Lead case study summary

Problem Statement

We have outlined 2 problem statements. Both are defined hereunder.

1. Problem Statement 1: There is an X Education Company, that delivers a bunch of online courses to various domain-specific professionals. They perform marketing of their online course through various search engines and their own website. If someone visits their webpage for any course, their Sales Team becomes proactive and reaches them either via a call or an email. The company marketing team aims to create a main strategy to convert all website visitors into potential leads.

2. **Problem Statement 2:** To increase their revenue every quarter, they also need to sell maximum online courses to these potential leads and convert them into hot leads. So, they don't waste their marketing team resources on unnecessary calling and use them for better resourcefulness n other activities.

Process

The process to identify who can be treated as the potential lead is as defined hereunder :

1. Data Cleaning

a. After reading the data set, we found that there are many columns with NULL, skewed, and single values in the various column.

b. After identifying these columns, we considered dropping all those columns. We again performed a review for the value = 'Select' or 'NAN' in the remaining columns. Now, in this assessment, if we found values that are less than 10% of the total values, we again dropped the record but, not the whole columns.

2. Data Dumification

a. After cleaning the data set , we now created dummy variables for the categorical columns.

B. We found a catch and that was the 'Specialization' column. For this column, we created a dummy with a select value, and later on, we dropped column 'specialization_select'.

c. Again, later on, for all these categorical columns for which the dummy values were created, we again dropped all the original columns.

3. Performing the Train-Test Split

a. We performed splitting of the data into the train and test with a 70:30 ratio.

b. Then used the standard scaler for the columns 'Total Time Spent on Website', 'TotalVisits', and 'Page Views Per Visit'.

c. We then performed 'Feature selection' for building the Model using 'RFE'.

4. Building Model

a. We build our Model 1 using 'Logistic Regression' and checked 'VIF' values.

b. We found that a total of 3 columns has VIF > 5.

c. We then dropped one of the columns, and thereafter build a second Model.

d. In Model 2, we discovered that the 'VIF' of all remaining columns is within a range of (<5). Thus, now we studied their 'p-values'.

e. Again in Model 2, we found that there are 2 columns that have p-value > '0.05'.

f. So, now we dropped one of these columns. Again, with the remaining column, we build our 'Model 3'.

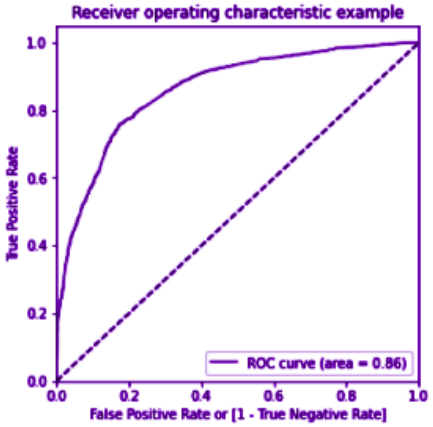
g. Now in Model 3, we saw that there is one column with a p-value > 0.05, so we removed this column and further build 'Model 4'.

h. The 'Model 4' has all VIF. Also, the p-value of columns is in the range. Thus, we now built y values for this model.

i. On the 'Model 4' data frame, we checked various factors such as the Precision, Accuracy, and Confusion Matrix. Also, compared converted and predicted_converted values along with the ROC curve, etc., and we found the below as a result.

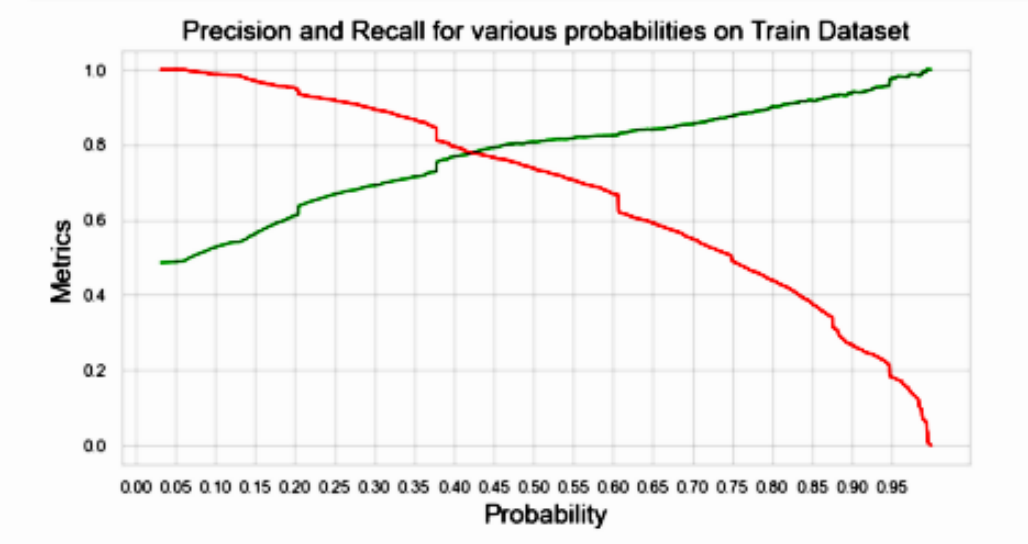
```
print(accuracy)
```

```
0.7894618834080718
```




```
Sensitivity: 0.7375058383932742
Specificity: 0.8374299266925399
FPR: 0.16257007330746012
PPR: 0.8072597137014315
NPR: 0.7755591054313099
```

j. Now, using all the chosen columns of our ‘Model 4’ on the test data, we finally calculated the test model's accuracy and precision. We found that the ‘Test model’ has an accuracy of 79% with a ROC cutoff as ‘0.42’.



5. Feature Analysis

Lastly, we performed some feature analysis, to identify important features responsible for converting all captured leads in the sales funnel into qualified and potential leads.