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.

2) 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.

3) 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.

4) 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.

5) 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.

6) 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.

7) Model Prediction

Model is tested on train dataset.

80% accuracy is achieved.

8) 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.

a. **Total Visits**
b. **Total Time Spent on Website**
c. **Lead Add Form**
d. **Olark Chat**
e. **Working Professional**

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

f. Page Views Per Visit
g. Converted to Lead
h. Email Bounced

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6220
Model:	GLM	**Df Residuals:**	6211
Model Family:	Binomial	**Df Model:**	8
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2910.4
Date:	Mon, 27 Nov 2023	**Deviance:**	5820.9
Time:	19:49:29	**Pearson chi2:**	8.49e+03
No. Iterations:	6	**Pseudo R-squ. (CS):**	0.3267
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.1103	0.101	-20.879	0.000	-2.308	-1.912
TotalVisits	0.8018	0.219	3.667	0.000	0.373	1.230
Total Time Spent on Website	4.4755	0.158	28.389	0.000	4.166	4.784
Page Views Per Visit	-0.5940	0.286	-2.080	0.037	-1.154	-0.034
Lead Add Form	4.0385	0.195	20.717	0.000	3.656	4.421
Olark Chat	0.8414	0.118	7.107	0.000	0.609	1.073
Converted to Lead	-1.4320	0.213	-6.723	0.000	-1.849	-1.014
Email Bounced	-2.3764	0.308	-7.708	0.000	-2.981	-1.772
Working Professional	3.0413	0.194	15.636	0.000	2.660	3.422