Lead Scoring Case Study

Summary:

We have completed this case study for X Education to find potential customers to join their courses. The dataset includes information on how the customers are visiting their website, spending time, and providing some details.

The company wants to find those potential leads with the highest lead scores so that they can make calls efficiently and improve the conversion rate significantly.

This process involves the following steps:

1. Reading and Understand the data
    - The given file is imported into a pandas data frame and required inspections have been done here to check information on rows, columns, data types, and some summary statistics.

2. Data Cleansing
    - The data frame is further inspected to check for the null values and fixing them. The columns which have more than 40% of null values have been dropped from the dataset.
    - New categorical values like “UnKnown” and “Others” are created for the null values in categorical features and used median values to impute the null values in numerical columns.

3. Exploratory Data Analysis (EDA)
    - Uni-Variate and Bi-Variate analysis is done here using the visualizations of distribution plots and count plots.
    - Outliers are present in the numerical columns, so dropped those rows since outliers affect the efficiency of the regression model.

4. Data Preparation
    - Few variables are dropped which are irrelevant to use in the model. New categorical values are imputed for some categorical features that have skewed data. All the binary columns that have “Yes” or “No” have been replaced with 1 and 0.
    - Created the dummy variables for all the categorical features, and rescaled the numerical features using the standard scaler.
    - The main data set is split into train and test data sets, and dependent and independent variables are separated and stored in different data frames.
5. Model Building
    - We used RFE to select the top 20 features and then manually removed other features one after other with reference to their p-values and VIF values.
    - After 12 iterations with different models created, finalized a model that remaining with 9 columns and having low p-values and VIF values.

- Created the predictions from the final model at the initial cut-off of 0.50

6. Model Evaluation
    - Following are the evaluation metrics for the final model at the cut-off of 0.5
      accuracy= 80.4%
      sensitivity = 66.1%
      specificity = 89.1%
    - Calculated the optimal cut-off point as 0.28, using accuracy, sensitivity, and specificity values at different probabilities, and the following are the metrics at the latest cut-off 0.28
      accuracy = 75.6%
      sensitivity = 84.5%
      specificity = 70.2%

      The conversion rate is at 84.45% on the train set.

7. Prediction on Test set and calculating the lead scores
    - Finally calculated the predicted values on the test set and calculated evaluation metrics
      accuracy = 76.6%
      sensitivity = 86.2%
      specificity = 70.8%
      The conversion rate is at 86.21% on the test set.

    - And, also calculated other metrics like Precision and Recall to check the relevance of the model.
8. Conclusions
   The following are the conclusions made out of this model
    - The Sensitivity, Specificity scores on the test set are very good and are closer to the train set scores at the optimal cut-off point
    - Hence the model is very good and in business terms, this model can adjust with the company's requirements in the coming future
    - The conversion rate on the final predicted model is greater than 80% on both train and test sets, which is the target given by the CEO of the company
    - The topmost 3 important features in the final model are
      1. Lead Source_Welingak Website
      2. Lead Origin_Lead Add Form
      3. What is your current occupation_Working Professional