Lead Score Case Study _ Summary

The analysis is done for X education and to find ways to get more industrial professionals to join their courses. The basic data provided gave us a lot of information and how potential customers visit the site, time spent on site, how they reached the site and the conversion rate.

The steps we used here are:

1. **Data Cleaning** – it is very important step since Model efficiency directly depend on the data, we removed the columns having more than 39 % missing values, also converted “Select” as NAN since available details not selected by leads,few object columns having low frequent level so created a sperate category as Others. We input missing value with mode for Object, median for Numerical columns.

2. **EDA**: A quick EDA is done to check the condition of data. It is found that a lot of elements in categorical variables cannot contribute to our analyses hence dropped Found outliers in numerical columns they are solved using capping method.

3. **Data preparation** :Dummys are created for categorical variables with multiple levels, Binary columns encoded with 0/1 and Numerical columns are scaled using standardization.

4. **Train_test split**: The split is done at 70-30 %for train and test data respectively

5. **Checking the correlations**: We plot heatmap to check most correlated variables and dropping the highly correlated dummy variables.

6. **Model building**: Firstly, RFE was done to attain the top 20 most relevant variables. Later the rest of the variables are removed manually depending on the VIF values and p-values (we kept p=value<0.05 and VIF =< 5), at last we got final model with 14 Variables.

7. **Model Evaluation**: A confusion matrix is made. And ROC curve is plotted which shows under curve area as 0.88 means our model is good, we plot Sensitivity, Specificity and Accuracy tradeoff and found optimal cutoff threshold as 0.35, we calculated accuracy, sensitivity, specificity which is 78%, 81% &80% respectively.

8. **Prediction**: Final Prediction is done on data set and using optimal cut off as 0.35 with Accuracy, Sensitivity, Specificity of 80%.

9. **Precision and Recall**: Calculated recession and recall values which are 79% and 69% respectively

10. **Predictions on the test set**: We find that model scores of train and test data set for Accuracy, Sensitivity, Specificity are approximately closer to 80%.

11. **Recommendation**:

1) Leads from Welingak Website and Reference are potential leads for conversion
2) Hot leads calculated to contact where Lead Score >90, which will help to increase the conversion.
3) If a company had phone conversation with Leads as last activity such leads are potential leads for conversation.

12. **Conclusion**: Comparing the model metrics score on both train and test dataset, we can ensure the model is functioning efficiently to increase the lead conversion around 80%, and able to adjust with the future requirement if any changes.