 25 most significant variables using recursive feature elimination. Post that, we fine-tuned the model by eliminating a variable one-at-a-time based on p-values and VIF. Eventually, we were left with 19 significant variables in our final model.

Post this, we used our model for prediction on the train set itself. To determine the optimal cut-off, we plotted the sensitivity, specificity and accuracy for different probabilities (from 0-1). The optimal cut-off point (intersection of all the three plots) turned out to be 0.4.

We then ran the model to predict the probability of lead conversion on the test dataset and used the cut-off probability as 0.4 to determine whether the lead converts or not.

Post the prediction, we went ahead to evaluate the performance of our model. With an area under the ROC curve as 0.86, the model was pretty good. The accuracy of our model was 0.8 and a precision score of 0.77.

Finally, we ran the model on the entire dataset to assign a lead score to every data point, as was requested by the CEO of X Education. Post this, we analyzed our predictions to come up with a number of leads that the company should reach out to in order to ensure 80% of lead conversions.