

Summary

This analysis for X Education aims to identify strategies to attract more industry professionals to enrol in their courses. The steps taken include data cleaning, exploratory data analysis (EDA), creating dummy variables, splitting the dataset, scaling, building, evaluating the model, and at last making predictions.

Actions:

Actions done on the provided dataset are:-

1. **Data Cleaning:** The initial dataset had a few null values and an uninformative "option select" entry. Null values were replaced with 'not provided' where necessary, and the data was categorized into 'India', 'Outside India', and 'not provided' to simplify analysis.
2. **Exploratory Data Analysis (EDA):** A preliminary EDA revealed that many categorical variable elements were irrelevant, while numeric values were clean with no outliers.
3. **Dummy Variables:** Dummy variables were created, excluding those with 'not provided' elements. Numeric values were normalized using the MinMaxScaler.
4. **Train-Test Split:** The dataset was split into 70% for training and 30% for testing.
5. **Model Building:** Recursive Feature Elimination (RFE) identified the top 20 relevant variables. Variables were further refined manually based on Variance Inflation Factor (VIF) and p-value criteria (keeping those with $VIF < 5$ and $p\text{-value} < 0.05$).
6. **Model Evaluation:** A confusion matrix and ROC curve analysis were used to determine the optimum cut-off value, achieving approximately 80% accuracy, sensitivity, and specificity.

7. Prediction: Predictions on the test dataset with an optimum cut-off of 0.38 maintained around 80% accuracy, sensitivity, and specificity.

8. Precision-Recall: Additional evaluation using precision-recall identified a cut-off of 0.42, with precision around 75% and recall around 76%.

Key Variables

The analysis identified the following variables as most influential in converting potential Leads:

1. Total time spent on the website.
2. Total number of visits.
3. Lead sources
 - Google
 - Direct traffic
 - Organic search
4. Last activities
5. Lead origin
6. Current occupation

Conclusion:-

Focusing on above mentioned variable, our model was predicting lead conversion quite nicely. So, X need to focus on them so that they can convince professionals to enrol in their course.