“LEAD SCORING CASE STUDY SUMMARY “

Problem Statement:

X Education provides online courses to industry professionals and wants to improve its lead conversion rate. The company aims to develop a lead scoring model to identify the most promising leads—those with a high likelihood of converting into paying customers. The goal is to assign a lead score to each lead, with higher scores indicating a higher probability of conversion.

The CEO has set a target lead conversion rate of approximately 80%. This means that leads with higher lead scores should have a higher chance of conversion, while those with lower scores are less likely to convert. By accurately predicting the conversion probability of each lead, X Education can prioritize its resources and focus its efforts on leads with the highest potential for conversion.

Solution Summary:

Step1: Reading and Understanding Data:

Read and inspected the data.

Step2: Data Cleaning:

a. Initially, we removed variables with unique values as they didn't offer any useful information for analysis or prediction.

b. Subsequently, we identified columns with 'Select' values, indicating that leads hadn't chosen any option. We replaced these values with nulls to ensure consistency in the dataset.

c. Following that, we dropped columns with a null value percentage exceeding 35%, aiming to maintain data integrity and reliability.

d. We then addressed imbalanced and redundant variables by imputing missing values with the median for numerical variables and creating new categorical variables where necessary. Outliers were also identified and removed. Additionally, we resolved issues such as identical labels with different cases by standardizing them to ensure uniformity.

e. To streamline the analysis, all variables generated by the sales team were eliminated to prevent confusion and ensure clarity in the final solution.

Step3: Data Transformation:

Changed the binary variables into ‘0’ and ‘1’

Step4: Dummy Variables Creation:

a. We transformed categorical variables into dummy variables to facilitate modeling and analysis.

b. We eliminated duplicated and redundant variables to streamline the dataset and enhance its clarity and efficiency.

Step5: Test Train Split:

The next step was to divide the data set into test and train sections with a proportion of 70- 30% values.

Step6: Feature Rescaling:

a. We used the Min Max Scaling to scale the original numerical variables.

b. Then, we plot the heat-map to check the correlations among the variables.

c. Dropped the highly correlated dummy variables.

Step7: Model Building:

a. Using the Recursive Feature Elimination, we went ahead and selected the 15 top important features.

b. Using the statistics generated, we recursively tried looking at the P-values in order to select the most significant values that should be present and dropped the insignificant values**.**

c. Finally, we arrived at the 11 most significant variables. The VIF's for these variables were also found to be good.

d. For our final model we checked the optimal probability cut off by finding points and checking the accuracy, sensitivity and specificity.

e. We then plot the ROC curve for the features and the curve came out be pretty decent with an area coverage of 86% which further solidified the of the model**.**

f. Then, checked if 80% cases are correctly predicted based on the converted column.

g. We checked the precision and recall with accuracy, sensitivity and specificity for our final model on train set.

h. Next, Based on the Precision and Recall trade-off, we got a cut off value of approximately 0.3**.**

i. Then we implemented the learning to the test model and calculated the conversion probability based on the Sensitivity and Specificity metrics and found out the accuracy value to be 77.52%%; Sensitivity= 83.01%; Specificity= 74.13%.

Step 8: Conclusion:

• The lead score calculated in the test set of data shows the conversion rate of 83% on the final predicted model which clearly meets the expectation of CEO has given a ballpark of the target lead conversion rate to be around 80%.

• Good value of sensitivity of our model will help to select the most promising leads.

• Features which contribute more towards the probability of a lead getting converted are:

i. Lead Origin_Lead Add Form

ii. What is your current occupation_Working Professional

iii. Total Time Spent on Website