Summary Report

We analyzed X Education's data to help them raise their conversion rate from 38% to 80%. We looked at lead details, where they came from, their website activity, email interactions, and their last contact with X Education. Using this data, we created a strategy to improve conversions. Process followed with rationale behind each step is as follows -

Loading and inspecting the data

We loaded the data, examined its structure, and identified 'Converted' as the target variable.

Understanding and cleaning the data (with EDA)

- **No Duplicates**: All records were unique.
- **Data Cleaning**: 'Select' labels in categorical columns were replaced with NaN. Columns with over 35% missing values were dropped.
- **Variable Analysis w.r.t to target variable**: We analyzed each column to decide how to impute variables. Low variance variables like 'Country' and 'What matters most to you in choosing a course' were dropped. Others were dropped as the data found in them was imbalanced.
- **Grouping**: We regrouped low frequency actions in 'Last Activity' and 'Last Notable Activity' into interpretable groups.
- **Outlier Treatment**: We treated outliers in numerical variables and evaluated their relation to the target variable.
- **Conversion Rate**: We found the current conversion rate to be approximately 38%.
- Deduced the current conversion rate to be ~38%

Data preparation

- Converted the Yes/No categorical variable into binary 1/0 variable
- Created dummy variables for retained categorical variables

Model Building

- We split the data into a 70% training set and a 30% test set.
- The correlation matrix revealed some collinearity among a few features. However, we decided to address this during model building due to the number of variables.
- We employed Recursive Feature Elimination (RFE) to identify the top 15 relevant variables. We fine-tuned the model by first manually eliminating variables with a p-value greater than 0.05, and then those with a Variance Inflation Factor (VIF) greater than 5 to remove highly collinear variables.
- We used the StandardScaler method to scale the numerical variables we retained.

Model Evaluation

- We used the ROC curve to evaluate the model's predictive performance, which was approximately 0.89.
- We determined the optimal cut-off point using accuracy, sensitivity and specificity to be 0.4. The evaluation metrics at this cutoff – accuracy is 81%, Sensitivity (recall) is 76.6%, Specificity is 84.4% and Precision is 75.9%
- Since the CEO of the company wants to improve the conversion rate to 80%, we continued to adjust the cutoff to get better Precision.

PREDICTION

Predictions were done on the rest of the test data at the optimal cut-off of 0.4 with minimal difference in evaluation metrics.

CONCLUSION

Increased odds of lead conversion	Decreased odds of lead conversion
1. Lead Origin a. Lead Add Form 2. What is your current occupation a. Working professional 3. Last Activity a. Digital or Direct Engagement – Clicked on the view link, Approach Upfront, Resubscribed Email, Had a phone conversation b. SMS Sent c. Email opened 4. Lead Source a. Welingak Website b. Olark Chat 5. Total time spent on website	1. Last Notable Activity a. Modified 2. What is your current occupation a. Unknown 3. Do not Email
	4.