Summary Report

Lead Scoring Case Study

Team Members: Poojitha Parupally, Nirajkumar Pithva, Prajwala

1) We started with understanding problem statement, desired outcome, and data dictionary.
2) Data understanding and Missing value treatment:
   It is very important that we start with understanding the data and address the missing values as it may otherwise give skewed results and weak model.
   a) There are many variables which have level 'Select'; these are replaced with null because it indicates no option was selected by the people.
   b) The variables with missing values > 40% was deleted.
   c) The columns with relatively high missing values and mostly one unique value was also dropped.
   d) The columns with very few missing values were imputed with mode.
   e) The specialization selected is evenly distributed. Dropping is not a good choice. Created another category as 'Others' to replace.
3) Outlier analysis was done for the numerical columns. TotalVisits","Page Views Per Visit": Both these variables found to have outliers and these variables have been treated to remove outliers.
4) Once, the data is cleaned, we carried out Univariate Analysis and Bivariate Analysis on categorical columns using bar plot, scatter plot and heat map. This gave many important insights.
5) Data Preparation
   Before performing Logistic Regression, the data was prepared in following way.
   a) Dummy variables were created for the categorical columns and the categorical variables were dropped to eliminate redundancy.
   b) The given data was split into Train and Test with 70% and 30% proportion respectively.
   c) Feature scaling was done using Standard Scaler.
6) Model Building
   a) For feature selection, we used Recursive Feature Elimination (RFE) and selected top 15 variables.
   b) The first Logistic Regression model was built using these 15 variables.
   c) The insignificant variables with p-values > 0.05 were removed one by one resulting in model 4. This was also showing VIF under 5, suggesting very low multi-collinearity. Hence this model 4 is finalized.
7) Model evaluation
   a) y values predicted for train dataset using model 4.
   b) The confusion matrix was obtained considering Threshold value of Probability at 0.5 to start with. Based on the same accuracy, sensitivity and specificity were determined. Here, low sensitivity and high specificity indicated the need to find optimal cut-off value.
   c) Before finding optimal cut-off value, ROC curve was plotted giving AUC of 0.88, which is considered to be very good value.
   d) To find optimal cut-off, we plotted accuracy, sensitivity, and specificity for various probabilities. The intersection was found to be at 0.345 suggesting optimal cut-off.

e) y values predicted again for train dataset using model 4 and cut-off probability of 0.345. The Lead score was calculated for each lead in train dataset. Also, confusion matrix was obtained. Based on this, accuracy, sensitivity and specificity were determined and found to be near 80% suggesting the good model.

f) Model 4 was used on test dataset for y value prediction and Lead score was calculated for each lead. Confusion matrix was obtained for this test data set. Here also, accuracy, sensitivity and specificity found to be near 80%.

g) These close values of evaluation metrics from train and test dataset suggest this to be a very good model.

____________***