

Summary

This analysis aimed to enhance X Education's lead conversion strategy by leveraging data insights to target industry professionals more effectively. The provided dataset offered comprehensive information on user interactions, such as site visits, time spent, and conversion rates. Here's a concise overview of the methodology and findings:

1. **Data Cleaning:** The initial dataset was mostly clean but contained some null values and a placeholder option 'Select' that was treated as null. To retain as much data as possible, null values were replaced with 'not provided.' Categorical elements were categorized into 'India,' 'Outside India,' and 'not provided' to standardize the data.
2. **Exploratory Data Analysis (EDA):** EDA revealed that many categorical variables had irrelevant elements, but numerical values were generally sound with no outliers detected. This step helped in understanding the data better and prepared it for further processing.
3. **Dummy Variable Creation:** Dummy variables were generated for categorical features, with 'not provided' categories removed to avoid skewed results. Numeric values were scaled using MinMaxScaler to normalize their impact on the model.
4. **Train-Test Split:** The dataset was divided into training (70%) and testing (30%) subsets to train and validate the model effectively.
5. **Model Building:** Recursive Feature Elimination (RFE) was utilized to select the top 15 relevant variables. Additional variable selection was based on Variance Inflation Factor (VIF) and p-values, retaining variables with $VIF < 5$ and $p\text{-value} < 0.05$ to ensure model robustness.
6. **Model Evaluation:** A confusion matrix was used to assess model performance, with an optimal cutoff value derived from the ROC curve. The model achieved accuracy, sensitivity, and specificity of approximately 80% each, demonstrating strong predictive capabilities.
7. **Prediction:** The model was applied to the test dataset with an optimal cutoff value of 0.35, maintaining accuracy, sensitivity, and specificity around 80%.
8. **Precision-Recall Analysis:** To further validate the model, precision-recall metrics were analyzed, resulting in a cutoff of 0.41, with precision around 73% and recall around 75%.

Key Variables Affecting Lead Conversion:

1. **Total Time Spent on the Website:** Leads who spend more time are more likely to convert.
2. **Total Number of Visits:** Increased visits correlate with higher conversion chances.
3. **Lead Source:**
 - Google
 - Direct traffic
 - Organic search
 - Welingkar website
4. **Last Activity:**
 - SMS
 - Olark chat conversation
5. **Lead Origin:** Lead Add form.
6. **Occupation:** Working professionals show higher conversion rates.

Based on these findings, X Education should focus on leads originating from high-engagement sources like Google and direct traffic, prioritize leads who interact frequently and spend more time on the website, and optimize communication strategies, particularly through SMS and chat. By targeting these high-potential leads and tailoring their approach based on occupation and lead source, X Education can significantly enhance its conversion rates and attract more industry professionals to their courses.

SUMMARY BY:- AMIT BHARTI

EMAIL:- amitmitu7235@gmail.com