SUMMARY REPORT

1. Steps Followed in the process:

1.1. Importing data and Conducting EDA

1.1.1. In this process we performed some primary data cleaning which involved steps such as identifying and removing rows/columns containing Null values. Along with removing null, the rows of categorical variable columns containing 'select' as values were also removed as they were equivalent to null with respect to our model.

1.1.2. They category for removing entire columns was percentage of Null values > 40%.

1.1.3. The next process was conducting univariate and bivariate analysis using data visualisation methods such as employing pair plots, box-plots, scatter plots and heatmap for correlation values between influencing variables.

1.1.4. In order to prepare the data for applying the logistic regression model, binary variables (Yes/No) were converted to (1/0), categorical variables were converted into dummy variables.

1.2. Splitting the data into train and test data Set:

1.2.1. The cleaned data was bifurcated into train and test data set using train_test_split function in sklearn library.

1.2.2. The numerical variables in the data were normalised using MinMaxScaler class from sklearn.preprocessing.

1.3. Building the Model:

1.3.1. Primary Logistic Regression model was built on the train data set using LogisticRegression class in sklearn.linear_model library

1.3.2. Since there were too many influencing variables affecting the target variable, RFE technique was utilised to select high priority variables based on rfe rankings

1.3.3. Then, Generalised Linear Model were built using statsmodels.api module.

1.3.4. The next step was to check the interdependency of the predictor variables using VIF values. Here all the cutoff of eliminating variables based on VIF value was kept at 5.

1.3.5. Similarly, subsequent models were built iteratively until all the variables had vif < 5.

1.3.6. The result of the final model are as follows:

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	3174
Model:	GLM	**Df Residuals:**	3161
Model Family:	Binomial	**Df Model:**	12
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1454.5
Date:	Mon, 27 May 2024	**Deviance:**	2908.9
Time:	19:40:11	**Pearson chi2:**	3.87e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

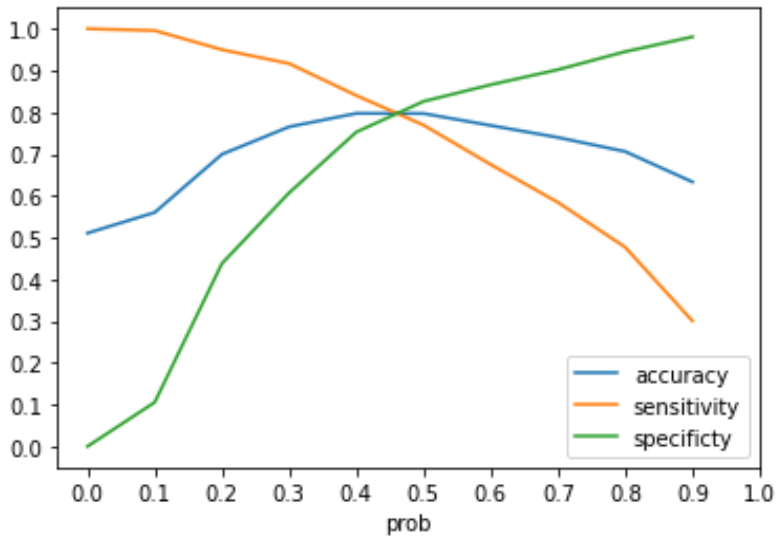
	coef	std err	z	P>|z|	[0.025	0.975]
const	0.2375	0.270	0.881	0.379	-0.291	0.766
TotalVisits	11.4496	3.523	3.250	0.001	4.545	18.354
Total Time Spent on Website	4.0298	0.194	20.807	0.000	3.650	4.409
Page Views Per Visit	-5.0842	1.574	-3.231	0.001	-8.168	-2.000
Lead Origin_Landing Page Submission	-1.3298	0.145	-9.176	0.000	-1.614	-1.046
Lead Source_Reference	2.6492	0.360	7.354	0.000	1.943	3.355
Last Activity_Converted to Lead	-1.3862	0.287	-4.824	0.000	-1.949	-0.823
Last Activity_Email Bounced	-3.5878	1.034	-3.470	0.001	-5.614	-1.562
Last Activity_SMS Sent	0.9708	0.099	9.810	0.000	0.777	1.165
What is your current occupation_Unemployed	-0.8612	0.241	-3.572	0.000	-1.334	-0.389
What is your current occupation_Working Professional	1.8827	0.314	5.992	0.000	1.267	2.499
Last Notable Activity_Email Bounced	2.6196	1.203	2.177	0.029	0.262	4.978
Last Notable Activity_Unreachable	2.9147	1.112	2.621	0.009	0.735	5.094

1.4. Model Evaluation:

1.4.1. In this step, the model outcomes generated in the form of a probability value. It means that with a certain probability the outcome could be positive or negative (In this case converted or not).

1.4.2. In order to find the optimal cutoff value based on which the predicted outcome will be produced, different probability values were tried to calculate the optimal cutoff point.

1.4.3. From our analysis the optimal cutoff point was calculated as 0.47, as given in the following figure:



1.4.4. The final step in the model evaluation is to find the accuracy of the model. It was done by forming confusion matrix. The accuracy value of the model was calculated to be around 78%.

2. Learnings from the Model

2.1. Most important variables

2.1.1. From the model, some of the most important variables affecting the probability of a lead turning into potential customers are **'Total Visits'**, **'Total time spent on websites'**, **'Lead Source (reference)', 'Current Occupation (Working Professional)'**

2.1.1.1. It means that X education Pvt. Limited should focus on the leads who have a high number of total visits and time spent on the website. Also, the leads who are working professional and/or are generated through reference have higher probability of converting into customers.

3. Model Outcomes

3.1. In the end, the model presents a 'Lead Score', which can help the company identify which particular lead they need to follow with a potential to turn into a customer.

3.2. The model accuracy was around 78% which is a standard acceptable value as per the industry norms.