Summary

X Education generates a substantial number of leads, but its lead conversion rate currently stands at a low 30%. The company has tasked us with developing a model to assign lead scores to each lead, aiming to prioritize those with higher scores due to their increased likelihood of conversion. The CEO's objective is to achieve a lead conversion rate of approximately 80%.

Data Cleaning:

- Columns containing over 40% null values were removed. For categorical columns, we examined value counts to determine the most suitable action: if imputation led to skewness, the column was dropped, a new category "others" was created, the most frequent value was imputed, or irrelevant columns were removed.
- Numerical categorical data were imputed using the mode, and columns with only one unique customer response were eliminated.
- Additional tasks included addressing outliers, rectifying invalid data, consolidating low-frequency values, and mapping binary categorical values.

EDA:

- Data imbalance checked- only 38.5% leads converted.
- Performed univariate and bivariate analysis for categorical and numerical variables. 'Lead Origin', 'Current occupation', 'Lead Source', etc. provide valuable insight on effect on target variable.
- Time spend on website shows positive impact on lead conversion.

Data Preparation:

- Created dummy features (one-hot encoded) for categorical variables
- Splitting Train & Test Sets: 70:30 ratio
- Feature Scaling using Standardization
- Dropped few columns, they were highly correlated with each other

Model Building:

- Used RFE to reduce variables from 48 to 15. This will make dataframe more manageable.
- Manual Feature Reduction process was used to build models by dropping variables with p – value > 0.05.
- Total 3 models were built before reaching final Model 4 which was stable with (p-values < 0.05). No sign of multi collinearity with VIF < 5.
- logm4 was selected as final model with 12 variables, we used it for making prediction on train and test set.

Model Evaluation:

- Confusion matrix was made and cut off point of 0.345 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 less performance metrics around 75%.
- As to solve business problem CEO 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 cut-off for final predictions
- Lead score was assigned to train data using 0.345 as cut off.

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:
  - Lead Source_Welingak Website
  - Lead Source_Reference
  - Current_occupation_Working Professional

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.