Summary

X Education gets many leads; its lead conversion rate is very poor at around 30%. The company requires us to build a model wherein we need to assign a lead score to each of the leads such that the customers with a higher lead score have higher conversion chance. CEO’s target for lead conversion rate is around 80%.

Data Cleaning

- Columns with >40% nulls were dropped.
- Value counts within categorical columns was checked to decide an appropriate action: if imputation causes skew, then column was dropped,
- Created new category (Unknown), impute high frequency value,
- Drop columns that do not add any value.
- Imputation of Numerical categorical data x 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 of binary categorical values was performed.

Exploratory Data Analysis (EDA)

- Data imbalance was checked: only 38.5% of leads converted.
- Univariate and bivariate analyses were performed, providing insights into variables influencing lead conversion.
- Time spent on the website showed a positive impact on lead conversion.

Data Preparation

- Label encoder was used for categorical variables.
- The data was split into training and test sets in a 70:30 ratio.
- Feature scaling was applied using standardization, and highly correlated columns were dropped.

Model Building

- RFE reduced the number of variables from 48 to 15.
- Manual feature reduction was used to build models, with variables having p-values > 0.05 being dropped.
- The final model, logm3, with 12 variables, was selected for making predictions on the training and test sets.

Model Evaluation

- A confusion matrix was created, and a cutoff point of 0.35 was selected based on accuracy, sensitivity, and specificity plots. This cutoff provided accuracy, specificity, and precision all around 78%.

Making Predictions on Test Data

- The final model was used for scaling and predicting, with evaluation metrics for training and test sets being close to 78%.
- Lead scores were assigned, identifying the top 3 features: Lead Source, Lead Origin, and Total time spent on the website.

Recommendations

- Increase budget/spending on the Websites for better advertising.
- Provide incentives/discounts for references that convert to leads.
- Aggressively target working professionals due to their high conversion rate and better financial situation.
