Lead Scoring Case Study Summary

PROBLEM STATEMENT:

An education company named X Education sells online courses to industry professionals. Company wants to know out of the potential customer who are visiting the website how many out of them are hot lead who can be converted into paying customers.

The company needs a model where hot lead is being generalised based on the scoring , higher the score higher the chance of lead conversion & lower the score less the chance of conversion.

Currently lead conversion rate at X education is around 30%. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

SOLUTION & EXPLANATION:

- READING AND UDERSTANDING DATA:
  Most important step is to read and understand data by checking data dictionary and by analyse the data summary.

- DATA CLEANING:
  Dropping variables with high percentage of missing value and handing missing value by imputing median for numeric variables and mode for categorical variables. Also checking and treating outliers by capping and flooring technique.

- EXPLORATORY DATA ANALYSIS:
  Here visualization plays and important role , plotted count plot and boxplot for better insights came across that many leads are reaching website from google platform so it's better to invest on it other than different social media platform.

- CREATION OF DUMMY VARIABLE:
  Created dummy variables for categorical columns

- TRAIN AND TEST DATA SPLIT:

  Divided the data into train and test set with a proportion of 70% - 30%.

- FEATURE RESCALING:

  Used Standard Scaler method for scaling numerical variables then created first model using stats lib, which gave complete statistical view of all parameters of model.

- FEATURE SELECTION USING RFE METHOD:

  Using RFE selected top 18 feature then checked significance of variable by p-value and multicollinearity between input variable by VIF method after dropping the insignificant variable arrived at 13 most significant variables.

  For probability arbitrary cut-off 0.5 were taken and based on it divided the data into 0 and 1 section.

  Based of above cut-off confusion metric were generated which further helped in generating accuracy, sensitivity and specificity to understand how reliable the model is.

- ROC CURVE:

  Then plotted the ROC Curve on TPR and FPR got AUC of 96% which indicates the goodness of model.

- FINDIND OPTIMAL CUT-OFF:

  Optimal cut-off was found by plotting Accuracy, Sensitivity, specificity on different probability value. Optimal cut-off will be the point of intersection which was 0.3. New value of Accuracy is 89.83%, sensitivity is 88.92% and specificity is 90.38%. Also created the lead score according to the probability values.

  Computed Precision and Recall which came out to be 84.98% and 88.92% on train set. Based on Precision & Recall trade off we got same 0.3 cut-off.

  Also, prediction of conversion rate increases from 30% to 88% on train set and 90% on test set

- PREDICTION ON TEST SET:

  Implemented the learnings to the test model and calculated the conversion probability based on the Accuracy, Sensitivity and specificity which came out to be 90.04%, 90.01% & 90.05%