Lead Scoring Case Study | Summary

Data analysis and model building exercise has been done for X Education to convert more industry professionals and make them join their courses.

Following steps were taken to build a LR model that assigns lead score between 0 and 100 to each lead that can be used by company to target potential leads –

1. **Understanding & Cleaning Data Set:**
   Data was inspected and cleaned by dropping columns having one unique value and removing rows for a particular column having high missing values.

2. **Data Preparation & EDA:**
   Next imputation was done for Null values (with Median for continuous and Mode for a categorical variables). For the drop-down option 'Select' values were handled by converting to Null as they did not give much useful information. All Nulls were binned to 'Unknown' category to not lose much data. 'Unknown' level was later removed during dummy encoding. Outlier treatment was done - outliers present in some columns were finally removed basis IQR analysis. Quick EDA was done as part of data preparation to understand data and accordingly prepare it for model building.

3. **Dummy Variables:**
   Binary variables (Yes/No) were converted to 0/1. For categorical variables with multiple levels, dummy features (one-hot encoded) were created.

4. **Train-Test Split:**
   Original dataframe split into Train (70%) and Test (30%) dataset. Train dataset was used to train model and test dataset was used to evaluate model.

5. **Model Building:**
   To initiate model building RFE was done to attain top 20 feature variables. Remaining 4 features were dropped manually on an iterative basis post reviewing VIF values and p-values - variables with VIF <5 and p-values <0.05 were retained.

6. **Model Evaluation:**
   A confusion matrix was generated. Later optimal threshold probability cut-off was determined. Basis cut-off of 0.34 Accuracy, Sensitivity, Specificity was calculated which came to around 90% each.

7. **Precision-Recall Trade-off:**
   This method was also used to check cut-off and an optimal threshold value of 0.39 was derived with F1 score of 0.873. However, given business requirement for 80% lead conversion rate proceeded with 0.34 cut-off from previous step.

8. **Prediction on Test Dataset:**
   Prediction was done on test dataset using optimum cut-off of 0.34 with Accuracy, Sensitivity, Specificity derived which came to around 90% each.

Finally based on model, some features are identified that X education can consider to improve success of Lead conversion. Conversion probability of a lead increases with increase in values of following features in descending order:

1. Tags_Lost to EINS
2. Tags_Closed by Horizzon
3. Tags_Will revert after reading the email
4. Lead Source_Welingak Website

5. Last Activity_SMS Sent
6. What is your current occupation_Working Professional
7. What is your current occupation_Unemployed

Conversely, conversion probability of a lead increases with decrease in values of following features in descending order:

1. Tags_switched off
2. Tags_Ringing
3. Tags_Already a student
4. Tags_Not doing further education
5. Lead Quality_Worst
6. Tags_opp hangup
7. Tags_Interested in full time MBA
8. Tags_Interested in other courses
9. Asymmetrique Activity Index_03.Low