# Lead Scoring Case Study

**By**
**Nguyen Minh Khanh, Yuvraj Goswami & Mudassir Imam**

## Problem Statement:

As a part of the Lead Scoring case study, we have been presented with the details how the company X Education pursues customer leads from various sources and tries to convert them to potential customers. The current conversion rate is quite low at 30%. So we have been tasked to analyse the data and come up with a model which can make predictions to the order to 80% Lead conversion, which will help the company to focus more on communication with the potential leads rather than making calls to every customer.

## Analysis Approach:

1. Understanding Data
2. Data Cleaning:
    a. Handling Missing Values
3. Exploratory Data Analysis:
    a. Univariate Analysis
    b. Multivariate Analysis
4. Data Preparation:
    a. Dummy variable creation
    b. Split data into train and test sets
    c. Feature scaling

5. Model Building:
    a. Creating different models until p-value and VIF is normalized
6. Model Evaluation:
    a. Evaluating the final model using accuracy, sensitivity and specificity
    b. Plotting ROC curve
    c. Evaluating based on precision and recall
7. Making Predictions on Test Data:
    a. Evaluating the final model using accuracy, sensitivity and specificity
    b. Plotting ROC curve
    c. Evaluating based on precision and recall
8. Assigning Lead Score to Data

# Final Model

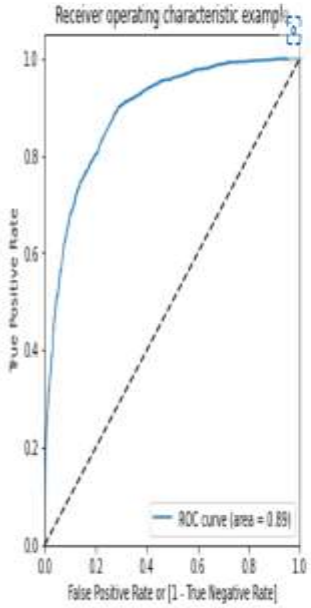
All the VIF values are good and all the p-values are below 0.05. So, we can proceed with making predictions using this model only

| | | | |
|---|---|---|---|
| **Dep. Variable:** | Converted | **No. Observations:** | 6468 |
| **Model:** | GLM | **Df Residuals:** | 6455 |
| **Model Family:** | Binomial | **Df Model:** | 12 |
| **Link Function:** | Logit | **Scale:** | 1.0000 |
| **Method:** | IRLS | **Log-Likelihood:** | -2637.5 |
| **Date:** | Sun, 19 Mar 2023 | **Deviance:** | 5274.9 |
| **Time:** | 19:39:31 | **Pearson chi2:** | 6.80e+03 |
| **No. Iterations:** | 7 | **Pseudo R-squ. (CS):** | 0.4059 |
| **Covariance Type:** | nonrobust | | |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **const** | -2.4640 | 0.100 | -24.693 | 0.000 | -2.660 | -2.268 |
| **TotalVisits** | 0.7479 | 0.154 | 4.863 | 0.000 | 0.446 | 1.049 |
| **Total Time Spent on Website** | 4.5251 | 0.166 | 27.339 | 0.000 | 4.201 | 4.849 |
| **Lead Origin_Lead Add Form** | 3.8446 | 0.206 | 18.662 | 0.000 | 3.441 | 4.248 |
| **Lead Source_Olark Chat** | 1.6343 | 0.121 | 13.487 | 0.000 | 1.397 | 1.872 |
| **Lead Source_Welingak Website** | 2.4312 | 1.028 | 2.364 | 0.018 | 0.416 | 4.447 |
| **Do Not Email_Yes** | -1.4827 | 0.166 | -8.913 | 0.000 | -1.809 | -1.157 |
| **Last Activity_Olark Chat Conversation** | -1.1466 | 0.160 | -7.182 | 0.000 | -1.459 | -0.834 |
| **Last Activity_SMS Sent** | 1.3721 | 0.075 | 18.386 | 0.000 | 1.226 | 1.518 |
| **What is your current occupation_Not provided** | -1.2886 | 0.088 | -14.702 | 0.000 | -1.460 | -1.117 |
| **What is your current occupation_Working Professional** | 2.4941 | 0.183 | 13.645 | 0.000 | 2.136 | 2.852 |
| **Last Notable Activity_Had a Phone Conversation** | 3.2761 | 1.183 | 2.768 | 0.006 | 0.957 | 5.596 |
| **Last Notable Activity_Unreachable** | 2.6148 | 0.695 | 3.760 | 0.000 | 1.252 | 3.978 |

