Summary

This analysis is done for X Education and to find ways to get more industry professionals to join their courses. The basic data provided gave us a lot of information about how the potential customers visit the site, the time they spend there, how they reached the site and the conversion rate.

The following are the steps used:

1. **Cleaning data:**
The data was partially clean except for a few null values and the option select had to be replaced with a null value since it did not give us much information. Dropped columns which has more than 40% data missing.

2. EDA:
A Exploratory Data Analysis was done which includes univariate, bivariate and multivariate analysis to check the condition of our data. It was found that a lot of elements in the categorical variables were irrelevant. The numeric values seems good except outliers in that we replace outliers with the most extreme tail value of the boxplot using formula Q3+1.5IQR. To reduce dimensionality we have converted categories with low value count into other (as a new category) with these we will achieve less columns after dummy variable creation.

3. **Dummy Variables:**
The dummy variables were created for the categorical columns and a heatmap was plotted to understand the correlation between features.

4. **Train-Test split and function creation:**
Train and test split was done we have taken 70% data for training and 30% data for test. Created different function like train, vif, model_summary, drop_column and calculate_metrics for code reusability.

5. **Model Building:**
Taken the logistic regression model and RFE from sklearn to select top 20 features. Then dropped features with high VIF (>5) and then dropped features with p-value > 0.05

6. **Model Evaluation:**
Matrix used to evaluate model:

1. Confusion matrics
2. Accuracy score
3. Precision
4. Sensitivity/Recall

5. Specificity

Later on the optimum cut off value was found using accuracy, sensitivity and specificity that is comes out to be 0.34.

7. **Prediction:**

Prediction was done on the test data frame and with an optimum cut off as 0.42(using the precision recall curve) with
Accuracy = 0.88
Precision = 0.84
Sensitivity/Recall = 0.84
Specificity = 0.90

8. **Precision – Recall:**

This method was also used to recheck and a cut off of 0.42 was found.

It was found that below variables are used to find potential buyers:

1. The total time spend on the Website.
2. Email_Call
3. Lead Origin_Landing Page Submission
4. Lead Source_Olark Chat
5. Lead Source_Reference
6. Last Activity_Email Bounced
7. Last Activity_Olark Chat Conversation
8. Specialization_Hospitality Management
9. Specialization_Other
10. What is your current occupation_Working Professional
11. Tags_Closed by Horizzon
12. Tags_Other
13. Tags_Will revert after reading the email
14. Last Notable Activity_Other
15. Last Notable Activity_SMS Sent

Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.