Lead Score Case Study Summary

Problem Statement:

X Education sells online courses to industry professionals. X Education needs help to identify the most promising leads, i.e. the leads that are most likely to convert into paying customers.

Although X Education gets a lot of leads, its lead conversion rate is very poor. Only 30% of all the leads X Education are able to convert. The company requires 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. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Solution Approach:

1. **Reading and Understanding Data:**
   In this, we understood what the data is about, checked the shape of the dataset, and checked the information of the dataset i.e. data types, non-null values, etc. Then we checked the statistical summary of the data.

2. **Data Cleaning:**
   We dropped variables with more than 40% null values and analyzed some variables to check if they are useful in analysis and imputed them, if not better to drop them. We dropped columns with high data imbalance and having only 1 value in a column. This step also involves outlier analysis and capping those outliers.

3. **Univariate and Bivariate Analysis:**
   Then we started univariate and bivariate analysis of features. Checked if there is data imbalance in target variable, correlation heatmap and countplot of different features.

4. **Data Preparation:**
   We created dummy variables for categorical columns.

5. **Train Test Split:**
   Then we split the data into train and test set in 70:30 ratio.

6. **Feature Scaling:**
   Then we normalised numerical values using MinMaxScaler.

7. **Feature Selection using RFE:**
   There were 70 columns in total. We used RFE for selecting best 15 features for our model.

8. **Model Building:**
   Using statsmodel we created our model and checked for p-value and vif value. If p-value is greater than 0.05 and vif value greater than 5.0, we dropped those columns one by one in order to find the best fit model. Once p-value and vif values are under 0.05 and 5.0, we stopped at that point selected that as our final model.

9. **Model Evaluation and Finding Optimal Cut-off Point:**
   Once our model is final, we checked the accuracy first which was 81% and after that we plotted the ROC curve to check the area under the curve which comes out 0.89. For optimal cut-off point we used sensitivity-specificity and precision-recall.

   Through sensitivity-specificity cut-off point is 0.36 and other metrics:
   Accuracy - 80%
   Sensitivity - 79%
   Specificity - 81%
   False Positive Rate - 19%
   Positive Predictive Value - 72%
   Negative Predictive Value - 86%

   Through precision-recall cut-off point is 0.42 and other metrics:
   Accuracy – 81%
   Precision – 76%
   Recall – 75%

   After comparing result from both cut-off points, we chose which gives best results.

10. **Making Prediction on Test Set:**
    Using cut-off point of 0.36 we made prediction on test set and got these results:
    Accuracy - 81%
    Sensitivity - 78%
    Specificity - 82%
    False Positive Rate - 18%
    Positive Predictive Value - 71%
    Negative Predictive Value - 87%

11. **Conclusion:**
    In this step we assigned scores to the leads and found out the 2046 hot leads having score above 70 and below are the top 3 variables contributing most in lead conversion are:
    - Total Time Spent on Website
    - Lead Origin_Lead Add Form
    - What is your current occupation_Working Professional