Lead Scoring Case Study Summary

Problem Statement:

X Education sells online courses to industry professionals. X Education needs help selecting the most promising leads, i.e., the leads most likely to convert into paying customers.

The company needs a model wherein a lead score is assigned to each of the leads such that the customers with higher lead scores have a higher conversion chance and the customers with lower lead scores have a lower conversion chance.

The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Summary of Solution:

1. The first step involved reading and understanding the data. The data was analyzed to get a better understanding of its contents.
2. The second step was data cleaning. This involved dropping variables that had a high percentage of NULL values. Missing values were imputed with median values for numerical variables and new classification variables were created for categorical variables. Outliers were identified and removed to improve the quality of the data.
3. The third step was data analysis. Exploratory Data Analysis was conducted to get a feel for how the data was oriented. During this step, three variables were identified that had only one value in all rows. These variables were dropped as they did not provide any useful information.
4. The fourth step involved creating dummy variables for categorical variables. This allowed for easier analysis of the data.
5. The fifth step was to split the data set into test and train sections with a proportion of 70-30%. This allowed for the model to be trained on one section of the data and tested on another.

6. The sixth step involved feature rescaling. Min Max Scaling was used to scale the original numerical variables. An initial model was then created using the stats model, which provided a complete statistical view of all the parameters of the model.
7. The seventh step involved feature selection using Recursive Feature Elimination (RFE). This method was used to select the top 20 most important features. The statistics generated were then used to recursively select the most significant values and drop insignificant values.
8. A data frame with converted probability values was created in the eighth step. An initial assumption was made that a probability value greater than 0.5 means 1, else 0.
9. In the ninth step, Confusion Metrics were derived and the overall Accuracy, Sensitivity, and Specificity of the model were calculated.
10. The tenth step involved plotting the ROC curve for the features. The curve came out to be pretty decent with an area coverage of 89%, which further solidified the reliability of the model.
11. In the eleventh step, the optimal probability cutoff point was found by plotting the probability graph for Accuracy, Sensitivity, and Specificity for different probability values. The intersecting point of these graphs was considered as the optimal probability cutoff point, which was found to be 0.37.

    Based on this new value, it was observed that close to 80% of values were correctly predicted by the model with an accuracy of 81%, sensitivity of 79.8%, and specificity of 81.9%.

12. In the twelfth step, the lead score was calculated and it was found that the final predicted variables gave a target lead prediction of approximately 80%.
13. In the thirteenth step, Precision and Recall metrics were calculated on the train data set with values of 79% and 70.5%, respectively.

14. Based on the Precision and Recall tradeoff, a cutoff value of approximately 0.42 was obtained in the fourteenth step.
15. In the fifteenth step, learnings from previous steps were implemented on the test model and conversion probability was calculated based on Sensitivity and Specificity metrics with an accuracy value of 80.8%, Sensitivity of 78.5%, and Specificity of 82.2%.