# SUMMARY

**The Ask:** X Education wants Data Science Team needs to build a Logistic Regression Model to increase the conversion rate to 80%. And help identify the potential Leads.

Approach:

- Importing Data
- Inspecting the Data frame and data cleaning:

  Read and analyse the data.

  We dropped the variables that had high percentage of NULL values in them. This step also included imputing missing values. Identifying columns having Select value and handling it, as Select is as good as missing value.
- Data Preparation: We visualized columns with less number of missing values and imputed with mode. We verified dataset has 0 duplicate values.

  .
- EDA : Performed univariate and bi-variate analysis, outlier detection by visualizing with Box Plot and Countplot,, fixed data imbalance, also dropped columns having highly skewed data.
- Dummy Variable Creation: Created dummy data for categorical variables
- Test-Train Split: Divided data set into test and train sections with a proportion of 70-30% values.
- Feature Scaling: We used StandardScaler to scale the original numerical variables. Using stats model we created our initial model, which would give a complete statistical view of all parameters of our model.
- Correlations: Dropped few columns having high correlations.
- Model Building: Using RFE we went ahead and selected 15 top important features. Using statistics generated, we recursively tried looking at P-values in order to select most significant values that should be present and dropped insignificant values. Finally, we arrived at 13 most significant variables. VIF's for these variables were also found to be good. We created data frame having converted probability values, we had an initial assumption that a probability value of more than 0.5 means 1 else 0. Based on above assumption, we derived Confusion Metrics and calculated overall Accuracy of model. We also

calculated ‘Sensitivity’ and ‘Specificity’ matrices to understand how reliable model is.

- Plotting ROC Curve:  We tried plotting ROC curve for features and curve came out be pretty decent with an area coverage of 88% which further solidified model.
- Finding Optimal Cut-off point: Optimal cutoff probability is that probability where we get balanced sensitivity and specificity. Plotting graph for ‘Accuracy’, ‘Sensitivity’, and ‘Specificity’ for different probability values, we determined the cutoff point = 0.38
- Based on Precision and Recall tradeoff, we got a cut off value of approximately 0.38

Basis this cut-off we got

- Accuracy  80.55 %
- Sensitivity  69.92 %
- Specificity  87.75 %
- Precision  78.15%
- Recall 77.04%

Thus we have achieved our goal of getting a ballpark of target lead conversion rate to be around 80% .

Recommendations:

The company should focus on below categories and sub-categories to enhance conversion rate as predicted by our model:

| Features | Sub-Categories |
|---|---|
| Lead Origin | Lead Add Form and Lead Import |
| Lead Add Form | SMS Sent and Others |
| Lead Source | Welingak Website and Olark Chat |
| more time on the websites | |

And Avoid calling the leads belonging as they are not likely to get converted:

| Features | Sub-Categories |
|---|---|
| Do Not Email | Yes |
| last Notable activity | Modified |
| last activity | Olark Chat Conversation |
| What is your current occupation | Students and Unemployed |