SUMMARY

The model is being built and predicted for the firm X Education in order to uncover strategies to convert potential consumers. We will further analyze and confirm the data in order to arrive at a conclusion on how to target the relevant group and enhance conversion rate. Let us go over the measures that were taken:

1. Data Reading:

Importing the necessary libraries and dataset.

2. Data Understanding:

- Routine Data Check: the number of rows and columns, the data type of each column, the distribution, mean and median for all numerical columns, and so on.
- Analysis of missing values
- Check for duplicate rows.

3. Data Cleaning:

In this case study, data cleansing is quite important. The data cleansing process determines the model's quality and efficiency. As a result, it must be strictly adhered to.

- The value "Select" is replaced with NAN.
- Computation of missing values for each column, with the Score and Activity variables removed.
- Removing columns that have a high percentage of missing values.
- Verifying that each column has a distinct category.
- If the columns are highly skewed with one category, such columns will be dropped. Combining different categories of the columns with less percentage values into "Others" category.
- Finally Checking for the number of rows kept after performing all the above steps.

4. EDA

- Quick check was done on % of null value and we dropped columns with more than 40% missing values.
- Then we saw the Number of Values for India were quite high (nearly 81% of the Data), so this column was dropped.
- We also worked on numerical variable, outliers and dummy variables.

5. Outlier Treatment:

We deal with the outliers present in the numerical columns using the Inter-Quartile Range approach.

6. Data Preparation:

In this step, the dummy variables were created. Performed train test data split and scaled the numerical columns.

7. Data Modelling:

We used both RFE and manual feature selection methods to get the final list of columns. In between the most insignificant, highly correlated columns are dropped and at last we had 15 columns in our final model

We know that the relationship between ln(odds) of 'y' and feature variable "X" is much more intuitive and easier to understand.

We chose the cutoff probability as 0.2 from Accuracy, Sensitivity, Specificity curve and calculated lead score for all the leads. The sensitivity of model was around 89%. ·A confusion matrix was created, and overall accuracy was checked which came out to be 90%.

CONCLUSION:

High sensitivity implies that our model will correctly identify almost all leads who are likely to Convert.

It will do that by overestimating the Conversion likelihood.

To follow an aggressive work flow choose a lower threshold value for Conversion Probability. This will ensure the Sensitivity rating is very high which in turn will make sure almost all leads who are likely to Convert are identified correctly and the agents can make phone calls to as much of such people as possible

X Education Company needs to focus on following key aspects to improve the overall conversion rate:

- **Tags_Closed by Horizzon**: When the current status of the lead is that he/she is closed by Horizzon, that's when there is the most chance that a lead will convert.
- **Tags_Lost to EINS** : When the current status of the lead is that he/she has lost to ENIS, there is a very high chance that the lead will convert