Summary

This analysis is done for X Education, who wishes to identify the most potential leads, also known as ‘Hot Leads’. 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.
2. **EDA:** A quick EDA was done 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 and not much outliers were found.
3. **Dummy Variables**: The dummy variables were created and later on the dummies with ‘not provided’ elements were removed. For numeric values we used the MinMaxScaler.
4. **Train-Test split**: The split was done at 70% and 30% for train and test data respectively.
5. **Model Building**: Firstly, RFE was done to attain the top 15 relevant variables. Later the rest of the variables were removed manually depending on the VIF values and p-value (The variables with VIF < 5 and p-value < 0.05 were kept).
6. **Model Evaluation**: A confusion matrix was made. Later on the optimum cut off value (using ROC curve) was used to find the accuracy, sensitivity and specificity which came to be around 80% each.
7. **Prediction**: Prediction was done on the test data frame and with an optimum cut off as 0.3 with accuracy, sensitivity and specificity of around 90%.
8. **Precision – Recall**: This method was also used to recheck and a cut off of 0.3 was found with Precision around 88% and recall around 91% on the test data frame.

Recommendations:

- The company **should make calls** to the leads:
    1. coming from the lead sources "Welingak Websites" and "Reference" as these are more likely to get converted.
    2. who are the "working professionals" as they are more likely to get converted.
    3. who spent "more time on the websites" as these are more likely to get converted.
    4. coming from the lead sources "Olark Chat" as these are more likely to get converted.
    5. whose last activity was SMS Sent as they are more likely to get converted.

- The company **should not make calls** to the leads:
    1. whose last activity was "Olark Chat Conversation" as they are not likely to get converted.
    2. whose lead origin is "Landing Page Submission" as they are not likely to get converted.
    3. whose Specialization was “Others" as they are not likely to get converted.
    4. who chose the option of "Do not Email" as "yes" as they are not likely to get converted.

The customers which should be contacted are the customers whose "Lead Score" is equal to or greater than 85. They can be termed as 'Hot Leads'. Thus we have achieved our goal of the target lead conversion rate to be around 90% . The Model seems to predict the Conversion Rate very well