LEAD SCORING CASE STUDY SUMMARY

Problem Statement

This is the case study for X Company to find the most prominent leads who can result in taking up their courses and be a payable customer. X Education sells online courses to industry professionals. X Education needs help in selecting the most promising leads, i.e., the leads that are most likely to convert into paying customers.

The company needs a model wherein you a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

Solution Approach & Steps Involved

Step 1: **Reading and Understanding Data**.

- Read and analyze the data.

Step2: **Data Cleaning**:

- We dropped the variables that had more than 40% of NULL values in them.
- Manipulating and Imputing data values wherever required like handling data having "Select" as values.

Step3: Exploratory Data Analysis

- We did Exploratory Data Analysis of the data set to get the details of how the data is oriented. In this step, there were we observed the importance of few features as well,

Step4: **Creating Dummy Variables**

- We went on with creating dummy data for the categorical variables.

Step5: **Test Train Split**

- The next step was to divide the data set into test and train sections with a proportion of 70-30% values.

Step6: Feature Rescaling

- We used the Min Max Scaling to scale the original numerical variables. Then using the stats model we created our initial model, which would give us a complete statistical view of all the parameters of our model.

Step7: **Feature selection using RFE**

- Using the Recursive Feature Elimination, we went ahead and selected the 20 top important features.
- Using the statistics generated, we recursively tried looking at the P-values in order to select the most significant features that should be present and dropped the insignificant values.
- We are dropping features having higher than 5 VIF and p-values higher than 0.05
- Finally, we arrived at the 9 most significant variables. The VIF's for these variables were also found to be good.
- Based on the above assumption, we derived the Confusion Metrics and calculated the overall Accuracy of the model.

We also calculated the '**Sensitivity**' and the '**Specificity**' matrices to understand how reliable the model is.

Step8: Plotting the ROC Curve

- We then tried plotting the ROC curve for the features and the curve came out be pretty decent with an area coverage of 83% which further solidified the of the model.

Step9: Finding the Optimal Cutoff Point

- We plotted the probability graph for the 'Accuracy', 'Sensitivity', and 'Specificity' for different probability values. The intersecting point of the graphs was considered as the optimal probability cutoff point. The cutoff point was found out to be 0.32

Step10: Computing the Precision and Recall metrics

- We also found out the Precision and Recall metrics values came out to be 73% and 72% respectively on the train data set.

Based on the Precision and Recall tradeoff, we got a cut off value of approximately 0.38

Step11: **Making Predictions on Test Set**

- We implemented the learnings to the test model and calculated the conversion probability based on the Sensitivity and Specificity metrics and found out the accuracy value to be 79%

Impacting Variable for our Conversion Rate

There are a few features which are most likely to be considered in order to get the prominent leads. The following insights were drawn after doing analysis on the given dataset :

1. **Total Time Spent on Website :** The people who spent more time on their website can be the prominent or probable positive leads. They might be either collecting some course information, checking the courses or might already be interested in some course.

2. **Total Number of Visits :** The leads who are most frequently visiting the website can be probably the positive leads.

3. **Lead Sources :** The people from following Lead Sources :

    - Google
    - Direct Traffic
    - Organic Search
    - Olark Chat Conversation
    - Referral Sites

4. **Lead Origins :** Lead Add Forms is the highest Lead converter among all the Lead origins.

5. The people who are **working professionals** are the most prominent Positive Leads.

6. **Last activity :** People whose Last activity is either Email Opened or SMS sent are possible Lead Converters.

7. **Last Notable Activity :** Enthusiasts with Last Notable Activity as SMS sent are also Lead Converters.

The above features must be considered in order to convert the most prominent leads and increase the sales as well for the X Company and get more people converted to a payable customer / learners.