Summary

X Education company has a lead conversion rate of only 30%, and the CEO's target is 80%. To achieve this target, we built a lead scoring model that assigns a score to each lead based on the likelihood of conversion. Our approach involved data cleaning, EDA, data preparation, model building, and evaluation.

Data Cleaning:

- Columns with >=40% nulls were dropped. Value counts within categorical columns were checked to decide appropriate action: if imputation causes skew, then column was dropped, created new category (others), impute high frequency value, drop columns that don't add any value.
- Numerical categorical data were imputed with mode and columns with only one unique response from customer were dropped.
- Other activities like outliers' treatment, fixing invalid data, grouping low frequency values, mapping binary categorical values were carried out.

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 encoding) for categorical variables
- Splitting Train and Test Sets: 70:30 ratio
- Feature Scaling using Standardization
- Dropped few columns which had high correlation 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 multicollinearity with VIF less than 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
    - LeadSource_Reference
    - current_occupation_Working Professional

Recommendations:

Based on these findings, X Education could increase spending on the Welingak website in terms of advertising, offer incentives or discounts for providing references that convert to lead, and aggressively target working professionals, who have a higher conversion rate and better financial situation to pay higher fees.