Lead Scoring Case Study Summary

Problem Statement:

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

Goals of Case Study:

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.
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 a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.

Solution Summary:

Step1: Reading and Understanding Data

We have checked below points -

- Number of rows and columns in data set
- Data types of each column in data set
- Checked first few rows in data set
- Checked the statistical information of data set
- Checked the column names and description and tried to understand business concept

i) Missing value treatment:

- Checked for null values and imputed them with appropriate methods if columns have null values less than 40%. We used mode imputation for categorical columns and used median imputation for numerical columns.
- Dropped columns having null values more than 40%.

ii) Outlier detected and removed:

The outliers were identified for numerical columns and it has been removed. We have used IQR method to treat the outliers in the data set.

Step 2: Visualizing the data

i) Univariate analysis:

- Performed overall univariate analysis using bar plot for categorical column.
- Performed overall univariate analysis using histogram for continuous column.

ii) Bivariate analysis:

- Performed bivariate analysis w.r.t Target variable using count plot (categorical column vs categorical column).
- Performed bivariate analysis w.r.t Target variable using scatterplot (numerical column vs numerical column)
- Performed bivariate analysis w.r.t Target variable using bar plot (numerical column vs categorical column)

iii) Multivariate analysis:

- Performed multivariate analysis using heat map.

Step 3: Data Preparation

i) Dummy Variables Creation:

- We created dummy variables for the categorical variables.
- Removed all the repeated and redundant variables.

ii) Data Transformation: Changed the categorical binary variables into '0' and '1'. '0' means No and '1' means Yes.

Step 4: Splitting the Data into Training and Testing Sets

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

ii) Feature Scaling: - Rescaled the variables using MinMaxScaler as they have large values as compared to the other variables of the dataset.

Step 5: Logistic Model Building

a. Using the Recursive Feature Elimination, we went ahead and selected the 20 top important features.
b. Using the statistics generated, we recursively tried looking at the P-values in order to select the most significant values that should be present. We dropped the insignificant values which have P-value greater than 0.05
c. Finally, we arrived at the 12 most significant variables. The VIF's for these variables were also found to be good.
d. We then plot the ROC curve for the features.
e. For our final model we checked the optimal probability cut off by finding points and checking the accuracy, sensitivity and specificity. Cut off was taken as 0.3 for sensitivity and specificity trade off.
f. We checked the precision and recall also for our final model.
g. Next, based on the Precision and Recall trade-off, we got a cut off value of approximately 0.42
h. Then we implemented the learnings to the test model and calculated the conversion probability based on the sensitivity, specificity metrics and precision, recall metrics.

Step 6: Model evaluation and performance

Below is our final model performance:

Sensitivity, Specificity approach:

Train Data Set: Accuracy 90.79%, Sensitivity 90.90%, Specificity 90.73%

Test Data Set: Accuracy 90.47%, Sensitivity 90.22%, Specificity 90.63%

Precision, Recall approach:

Training Data Set: Accuracy 91.04%, Precision 88.07%, Recall 88.52%

Test Data Set: Accuracy 90.62%, Precision 88.25%, Recall 87.56%

Step 7: Conclusion:

• Sensitivity was calculated in the test set of data which is more than 90% for final predicted model which clearly meets the expectation of CEO has given a ballpark of the target lead conversion rate to be around 80%.
• Good value of sensitivity and recall of our model will help to select the most promising leads.
• Top 3 features which contribute more towards the probability of a lead getting converted are:

- **Total Time Spent on Website**
- **Lead Origin_Lead Add Form**
- **Lead Source_Welingak Website**