

X Education generates a substantial number of leads, but its lead conversion rate is currently low, hovering at about 30%. The company has tasked our group with developing a model to assign lead scores to these prospects, prioritizing those with higher scores for a better chance of conversion. The CEO's goal is to significantly improve the lead conversion rate, aiming for an ambitious target of approximately 80%.

Data Cleaning:

- Columns with >40% nulls were dropped.
- Dealing with categorical columns:
 - o Value counts of the categorical columns were checked
 - o If imputation causes skew, then column was dropped
 - o Created a new category called 'Others' to group smaller categories that didn't add any value
 - o Dropped columns that don't add any value
- Dealing with numerical columns:
 - o Imputed with mode
 - o columns with only one unique response from customer were dropped.
- Other activities:
 - o Fixing invalid data
 - o Mapping Binary Categorical variables

EDA:

- Data imbalance checked- only 38.5% leads converted.
- Performed univariate, bivariate and multivariate analysis for categorical and numerical variables.

Data Preparation:

- Creation of dummy variables for categorical variables
- Split Train-test using 70:30 ratio
- Used Standardization to feature scale
- Dropped columns that were highly coorelated

Model Building:

- Using RFE reduced variables to 20.
- Manual Feature Reduction process was used to build models by dropping variables with $p - \text{value} > 0.05$.
- Total 3 models were built. No sign of multicollinearity with $VIF < 5$.
- Logm3 was selected as final model with 19 variables, and used it to predict train and test set.

Model Evaluation:

- Based on accuracy, sensitivity and specificity plot, cut off point of 0.348 was selected which gave approximately 80% accuracy, specificity and precision.
- precision recall view gave less performance metrics around 75%.
- Chose sensitivity-specificity view for our optimal cut-off for final predictions as it aligned better with the 80% goal set by the CEO.

Making Predictions on Test Data:

- 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
 - Last Activity_SMS Sent

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.
- More time and effort spent on SMS campaigns