LEAD SCORING CASE STUDY SUMMARY

Problem Statement:

X Education sells online courses to industry professionals, the company wishes to identify the most potential leads.

The company wants to build a model with lead scores assigned to it, the Higher the lead score means higher chances of conversion and vice versa.

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

Solution/Approach:

Step 1: Reading and Understanding Data

Reading and understanding the data is first and foremost step as it gives proper idea about the different columns in given data and enable us to take better decisions through out model building.

Step 2: EDA and Data Cleaning

Data cleaning plays a vital role for better model building. We have done it in different steps like:

- To treat missing values we either dropped or imputed with mode, median or Other value
- Dropping the columns with high percentage of imbalance. Since the imbalance leads to biased decisions/understanding from Model side.
- Created new classes by merging existing classes in Categorical columns based on their nature and frequency to reduce the no. of dummy variables in further steps.

Step 3: Data Visualization

We have used different plots like count plot and heatmaps to get the feel of how data is oriented. Through count plots we can find How many leads were converted and how many were not across different classes.

Step 4: Creating Dummy Variables

We went with creating dummy variables for Categorical Columns in the given Dataset, we have also dropped the first column during dummy variable creation for Categorical Columns to reduce the Multicollinearity in model building.

Step 5: Train and Test Split

Later, we had split the data into Train and Test sections with 70% - 30% proportions.

Step 6: Rescaling

We have used Minmax Scaling technique to scale the Numerical Variables in data then built our initial model using Stats Model to get a complete statistical view of all parameters.

Step 7: Feature selection using RFE.

Using the Recursive Feature Elimination technique we have selected 20 topmost features. Then we looked at P-Values to keep the significant columns by dropping insignificant columns for a better performing model.

Then we checked the VIF scores of the better-known model to make sure all the VIF values are good and within limits.

We created a data frame with Converted Probability column with an initial assumption of Probability Cut-off point 0.5

Step 8: Plotting ROC Curve

We plotted the ROC curve for the remaining features and found the curve is decent with 89% area coverage.

Step 9: Finding optimal Cut-off point.

Then we plotted the probability graph of 'Accuracy', 'Sensitivity' and 'Specificity' at different probability values.

The intersection point of the lines would be considered as Optimal cut-off point and it was found out to be 0.38, based on it we found that near to 80% of the values were correctly predicted by the model.

Also, we have calculated the lead score based on the converted probability values.

Step 10: Computing and checking Precision and Recall metrics

We calculated precision and recall values and found out to be 73%, 79% respectively.

Based on the precision and recall curve, we got the cut-off value at 0.41

Step 11: Making Predictions on Test Data set

Then finally we implemented the Model on Test Data set and found out that the accuracy value is 80.26%, sensitivity 85.38% and specificity is 76.92%