Objective:

X Education Company facing a poor converstion rate problem regarding which a logestic regression model is build which can predict the probablity of conversion base on existing data set.

Method used to Build the Model:

DATA CLEANING:

1. All the entries with ‘Select’ record were convered to NaN value.
2. Columns with >40% null values were dropped.
3. Highly Skewed columns were also dropped
4. Columns with <40% were handled indivisially and considering their distribution the null values were replace by mode or median values.

EDA

1. Univariate analysis in categorical variables showed the maximum and minimum occurrence of categories
2. Numerical columns showed that there were outliers in 2 columns so the values in these columns were capped to 95%
3. Bivariate analysis was done using ‘Converted’ as target variable.
4. Multivariate analysis using correlation matrix showed the most correlated variables such as total visits and page views per visits.

DATA PREPARATION

1. Dropped 7 columns which would contribute less in the analysis.
2. Convereted 3 categorical columns into binary variables with yes = 1 and no = 1.
3. Created Dummy Variables for Categorical variables
4. Split the Data into test-train with 70-30 ratio.
5. Scaled the numeric columns using minmaxscaler

BUILDING MODEL

1. Used RFE for Feature Selection and +20 variables were selected
2. Regression models were build until a low VIF score and 0 P-Value of variables were achived.
3. Model evaluation gave accuracy 81%, Sensitivity 70% and Specificity 88% .
4. Found optiomal cut-off using ROC curve that is 0.35 and all the TP, TN FP, FN are balanced at 80% and precision is 72% and recall is 81%
5. Made prediction on test data set with accuracy , senstivity and specificity as 80%.

Conclusion and Learnings:

- The sales team of the X-Education should focus on the leads having lead origin: lead add form, occupation: Working Professional, Lead source: Wellingak website.
- Hot Leads are identified as Customers having lead score above 35. Sales Team of the company should first focus on the Hot Leads.
- There are many important variables like city, specialization, occupation which can potentially explain Conversion better. It is important for the management to make few of these information mandatory to fill, so that we can use in our model and suggest more important decisions for the business.
- High Sensitivity will ensure that almost all leads who are likely to Convert are correctly predicted where as high Specificity will ensure that leads that are on the brink of the probability of getting Converted or not are not selected.
- If the Last Notable Activity is Modified, the individual may not be the potential Lead.