Summary Of Case Study

This analysis is done to find ways to get more industry professionals to join courses offered by X Education Company . The data provided contained information about how the potential customers visit the site, the time they spend in the site , their reference , their conversion rate etc .

Steps Involved in analysis are :

- Cleaning the given data : Dataframe has many categorical variables which have a level called 'Select' which needs to be handled because it is as good as a null value.Columns with more than 40 % null values were removed . Other missing values were imputed using mode . Unwanted columns were removed since they did not have any significance in our analysis .Standardised data in columns and fixed invalid values . Lead Score and Last Activity columns have very few records. To prevent ending up with a bunch of unnecessary columns when we create dummy variables.
- Data Analysis (EDA) : Data is imbalance when one value is present in majority and other is in minority meaning an uneven distribution of observations in dataset . Target variable is 'Converted' which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn't convertedConversion rate is of 38.5%, i.e only 38.5% of the people have converted to leads. In Categorical Univariate Analysis we get to know the value counts percentage in each variable that how much is the distribution of values in each column .
- Dummy Variables : For categorical variables with multiple levels, create dummy features .
- Train Test split : The split was done at 70 % for train and 30 % for test .
- Model Building : We will Build Logistic Regression Model for predicting categorical variable . Feature Selection Using RFE Coarse tuning , Manual fine-tuning using p-values and VIFs , RFE was done to attain the top 15 relevant variables . Columns with high p- value were removed from model .
- Evaluating the model :A confusion matrix was made. Later on the optimum cut off value (using ROC curve) was used to find the accuracy, sensitivity and specificity . Area under ROC curve is 0.88 out of 1 which indicates a good predictive model.

- Making Prediction : *0.345 is the approx. point where all the curves meet, so 0.345 seems to be our Optimal cutoff point for probability threshold .* The model achieved a sensitivity of 80.05% in the train set and 79.82% in the test set, using a cut-off value of 0.345.
- Conclusion : The CEO of X Education had set a target sensitivity of around 80%.The model also achieved an accuracy of 80.46%, which is in line with the study's objectives.
- Top three variables in our model which contribute most towards the probability of a lead getting converted are :

  Lead Source_Welingak Website
  Lead Source_Reference
  Current Occupation_Working Professional