1. Problem Statement and the brief on analysis approach

>X Education company has poor lead conversion rate of say 30%. In order to get higher lead conversion rate, company wants us to build a Machine Learning model which will satisfy this requirement of higher lead conversion rate (80% target from CEO)

>Here the dataset consists of 9000 data points, with attributes like Total time spent on website, Total Visits, etc. The Target variable is ‘Converted’

>Goals: 1. Build a logistic regression model which assigns the lead scores between 0 to 100. Higher number showing better possibility of conversion.

1. Contd…

>EDA:

Used sweetviz an automatics EDA technique for Data Visualization.

Data Cleaning : Handled the missing values in the DataSet. Dropped Null values columns. Removed the columns ‘City’, ‘Country’ having lesser impact on the ML model based on our Business case understanding. Dropped the columns having major missing values by identifying the value ‘Select’. Dropped other columns having only one value majorly as a data point (showing lesser variance in EDA plot) eg. I agree to pay the amount by cheque. Dropped null value rows.

Dummy Variables: Employed dummy variables to convert categorical variables into a format suitable for modeling.

>Train-Test Splitting & Scaling:

Splitted the DataSet into 70% train and 30% test. Scaled few numeric variables having different scales.

1. Contd(2)…

>Model Building:

Used RFE to select top 15 relevant variables from the DataSet. Manually removed the variables based on VIF and p-value. Evaluation of model using confusion matrix

>Model Evaluation:

Used ROC curve for finding the optimal cut-off value. Used sensitivity and specificity tradeoff for optimal cutoff point.

> Making Predictions on the Test Data:

Checked for the predictions on the test data. It shows ~79% accuracy and is acceptable.

2. Explain the results in business terms

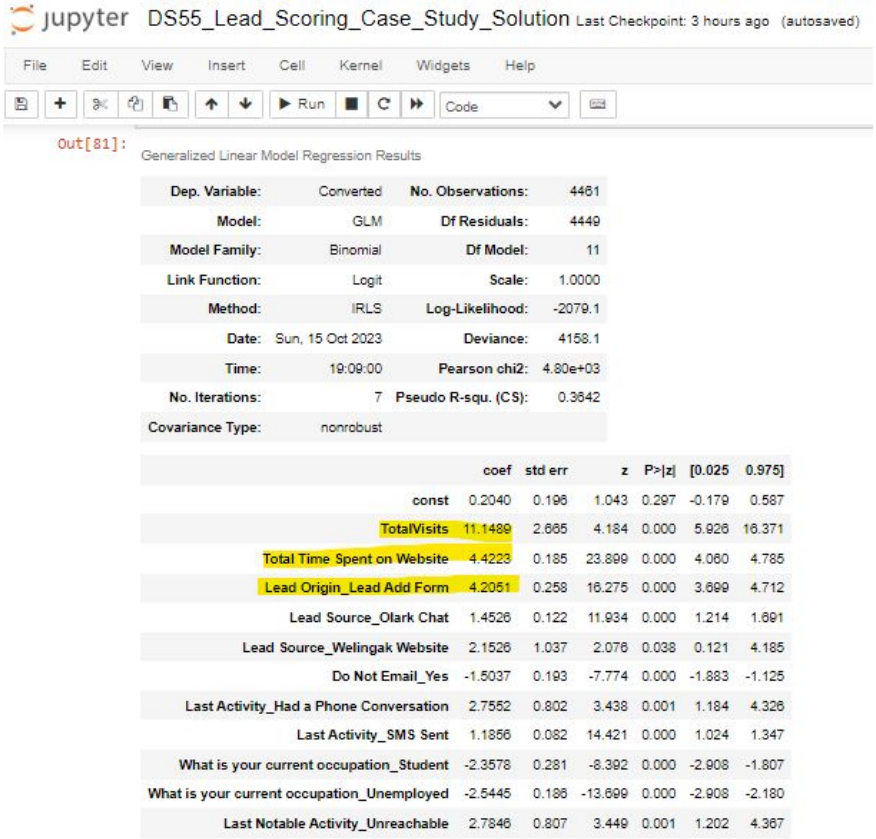
>The Machine Learning Model built gives predictions which are ~80% correct. It rightly helps the X Education company to funnel the potential leads. If we give 100 Data points the this ML model it will give the output of say 30 leads then the probability of these leads getting converted is 80% i.e 24 leads.

>The model thus reduce the manual efforts. Reduces company's efforts, resources, revenue required for cold calling thus saving the lot of revenue.

>This ML model also gives suggestions to the CEO regarding the top variables contributing to the conversion rate. This will also help the CEO to understand the marketing possibilities and direct the company's revenue & resources to the right marketing source.

>This model thus helps a lot to the company and the CEO in making right decisions.

