LEAD SCORING CASE STUDY

Problem Statement

An education company named X Education sells online courses to industry professionals. X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers.

There are a lot of leads generated in the initial stage, but only a few of them come out as paying customers. The company requires you to build a model wherein you need to assign a lead score 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.

The CEO has given a ballpark of the target lead conversion rate to be around 90%.

Working Professionals going for the course have high chances of joining it.
Maximum number of leads are generated by Google and Direct traffic.
Conversion Rate of reference leads and leads through welingak website is high.
To improve overall lead conversion rate, focus should be on improving lead conversion of olark chat, organic search, direct traffic, and google leads and generate more leads from reference and welingak website.
API and Landing Page Submission bring higher number of leads as well as conversion.
In order to improve overall lead conversion rate, we have to improve lead converion of API and Landing Page Submission origin and generate more leads from Lead Add Form.
Leads spending more time on the website are more likely to be converted.
Website should be made more engaging to make leads spend more time.

Solution approach summary

1. Read and understand the data.
2. Data Cleaning
   We dropped the variables that had high percentage (> 45%) of NULL values in them. This step also included imputing the missing values as and where required with median values in case of numerical variables and creation of new classification variables in case of categorical variables. The outliers were identified and removed.
3. Data Analysis
   EDA is an approach of analysing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods. We have found 3 variable having unique data and drop from the data set.
4. Create Dummy Variables
   Dummy variable is one that takes only the value 0 or 1 to indicate the absence or presence of some categorical effect that may be expected to shift the outcome. We have around 10 dummy variables.
5. Train and Test We will split the data into 2 parts.
   - Train Data (On which model will be build and is almost 70% of total data).
   - Test Data (On which build model will be tested and is almost 30% of total data).
6. Feature Rescaling

StandardScaler removes the mean and scales each feature/variable to unit variance. This operation is performed feature-wise in an independent way. We have used StandardScaler for rescaling the numerical variables.

7. Feature selection using RFE.

   Using the Recursive Feature Elimination, we went ahead and selected the 15 top important features. Using the statistics generated, we recursively tried looking at the P-values in order to select the most significant values that should be present and dropped the insignificant values.

   Finally, we arrived at the 13 most significant variables. The VIF's for these variables were also found to be good.

8. Plotting the ROC Curve

   We then tried plotting the ROC curve for the features and The ROC Curve should be a value close to 1. We are getting a good value of 0.97 indicating a good predictive model.

9. Finding the Optimal Cut-off Point, Precision & Recall

   We can see above the model seems to be performing well. The ROC curve has a value of 0.97, which is very good. We have the following values for the Train Data:

   - Accuracy: 92.29%
   - Sensitivity: 91.70%
   - Specificity: 92.66%

   Some of the other Stats are derived below, indicating the False Positive Rate, Positive Predictive Value, Negative Predictive Values, Precision & Recall.

10. Prediction on Test Sets

    After running the model on the Test Data these are the figures we obtain:

    - Accuracy: 92.78%
    - Sensitivity: 91.98%
    - Specificity: 93.26%

11. Final Observation

    Train Data:

    - Accuracy: 92.29%
    - Sensitivity: 91.70%
    - Specificity: 92.66%

    Test Data:

    - Accuracy: 92.78%
    - Sensitivity: 91.98%
    - Specificity: 93.26%