Summary

X Education receives various leads, but its current lead conversion rate is relatively low at about 30%. The company needs us to develop a model that assigns a lead score to each prospect. This scoring should ensure that leads with higher scores are more likely to convert. The company wants to increase the lead conversion rate to approximately 80%.

Data Cleaning:

- Columns with more than 3000 missing values were removed. For categorical columns, value distributions were reviewed to determine the best course of action: if imputation led to skewed data, the column was either removed, a new category labeled "others" was created, high-frequency values were used for imputation, or columns that offered no additional value were dropped.
- Additional preprocessing steps included handling outliers, correcting invalid data entries, consolidating low-frequency values, and mapping binary categorical variables.

EDA:

- Checked and analyzed the data imbalance, finding that only less leads were converted.
- Conducted both univariate and bivariate analyses for categorical and numerical variables. Variables such as 'Lead Origin', 'What is your Occupation', and 'Lead Source' were analyzed for their influence on the target variable.
- The amount of time spent on the website was found to have a positive effect on lead conversion.

Data Preparation:

- Created dummy features for categorical variables
- Splitting Train & Test Sets ratio is 70 :30
- Feature Scaling using Min-Max Scaler
- Removed several columns due to their high correlation with one another.

Model Building:

- Used RFE and VIF to reduce variables from 41 to 15. This will make data frame more manageable.
- A manual feature selection process was employed to construct the models, which involved excluding variables with a p-value greater than 0.05.
- Total 3 models were built and the 3rd model logm3 was stable with (p-values < 0.05). No sign of multicolinearity with VIF < 5.
- Logm3 was selected as final model with 13 variables, we used it for making prediction on train and test set.

Model Evaluation:

- Confusion matrix was made and cut off point of 0.41 was selected based on accuracy, sensitivity and specificity plot. This cut off gave accuracy, specificity and precision all around 80%. Whereas precision recall view gave performance metrics around 79%.
- As to solve business problem, company asked to boost conversion rate to 80%, but metrics dropped when we took precision-recall view. So, we will choose sensitivity-specificity view for our optimal cutoff for final predictions.
- Lead score was assigned to train data using 0.41 as cut off.
- Accuracy is 0.79, Sensitivity is 0.73 , and specificity is 0.84 for Trained dataset
- For Test dataset accuracy is 0.78 , sensitivity is 0.76 and specificity is 0.80

Making Predictions on Test Data:

- Making Predictions on Test: Scaling and predicting using final model.
- Evaluation metrics for train & test are very close to around 80%.
- Lead score was assigned.

- Top 3 features are:
  - Total Time Spent on Website -0.43
  - Lead Origin_Lead Add Form -0.27
  - What is your current occupation_Working Professional-0.11

Recommendations:

- Focus more on the features with positive correlation.
- Allocate more funds for improving advertising
- Provide incentives/ discounts for providing reference that convert to lead.
- As areas of improvement review landing page submission