**Approach Followed:**

* Clean the dataset we have
* Find the important field from the dataset
* Built as logistic regression model to predict the probability of conversion
* Find the optimal cutoff point based on business requirement

**Data Cleaning & EDA:**

* Removed the columns which had high null percentage (more than 45%)
* Removed highly skewed columns
* Combined the values which had very less contribution in their field.
* Imputed the missing numerical fields with the median value

**One Hot Encoding:**

* Created Dummy variables for the categorical fields
* Dropped the dummy variable with least business importance
* Dropped the original categorical fields

**Feature Elimination:**

* Used the RFE approach for feature elimination
* Used manual elminiantion to optimize the model

**Final Model:**

* Fields and their coefficient in the final model are:

Total Time Spent on Website : 0.8316

Lead Source\_Direct Traffic : -0.8834

Lead Source\_Organic Search : -0.7659

Do Not Email\_Yes : -1.3356

Occupation\_Unemployed : 1.0018

Tags\_Interested in other courses : -3.1007

Tags\_Ringing : -3.6890

Tags\_Will revert after reading the email : 4.0046

Last Notable Activity\_SMS Sent : 1.8451

**Optimal Cutoff:**

* Calculated ROC curve
* Calculated the confusion matrix
* Final Optimal Cutoff comes out to be 38.

**Model Evaluation:**

* Sensitivity : 84.42%
* Specificity : 86.72%
* False Positive Rate : 13.27%
* Positive Predictive Value : 89.92%
* Negative Predictive Value : 83.87%
* Precision Score : 83.88%
* Recall Score : 81.15%
* F1 Score : 82.89%

**Overall Model accuracy is 86.61%**

**Conclusion:**

* We have built a logistic regression model which provides the Lead Score to each lead.
* Higher the Lead Score higher the probability of conversion for the lead
* The cut off Lead Score is 38.
* We can change the cut off lead based on the business requirement in different time intervals.