Summary Report

* Submitted by Rahul Gupta, Rahul Davaskar, Rinky Juneja
* DSC56

X-Education has appointed us to help them select the most promising leads, who are most likely to convert. We are required to build a model where we need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance. To achieve the above result, we performed below steps and achieved 80% conversion goal.

1. **Data Understanding**

First, we went through the data to understand what it contains what is our target variable.

Second, we tried to understand what each column signifies via Data Dictionary.

1. **Data Cleaning**

Dropped all the columns where we have more than 40% null values.

Removed duplicate rows.

Identified Categorical and Numerical variables.

With the count plot of Categorical variables, we dropped skewed columns.

Performed data imputation for both types of variables.

Outliers in numerical variables are handled via IQR.

1. **Exploratory Data Analysis**

Performed univariate analysis for both variables.

Performed bivariate analysis with Converted columns for each type of variables.

Performed multivariate analysis post creating dummy variables.

1. **Scaling and Train-Test split**

Categorical variables are arranged to contain max 6 categories.

Dummy variables are created for Categorical variables.

Performed train test split on dataset with 70% train data.

Performed Standard Scaler for both train and test dataset.

1. **Building Model**

Used RFE from Logistic Regression to select 15 feature variables.

Model is created using GLM.

VIF is calculated and p score is observed.

Features having VIF more than 5 and p score more than 5% are dropped.

1. **Evaluating Model**

First confusion matrix is created.

Optimum cut off values is fetched via ROC curve.

Accuracy, Sensitivity and Specificity is calculated for each cut off till 80%.

Cuff off is calculated i.e., 0.31 using precision recall.

Lead score is calculated for each row of data.

1. **Model Prediction**

Model is tested on train dataset.

80% accuracy is achieved.

1. **Final Feature Model**

Eventually, the final model is created, and important features are noted down.

Positive Impacted Features are given below, these need to be considered while making a phone call as they are most likely to convert.

* 1. **Total Visits**
  2. **Total Time Spent on Website**
  3. **Lead Add Form**
  4. **Olark Chat**
  5. **Working Professional**

Negative Impacted Features are given below, these parameters negatively impact lead conversion.

* 1. **Page Views Per Visit**
  2. **Converted to Lead**
  3. **Email Bounced**

