Summary – Lead Score Assignment

The model building and prediction is being done for company X Education and to find ways to convert potential users. We will further understand and validate the data to reach a conclusion to target the correct group and increase conversion rate. Let us discuss steps followed:

Step 1: EDA

- Large number of columns had Null values and columns with > 40% Null values were dropped.
- We noticed there were some critical columns with high % of Null values and replaced the Null values with “Not Available”.
- We dropped few other categorical columns post univariate analysis as we noticed high skewness.
- We analysed numerical variables and checked for correlations.
- We saw outliers in the numerical columns and handled them by removing the top and bottom 1% values.
- Created dummy variables for categorical columns and dropped the first column.

Step 2: Train-Test split & Scaling

- The split was done at 70% and 30% for train and test data.
- We used standard scaling on the variables ['TotalVisits', 'Total Time Spent on Website', 'Page Views Per Visit']

Step 3: Model Building

- RFE was used for feature selection and select top 15 variables.
- Created the model and dropped columns where p value was > .05. We continued with this process till p value was below 0.05.
- We checked for VIFs and we did not have any variable with > 5, so we did not have to drop any columns based on VIF value.
- Model accuracy was 92% at this point.

Step 4: Metrics beyond simply accuracy

- Confusion matrix was created.
- Specificity and Sensitivity was 96% and 85% respectively.
- False positive rate is ~4%, positive predictive value is 93% and negative predictive value was 91%, suggesting strong predictive capability of the model.

Step 5: Plotting the ROC Curve

- From the curve above, 0.3 is the optimum point to take it as a cutoff probability.
- Created the final predictive model with 0.3 as probability.
- Overall accuracy of the model was 90.15%.
- Specificity and Sensitivity were 91.6% and 87.6% respectively.
- Precision and Recall were at 86% and 87%.
- Area under ROC curve is 0.96, which is close to 1.

Step 10: Making predictions on the test set

- Test Data shows similar results as Train Data:
- a. Accuracy: 90.35% b. Sensitivity : 88.23% c. Specificity: 91.66%

CONCLUSION:

Below are the variables that can increase leads:

- Last Activity_SMS Sent
- Last Notable Activity_Modified
- Tags_Busy
- Tags_Closed by Horizzon
- Tags_Lost to EINS
- Tags_Not Available
- Tags_Ringing
- Tags_Will revert after reading the email
- Tags_in touch with EINS
- Tags_switched off

Model comparison for Train and Test Data:

Train Data:

1. Accuracy: 90.15%
2. Sensitivity: 87.61%
3. Specificity: 91.68%

Test Data:

1. Accuracy: 90.35%%
2. Sensitivity: 88.23%
3. Specificity: 91.66%

The model is able to predict the Conversion Rate and can be used by X Education to target potential leads