SUMMARY

We essentially began with a lot of features that looked like they can be particularly insightful, in order for us to understand how and why leads are getting converted. This would help us in conveying the factors to X education, so that they can use this information for choosing hot leads.

So, we shaped our analysis around this idea only. We began with more than 9000 data points, and 37 features, that had to be reduced to the most pertinent features. Therefore, we cleaned the data by treating null values by dropping a few columns, dropping NaN values, dropping columns that had most values as 'Select', and so on. From there we moved on to mapping variables to binary values (Yes or No type of entries).

Categorical variables were hot encoded then. Thereafter, the data was split into test and train sets by keeping 'Converted' as the dependent variable. Data was then scaled down using StandardScaler.

Now we moved on to creating a model (logistic regression model) using Recursive Feature Elimination. Model1 was further improved by analysing VIF and p-values, which led to us dropping another 2 features to create Model2. Now moving on to model evaluation, prediction values were used to create a confusion matrix, first using an arbitrary cut-off and then ROC curve was used to find a more appropriate cut-off threshold. Comparison between Accuracy, sensitivity, & specificity. Precision and recall scores were also calculated.

Top three features for X education to focus on: ‘What is your current occupation_Working Professional’, ‘What is your current occupation_Other’, ‘Lead Origin_Lead Add Form’.

Recommendations: That ideal hot leads should have maximum values for the above mentioned three features. So, such leads should be focussed on as much as possible. This can be done by informing them about new offers, newly added course, course application deadlines, course application status, job offers, and so on. The information provided to the leads would have to be designed according to the lead being followed, and for doing so the necessary data can be collected when they are being reached out to. Thereon, the leads should be monitored closely for the best possible outcome.