Summary Report

This case study focused on applying our knowledge on Logistic Regression. Our goal was to help X Education firm select the most promising leads by creating a logistic regression model that assigns lead score (1-100) to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

Below are the steps that we followed for achieving this requirement:

Data Cleaning and Manipulation

Drop columns that have single values and redundant columns: There were a few columns with single values that would not have added any value add to the model we dropped those columns.

Handle columns with null/missing values by dropping or imputing: We first imputed columns with 'Select' value with NaN null value. Dropped columns with 45% above null values. For few columns we imputed null values with 'Not Given' value. Tags column has more than 35% null value, we dropped this column.

Outlier Detection: Used boxplots to identify outliers, there were few outliers in Total Visits and Page Views per visit however since these both variables seems to be not correlated to Converted we dropped them

Perform EDA

Performed Bivariate and Heatmap analysis to find relation of variables with target variables to decide if they should be kept or dropped. Found that there were couple columns that post analysis did not seem to relate with target variable and so we dropped them

Data Preparation

There were few Yes/No columns converted them to Binary value, created Dummy Variable for categorical columns and did feature Scaling using StandardScaler.

Model Building

Applied RFE to get top 15 best performing features and created Logistic Regression model. Then used high P-values and VIF to eliminate low performing features. Features with VIF < 5 were kept rest were eliminated.

Model Validation

Created confusion matrix and identified Sensitivity, Specificity, Pression and Recall. ROC curve was generated with 88% AUC and specificity above 80%. Also determined the best cutoff to be 0.37

Apply model on test data and assigned lead score based on predicted probability multiplied by 100

Learnings

We understood the significance of using RFE and usage of P-value and VIF to do feature selection and eliminations and its effect on the model.

Learnt about significance of confusion matrix, accuracy, specificity, sensitivity, and precision in Logistic Regression.

Role of ROC curve and plotting accuracy, sensitivity, specificity to get optimal Cut-Off value and create a good model to solve business problem