Summary Report

This analysis is done for X Education and to find ways to get more industry professionals to join their courses. The basic data provided gave us a lot of information about how the potential customers visit the site, the time they spend there, how they reached the site and the conversion rate.

The Motive behind this Analysis is to increase the number of learners practically taking admission rather than Just visiting the portal or calling help desk.
Hence To overcome the issue and increase the leads, we prepared a model to enhance the business and increase the sales.

The following are the steps followed in the Logistic Regression Model:

1. Importing & Cleaning data:
The data was partially clean except for a few null values and the option select had to be replaced with a null value since it did not give us much information. Few of the null values were changed to 'not provided' so as to not lose much data. Although they were later removed while making dummies.
Since there were many from India and few from outside, the elements were changed to 'India', 'Outside India' and 'not provided' and later Country was dropped.

2. EDA:
A quick EDA was done to check the condition of our data. It was found that a lot of elements in the categorical variables were irrelevant. The numeric values seem good and no outliers were found.
And Univariate Analysis and Bivariate Analysis was performed to plot graphs and show visualizations of the Variables.

3. Dummy Variables:
The dummy variables were created and later on the dummies with 'not provided' elements were removed. Most of the categorical variables were domified.

4. Train-Test split:
The split was done at 70% and 30% for train and test data respectively.

5. Feature Scaling:
To scale and standardize the variables we used Min-Max-Scaler.

6. Looking at Correlations
We used Heatmap to visualize the correlation between variables. And dropped highly corelated variable.

7. Model Building:

Firstly, RFE was done to attain the top 20 relevant variables. Later the rest of the variables were removed manually depending on the VIF values and p-value (The variables with VIF < 5 and p-value < 0.05 were kept).

So Finally, by dropping 5 variables final model was left with 15 Variables.

8. Model Evaluation:
A confusion matrix was made. Later on the optimum cut off value (using ROC curve) was used to find the accuracy, sensitivity and specificity which came to be around 80% each. In fact sensitivity responded well with 81.4%

9. Precision – Recall:
This method was also used to recheck and a cut off of 0.41 was found with Precision around 74% and recall around 75% on the test data frame.

10. Prediction:
Prediction was done on the test data frame and with an optimum cut off as 0.35 with accuracy, sensitivity and specificity of 80%.

After Building the model we conclude that :

It was found that the variables that mattered the most in the potential buyers are (In descending order) :

- 1. Total number of visits.
- 2. The total time spend on the Website.
- 3. Lead Origin through Add Form
- 4. When the lead source was:
    a. Welingak Website
    b. Olark Chat
- 5. When their current occupation is as a working professional.
- 6. Last Activity of the user.

Recommendations

Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses by following some different methods and approaches and increase the sales by :

- Automated emails and SMS
- Campaigns on Social Networking sites
- Some Additional add-on prep courses

Summary of the Final Logistic Regression Model Built :

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6468
Model:	GLM	**Df Residuals:**	6452
Model Family:	Binomial	**Df Model:**	15
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2598.2
Date:	Mon, 07 Sep 2020	**Deviance:**	5196.4
Time:	17:09:02	**Pearson chi2:**	7.11e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.5154	0.128	-19.640	0.000	-2.766	-2.264
TotalVisits	6.1617	1.951	3.158	0.002	2.338	9.985
Total Time Spent on Website	4.5357	0.168	27.046	0.000	4.207	4.864
Lead Origin_Lead Add Form	3.5943	0.193	18.595	0.000	3.215	3.973
Lead Source_Olark Chat	1.4507	0.119	12.221	0.000	1.218	1.683
Lead Source_Welingak Website	2.0432	0.746	2.738	0.006	0.580	3.506
Do Not Email_Yes	-1.2283	0.179	-6.866	0.000	-1.579	-0.878
Last Activity_Email Opened	0.5923	0.106	5.579	0.000	0.384	0.800
Last Activity_Other	1.7349	0.516	3.365	0.001	0.724	2.745
Last Activity_SMS Sent	1.7086	0.108	15.874	0.000	1.498	1.920
Specialization_Hospitality Management	-0.9460	0.328	-2.885	0.004	-1.589	-0.303
Specialization_not provided	-0.2544	0.088	-2.877	0.004	-0.428	-0.081
What is your current occupation_Working Professional	2.5095	0.191	13.140	0.000	2.135	2.884
What matters most to you in choosing a course_not provided	-1.0823	0.088	-12.340	0.000	-1.254	-0.910
Last Notable Activity_Modified	-0.6385	0.088	-7.250	0.000	-0.811	-0.466
Last Notable Activity_Other	1.7829	0.352	5.062	0.000	1.093	2.473