Lead Scoring Case Study

This analysis is done for X Education to increase their business by getting industry professionals to join their courses and increase their revenue.

This is done by analyzing potential customers attributes. For example -who visit the site, the time they spend there, lead's source and the conversion rate etc.

The following are the steps used:

Data Cleaning:

As part of Data Cleaning, we started with handling missing values -

1. **Handling Missing Values:**
    a. **Explicit Missing Values:**
        i. There are columns with more than 30% of missing values. Identified all such columns and dropped them.
        ii. For all the columns with less than 30% of missing data, dropped all such rows with missing / NA values.
    b. **Implicit Missing Values:** There are a few columns where the option was not selected by the user. The column value is "Select".
        i. Dropped all such rows which has the value as "Select".

2. **Class Imbalance:** There are a couple of categorical columns with a huge imbalance in the data. Dropped these columns as these could impact the model.
    a. For Example: - The variable 'What matters most to you in choosing a course' has the value 'Better Career Prospects' 99.9% of the times.
    b. Since this could impact the model performance, dropped this.

Exploratory Data Analysis: Performed a quick EDA on the dataset to understand the following aspects :

1. Collinearity in the data
2. **Outlier Treatment:**
    a. The columns "TotalVisits", "Page Views per visit" have outlier values.
    b. Outliers in a logistic model are very sensitive, and could impact the model performance.
    c. Hence, Removing 0.1% of the values from the upper quartile region.

Data Transformation:

1. Binary Variables: Converted all the binary columns to 0/ 1 encoding.

2. Categorical Variables: For all the categorical variables, created dummy variables, and dropped the first column.

a. For e.g.: If there are n categories in a variable, this procedure gives n-1 variables.

Data Preparation:

1. Split the dataset into train and test dataset in the ratio 70:30.
2. Data is scaled using the minmax scaler.

Model Building:

1. Firstly, RFE was done to attain the top 15 relevant variables.
2. Later the rest of the variables were removed manually depending on the VIF values and p-value (The variables with VIF < 5 and p-value < 0.05 were kept).
3. The step 2 is repeated until all the p-values and VIF values are in range.

Model Evaluation:

1. Using the model built in the previous step, Y values are predicted
2. Using threshold as 0.5, y-values are categorized into 0 or 1.
3. Created a Confusion matrix.
4. Using these values, we have evaluated the accuracy, sensitivity and specificity which came to be around 79%, 76, 80% respectively

The threshold of 0.5 was chosen randomly to check the model performance. But in order to get good results, we need to optimize the threshold by plotting an ROC curve to see what the (area under the curve) AUC is.

The area under the curve of the ROC is 0.86 which is quite good. So, we seem to have a good model.

Finding the optimal Threshold Value:

1. Sensitivity-Specificity Trade-Off:
    a. Using different values of probability cut-offs, we have computed the values of accuracy, sensitivity, and specificity and stored the values in a data frame.
    b. After plotting, we see that around 0.42 there are optimal values of the three metrics.
2. Precision-Recall Trade-Off:
    a. We checked the precision and recall along with accuracy, sensitivity and specificity for our final model after changing the threshold value and checked the tradeoffs.

Evaluating Model Performance on the Test Data Set:

1. Scaled the test data using the scaler fitted on the train data set.
2. Prediction was done on the test data frame and with an optimum cut off as 0.42.
3. accuracy, sensitivity and specificity of the model is ~ 80%.

We found the score of accuracy and sensitivity from our final test model are in acceptable range.

Conclusion

1. Test set is having accuracy, recall/sensitivity in an acceptable range.

2. In business terms, our model is having stability and accuracy with adaptive environment skills.
3. Top features for good conversion rate:
    a. Lead Origin_Lead Add Form
    b. Total Time Spent on Website
    c. Last Notable Activity_Had a Phone Conversation
    d. What is your current occupation_Working Professional
    e. Last Notable Activity_Unreachable