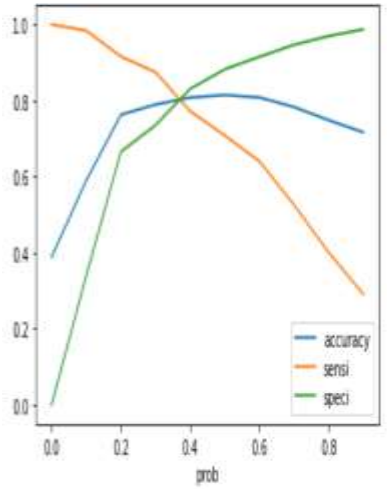
| | Features | VIF |
|---|---|---|
| 0 | TotalVisits | 2.09 |
| 1 | Total Time Spent on Website | 1.95 |
| 3 | Lead Source_Olark Chat | 1.57 |
| 8 | What is your current occupation_Not provided | 1.48 |
| 7 | Last Activity_SMS Sent | 1.44 |
| 6 | Last Activity_Olark Chat Conversation | 1.41 |
| 2 | Lead Origin_Lead Add Form | 1.40 |
| 4 | Lead Source_Welingak Website | 1.24 |
| 9 | What is your current occupation_Working Profes... | 1.20 |
| 5 | Do Not Email_Yes | 1.07 |
| 11 | Last Notable Activity_Unreachable | 1.01 |
| 10 | Last Notable Activity_Had a Phone Conversation | 1.00 |

# Model Evaluation Metrics



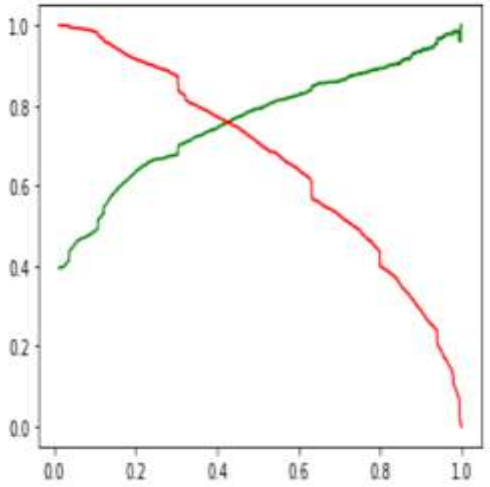

- The ROC Curve should be a value close to 1. Area under ROC is 89%, which indicates a good predictive model

**Sensitivity – Specificity:**




- From the curve above, 0.35 is the optimum point to take it as a cutoff probability.

**Precision – Recall**



- From the curve above, 0.40 is the optimum point to take is as the cutoff probability

1. **Sensitivity – Specificity:**

   a. **On Training Data:**

      - The optimum cut off value was found using ROC curve. The area under ROC curve was 0.89.
      - After Plotting we found that optimum cutoff was 0.35 which gave
        - Accuracy = 80.24%
        - Sensitivity = 79.91%
        - Specificity = 80.45%.

   b. **On Test Data:**
        - Accuracy 80.95%
        - Sensitivity 80.80%
        - Specificity 81.04%.

2. **Precision – Recall:**

   a. **On Training Data:**

      - With the cutoff of 0.35 we get the Precision & Recall of 79.25% & 70.80% respectively.
      - So, to increase the above percentage we need to change the cut off value. After plotting we found the optimum cut off value of 0.40 which gave

        - Accuracy 80.76%
        - Precision 74.32%
        - Recall 77.16%

   b. **On Test Data:**

        - Accuracy 81.20%
        - Precision 73.95%
        - Recall 77.55%

## Conclusion

1. The Model seems to predict the Conversion Rate very well and we should be able to give the CEO confidence in making good calls based on this model
2. Important features responsible for good conversion rate or the ones' which contributes more towards the probability of a lead getting converted are:
    a. The total time spend on the Website
    b. Lead Origin_Lead Add Form
    c. Last Notable Activity_Had a Phone Conversation