SUMMARY REPORT

The Leads dataset have some columns which have been dropped as they don't add any value to the model. We dropped Lead Profile as 74% of the values were null.

MISSING VALUE TREATMENT

We move on to the treatment of missing values. The Numerical columns which we took into consideration were 'Total website visits', 'Time spent on website'. The Categorical columns taken into consideration were 'Lead Origin', 'Lead Source', 'Occupation', 'City'.
Since 'City' had around 60% of the data from Mumbai so we imputeD Mumbai in the missing values. In TAGS we imputed the missing values with 'Will revert after reading the email'. For 'Country' we imputed the missing values with India and 'Occupation' with Unemployed.

DROPPING VARIABLES BASED ON THEIR DISTRIBUTION

In the next step, we again drop a few variables based on their distribution (maximum entries in one field).

OUTLIER TREATMENT

Next, we perform some outlier treatments for our continuous variables and mostly capped the outliers at **95%**. We capped TotalVisits and Page Views Per Visit for analysis.

EXPLORATORY DATA ANALYSIS

We do some EDA and after completing the univariate analysis we found that many columns are not adding any information to the model, hence we dropped them for further analysis like 'Lead Number', 'Magazine', 'Newspaper Article', 'Receive More Updates About Our Courses',

DATA PREPARATION

We the prepared the data by converting some binary variables (Yes/No) to 1/0, by creating a dummy variable and dropping the first one and also adding the results to the master data frame and dropping the columns not required.

MODEL BUILDING

We find that our target variable is not imbalanced and then proceed to the model building part. We divide our data into the **training** and **testing** sets and start working on building a robust model on the training data.

RFE AND BUILDING A ROBUST MODEL USING P-VALUES AND VIF

We run various steps like feature scaling for the continuous variables. We also performed a **Recursive Feature Elimination (RFE)** to select the **15 strongest feature variables**. We look at **p-values** and drop some variables which turn out to be insignificant. We also ran a check for multicollinearity using **VIF** and when we are satisfied with the p-values and VIF's of our feature variables, we generate the regression equation.

GENERATING LEAD SCORE COLUMN TO IDENTIFY HOT LEADS

We do the VIF and drop couples of columns and we get the predicted values on the train set. We create data frame with the actual churn flag and the predicted probabilities. Then we create new column 'predicted' with 1 if Churn_Prob > 0.5 else 0. Then checking Confusion matrix, VIF's and ROC and then assigning Lead_Score

MAKING PREDICTIONS ON THE TEST SET

We then work on the test set similarly to get a Lead_Score. We check the evaluation metrics on the **test set** and find them to be **nearly like our metrics from the train set**. We conclude that our model is robust and will provide the best of results to X Education.