***Submitted By : Raj Pratap Pandey, Sushma Mahagaonkar, Kaveri Namdeorao Deotkar***

# Lead Scoring – Summary

The model is being built and predicted for the business X Education in order to discover ways to convert potential users. We will further understand and verify the data to reach a conclusion to target the correct group and increase conversion rate. Let us go over the measures that were taken:

## 1. EDA:

I. We ran a quick check on the percentage of null values and deleted columns with more than 45% missing values.

II. We also discovered that rows with null values would cost us a lot of data, despite the fact that they were essential columns. So we replaced the NaN values with 'not supplied' instead.

III. Because India was the most frequent occurrence among the non-missing values, we attributed all missing values to India.

IV. When we saw that the number of values for India was quite large (nearly 97% of the data), we dropped this column.

V. Additionally, we focused on numerical variables, outliers, and dummy variables.

## 2. Train-Test split & Scaling :

I. The train and test data were divided at 70% and 30%, respectively.

II. We will use min-max scaling to determine the factors ['TotalVisits,' 'Page Views Per Visit,' and 'Total Time Spent on Website.']

## 3. Model Building

I. RFE was used for feature selection, followed by RFE to determine the top 15 pertinent variables.

II. The remaining variables were then carefully removed based on the VIF values and p-value.

III. A confusion matrix was developed, and overall accuracy was determined to be 80.91%.

## 4. Model Evaluation

- **Sensitivity – Specificity**

If we use the Sensitivity-Specificity Assessment method. We will receive:

Using Training Material

- The ROC curve was used to determine the best cut off number. The region beneath the ROC curve was 0.88.
- After plotting, we discovered that the best limit was 0.35, which resulted in

    - The accuracy is 80.91%.
    - 79.94% Sensitivity
    - 81.50% specificity.

- **Prediction on Test Data**

    - Accuracy 80.02%
    - Sensitivity 79.23%
    - Specificity 80.50%

- **Precision – Recall:**

- On **Training Data**

    - With a limit of 0.35, the precision and recall are 79.29% and 70.22%, respectively.
    - In order to raise the above percentage, we must modify the cut off value. After plotting, we discovered that 0.44 was the best cut off number.

    - Accuracy 81.80%
    - Precision 75.71%
    - Recall 76.32%

- **Prediction on Test Data**

    - Accuracy 80.57%
    - Precision 74.87%
    - Recall 73.26%

- *So, if we use Sensitivity-Specificity Evaluation, the best cut off number is 0.35. And, if we use Precision - Recall Evaluation, the optimal cut off number is 0.44.*

## CONCLUSION

Top variable contributing to conversion:

- Lead source:
    - Total visits
    - Total time spent on website
- Lead origin:
    - Lead add form
- Lead source:
    - Direct traffic
    - Google
    - Welingak website
    - Organic search
    - Referral sites

Last activity:

- Do not email_yes
- Last activity_email bounced
- olark chat conversation

- *The model appears to accurately forecast conversion rates, and we should be able to give the business confidence in making good decisions based on this model.*