

Summary

An education company named X Education sells online courses to industry professionals; After implying several marketing strategies their current lead conversion rate is less around 30%; We need to come up with the model which predicts more potential leads and increase the conversion rates to nearby 80%.

Steps for creating a successful predicting model are as follows:

- **Data Cleaning:**
 - *Dropped Columns* which has null values more than 30 %, having unique values in all rows, having one unique categorical value, having skewed data
 - *Manipulated Data* in which column has misspelled data, Replaced “Select” value with NULL, clubbed categorical values with lower frequencies to one category
 - *Data Imputation* performed in which NULL values in numerical columns were replaced with their Median and NULL in categorical columns were replaced with their Mode values
 - *Outliers* were detected and handled by capping the extreme outliers with 99th percentile value
- **EDA:**
 - Converted Column Data imbalance ratio is 1:1.59; 38.5 % customers have converted
 - Performed Univariate, Univariate Segmented, Bivariate, Multivariate analysis
 - Columns like **Lead Origin, Lead Source, Current Occupation, Time Spent on Website, Total Visits** seems to be potential indicators for this analysis
- **Data Preparation for Logistic Regression:**
 - Dummy variables were created using Encoding for categorical values with more than two values
- **Model Building:**
 - Train – Test Split in the ratio 70:30
 - Feature Scaling using Min-Max Scaler for numerical columns
 - Used Recursive Feature Elimination to select 15 variables for reducing complexity of dataset
 - Fine tuning was performed by eliminating features having p-values more than 0.05 and variance inflation factor more than 5
 - Final model was created with good vif and p-values
- **Model Evaluation:**
 - Confusion matrix was created; Using ROC curve, Accuracy, Sensitivity, Specificity plot, Precision-Recall Trade-off plot , we have arrived to an optimum cut-off value as 0.35
 - Accuracy, Sensitivity, Specificity, Recall came around 80% on Train data
- **Prediction on Test Data:**
 - Using optimum cut-off 0.35 we got Accuracy, Sensitivity, Specificity, Recall around 80% on Test data too.

➤ **Lead Score Generation:**

- Lead Score was calculated using probabilities values; Higher the value higher the chance of conversion

➤ **Results and Recommendations:**

- Have achieved model predicting HOT LEADS with accuracy of 80%
- Top 3 Variables influencing the Conversion is "Total Time Spent on Website", "Lead Origin_Lead Add Form", "Lead Source_Welingak Website"
- Sales team should concentrate on leads depending on total time spent on website, lead source generated through Welingak website and lead originated from lead add form.
- Leads contacted through SMS and email have better conversion rate and hence sales team should share promotional campaign like new courses or discount offers on courses to leads who have higher lead score through SMS and email.
- The target audience should be Indian working professionals as conversion rate is higher in this category