Lead Score Summary Report

This process was carried out by studying and preparing the Dataset of X Education by removing columns where we have null values greater than **40%** and imputing or replacing the rest of the values in other columns. Then we checked the outlier values in our numerical columns and took the IQR where values less than **5%** and greater than **95%** were automatically removed from those columns. The above steps covered our EDA process. After this, we proceeded towards making dummy variables for our categorical columns and also converted our binary variables into Yes and No. Now, we split our Model into the Train and Test set for Model Building Process and the split percentage is taken to be around **70% and 30%.** We use fit transform on the train set. We have a conversion rate up to this point around **38.5%.** We build our 1st Model using RFE and our aim here is to have a Model in which we have the P-value of every variable less than **0.05 and VIF of all the variables less than 5.** We get our desired results in the 4th Model

Now, once we have developed our model our next step was to find out the **Accuracy (92%), Sensitivity (82.2%), and Specificity (97%)** of the confusion matrix that we have from our train set. We plot the ROC curve to see how much **percentage of the area** we have **under the cure** which comes to be around **90%.** This gives the relation we have between the Sensitivity and the Specificity of the model. Now, we will calculate our **cut-off probability** which comes around **(0.33)** based on the intersection of Sensitivity, Specificity, and Accuracy. Now, where ever we have the converted probability greater than **33%** we assume that the lead is converted. Then again we calculate **Accuracy (91%), Sensitivity (93.2%), and Specificity (89.8%).** We also calculate the **Recall and the Precision percentage** of the Train set and it comes **95%** and **82.2%.** We plot the Recall and Precision and Trade of the curve and found the cut-off to be 0.36 which we used on our Test set.

We begin our Model for **Test set** in which leads having a **probability greater than 33%** are marked as converted and we get the **Accuracy (91%), Sensitivity (93%), Specificity (89.66%), Precision Score (85.8%), and Recall Score (93**%).

Both Models when compared gave us more or less similar results. Based on the final model which we have variables such as:

1. **Tags_Closed by Horizon**
2. **Tags_Lost to EINS**
3. **Tags_Will revert after reading the email**
4. **Lead Source_Welingak Website**
5. **Last Activity_SMS Sent**

The above variables will contribute more towards lead conversion.

And as per our final model, leads that have a lead score greater than 33 have a good chance of **conversion with a conversion percentage of around 93%.**