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.

The company requires a model wherein we 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 Summary:

Step 1: Reading and Understanding data:

First step is to understand the features and its purpose. Understanding it from given data dictionary. Also explored the csv file to the different aspects of it.

Step 2: Importing the data into Data Frame and inspecting it:

- Imported related libraries to import csv file and read it from data frame.
- Analysing the data frame to check shape, datatypes and null values.

--EDA

Step 3. Data Preparation :

- Data Cleaning- dropped the features having null values more than 70% and features having unique values.
- Converted 'Select' value to null values as these are the data that leads didn't choose and it got populated by default.
- Imputed null columns with median value for numerical and mode for categorical features
- Grouped and renamed all categories that has insignificant leads into one group.
- Encoding Categorical variables (Yes-1, No-0)
- Capping the outliers to 95percentile for numeric features.
- Dummy variable creations from categorical features to feed numerical data to our logistic model.

Step 4: Test-Train Split:

- Then divide the data set into train and test sections with a proportion of 70-30% values.

Step 5. Model Building:

- Used Recursive Feature Elimination to select top 15 important features to feed in our first train model.
- Reviewed Statsmodel summary report and dropped the most insignificant feature by looking at p value >.05.
- After dropping 2 features we got 13 most significant variables. Their VIF were also found to be good
- Created ROC curve to get the AUC of our model. The area coverage of 95% which is very good.
- We also checked accuracy score, Sensitivity and Specificity of our model on train set.

```
Accuracy: 0.892114250232847
Sensitivity/TPR: 0.8780487804878049
Specificity: 0.9012820512820513
FPR: 0.0987
```

- Then we found the optimal Cutoff Point of .27 by using accuracy score, Sensitivity and Specificity

- We implemented the learnings to the test model and evaluated the score which came out pretty close to our train dataset

```
Accuracy: 0.8964518464880521
Sensitivity: 0.8909090909090909
Specificity: 0.899548532731377
FPR: .1005
```

Step 6. Conclusion:

- The Lead Score calculated in the test set of data shows the conversion rate of 89% on the final predicted model which clearly meets the expectation of CEO of 80% ballpark target.
- Good value of sensitivity of our model will help to select most promising leads.
- Features used in Final Model are:

```
['Lead Source_Welingak Website',
'Last Activity_SMS Sent',
'What is your current occupation_Working Professional',
'Tags_Busy',
'Tags_Closed by Horizzon',
'Tags_Lost to EINS',
'Tags_Ringing',
'Tags_Will revert after reading the email',
'Tags_invalid number',
'Tags_switched off',
'Lead Quality_Not Sure',
'Lead Quality_Worst',
'Last Notable Activity_Modified']
```