Summary

This analysis and model building exercise is for X Education company, which sells its courses to industry professionals. They are trying to find ways to convert more leads (potential customers) to join their courses. The data gives information like how these leads reached the site, time spent on site and conversion rate.

Steps followed in making the logistic regression model:

1. **Data cleaning**: Data looked more or less clean except few null values. Also, few columns had “Select” which we replaced with null values because it did not give much information. Columns like country were dropped because majority of data was from India and there was no variance in data to add value to model.
2. **EDA**: EDA was done on the data and many categorical data were removed which were not relevant. Also some null values were changed to “Not provided” and very low frequencies were clubbed together before creating of dummies to reduce number of columns. Missing values were imputed with mode in categorical data. Numeric data outliers were removed though there were not many outliers in numeric data.
3. **Dummy variables**: Dummy variables were created and “not provided” were removed.
4. **Train test Split**: Data was split in 70% and 30% for train and tests respectively.
5. **Scaling**: Standard scaler was used for numeric variables.
6. **Model building**: Top 15 variables were selected using RFE. Other variables were removed manually looking at their p-value and VIF.
7. **Model Evaluation**: Confusion matrix was made. ROC curve was used to find optimum cutoff at 0.3. Accuracy,

sensitivity, specificity were calculated. Precision and recall were also calculated.

8. **Prediction**: Prediction were done on test data with optimal cutoff of 0.3. And accuracy, sensitivity and specificity were calculated on test data. Precision and recall were also calculated.
9. **Result**: Train Data: Accuracy : 89.80% Sensitivity : 88.94% Specificity : 90.33%

Test Data: Accuracy : 88.94% Sensitivity : 87.89% Specificity : 89.57%

Recall rate of 81.65% on test data and train data.

Important variables:

i. Lead Origin_Landing Page Submission,

ii. Tags_Will revert after reading the email

iii. Last Activity_SMS Sent

iv. Lead Origin_Lead Add Form

v. Tags_Not Specified

vi. Last Notable Activity_Modified

vii. Lead Source_Welingak Website

viii. Tags_Ringing

ix. Total Time Spent on Website

x. Tags_Busy

xi. Tags_Lost to EINS

xii. Specialization_Travel and Tourism

xiii. Last Notable Activity_Olark Chat Conversation