Summary

We have asked to solve the business problem of improving conversion rate of X Education to find ways to increase their current conversion rate of 30% to 80% by finding out hot leads which is nothing but potential leads having high probability of getting converted by using their past year data. So that sales team can focus more time on those leads. Which will allow them to increase their leads to conversion ratio.

Following are the steps used to solve the business problem:

- **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. Few of the null values were changed to 'not provided' so that we do not lose much data. Although they were later removed while making dummies. Since there were many from India and few from outside, the elements were changed to 'India', 'Outside India' and 'not provided'.
- **EDA**: A quick EDA was done to check the condition of our data both using Sweetviz library and manually. It was found that a lot of elements in the categorical variables were irrelevant. The numeric values seem good, and no outliers were found.
- **Dummy Variables**: The dummy variables were created and later the dummies with 'not provided' elements were removed. For numeric values we used the MinMaxScaler to normalize the values.
- **Train-Test split:** The split was done at 70% and 30% for train and test data respectively.
- **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 < 4 and p-value < 0.05 were kept).
- **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 81.1%, 77.7% and 83% respectively.
- **Prediction:** Prediction was done on the test data frame and with an optimum cut off as 0.38 with accuracy, sensitivity, and specificity of 80%.
- **Precision – Recall:** This method was also used to recheck and a cut off of 0.42 was found with Precision and recall around 74% on the test data frame.

It can be seen in our final model that there are lot of features which will impact the potential buyer's decision to enrol in the course we will see few variables which have positive effect on our target variable in descending order:

1. Total number of visits made by the customer on the website
2. The total time spent on the Website.
3. When leads current occupation is working professional
4. Lead origin is Lead Add Form
5. Lead source is Welingak Website
6. Lead activity is had a phone conversation

Keeping these in mind the X Education can investigate their leads and focus more time on clients with these features in the data to increase their conversion rate.