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:
   - Columns with >30% nulls were dropped.
   - Replaced 'Select' with null values since it did not give much information.
   - Value counts within categorical columns were checked to decide appropriate action: if imputation causes skew, then column was dropped, columns that don't add any value were also dropped.

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. Data imbalance checked and only 38% leads converted.
   - Performed univariate and bivariate analysis for categorical and numerical variables.

3. Dummy Variables:
   - Created dummy features (one-hot encoded) for categorical variables
   - Splitting train & test sets: 70:30 ratio
   - Feature Scaling using MinMaxScaler

4. 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.
   - Manual Feature Reduction process was used to build models by dropping variables with $p-value > 0.05$.

- Total 5 models were built before reaching final Model 5 which was stable with (p-values < 0.05). No sign of multicollinearity with VIF < 5.
- Res was selected as final model with 13 variables for making prediction on train and test set.

5. Model Evaluation:
   - A confusion matrix was made.

   [3246, 707],

   [521, 1898]
   - Cut off point of 0.37 was selected based on accuracy, sensitivity and specificity plot.
   - Training model metrics:
     - Accuracy: 81%
     - Sensitivity: 78%
     - Specificity: 82%
     - Precision: 73%
     - Recall: 78%
   - Lead score was assigned to train data using 0.42 as cut off.

6. Prediction
   - Making Predictions on Test: Scaling and predicting using final model.
   - Test model metrics:
     - Accuracy: 81%
     - Sensitivity: 79%
     - Specificity: 82%
     - Precision: 73%
     - Recall: 79%
   - Evaluation metrics for train & test are very close to around 80%.
   - Lead score of 0.42 was assigned.
   - Top 3 features are:
     - Lead Origin_Lead Add Form
     - Last Notable Activity_Had a Phone Conversation
     - What is your current occupation_Working Professional

7. Recommendations
   - More budget/spend can be done on Welingak Website in terms of advertising, etc.
   - Incentives/discounts for providing reference that convert to lead, encourage to provide more references.
   - Working professionals to be aggressively targeted as they have high conversion rate and will have better financial situation to pay higher fees too.

As per the above-mentioned points we can say that X education can increase all the potential buyers to buy their courses.