3. Summarize the most important results



jupyter DS55_Lead_Scoring_Case_Study_Solution Last Checkpoint: 3 hours ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help

Run Code

Out[81]:

Generalized Linear Model Regression Results

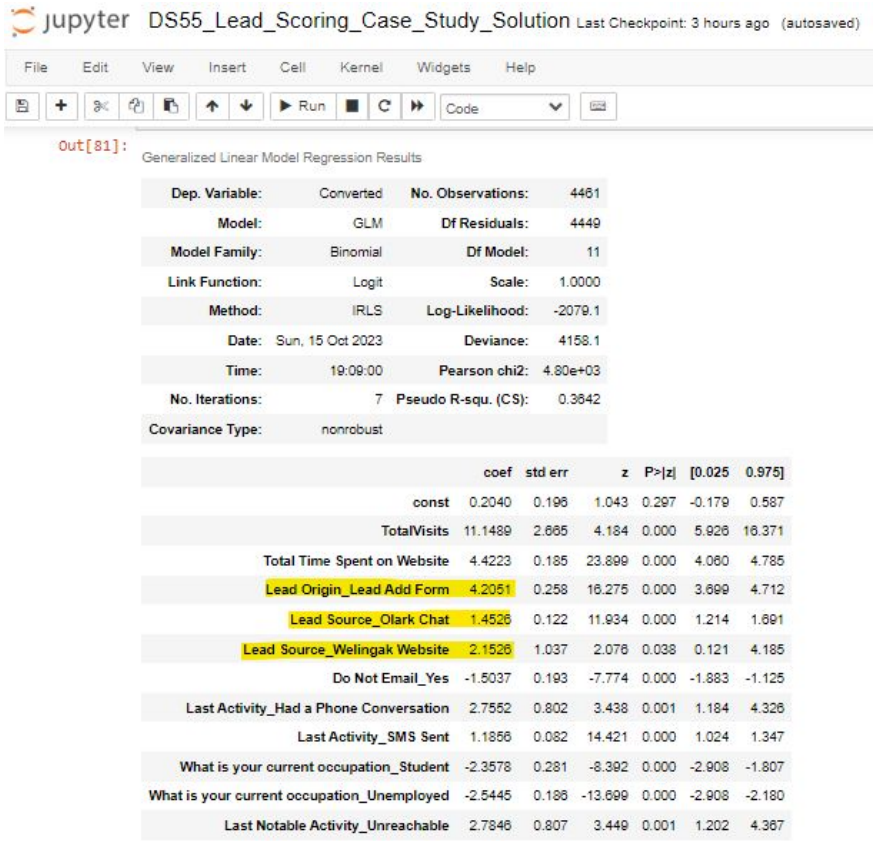
Dep. Variable:	Converted	No. Observations:	4461
Model:	GLM	Df Residuals:	4449
Model Family:	Binomial	Df Model:	11
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2079.1
Date:	Sun, 15 Oct 2023	Deviance:	4158.1
Time:	19:09:00	Pearson chi2:	4.80e+03
No. Iterations:	7	Pseudo R-squ. (CS):	0.3642
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.2040	0.196	1.043	0.297	-0.179	0.587
TotalVisits	11.1489	2.665	4.184	0.000	5.926	16.371
Total Time Spent on Website	4.4223	0.185	23.899	0.000	4.060	4.785
Lead Origin_Lead Add Form	4.2051	0.258	16.275	0.000	3.699	4.712
Lead Source_Olark Chat	1.4526	0.122	11.934	0.000	1.214	1.691
Lead Source_Welingak Website	2.1526	1.037	2.076	0.038	0.121	4.185
Do Not Email_Yes	-1.5037	0.193	-7.774	0.000	-1.883	-1.125
Last Activity_Had a Phone Conversation	2.7552	0.802	3.438	0.001	1.184	4.326
Last Activity_SMS Sent	1.1856	0.082	14.421	0.000	1.024	1.347
What is your current occupation_Student	-2.3578	0.281	-8.392	0.000	-2.908	-1.807
What is your current occupation_Unemployed	-2.5445	0.186	-13.699	0.000	-2.908	-2.180
Last Notable Activity_Unreachable	2.7846	0.807	3.449	0.001	1.202	4.367

Following three variables contribute most towards the probability of getting a lead to be converted and have maximum co-relation coefficient

1.) **TotalVisits = 11.1489**
2.) **Total Time Spent on Website = 4.4223**
3.) **Lead Origin_Lead Add Form = 4.2051**

3. Contd..


jupyter DS55_Lead_Scoring_Case_Study_Solution Last Checkpoint: 3 hours ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help

Run Code

Out[81]:

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	4461
Model:	GLM	Df Residuals:	4449
Model Family:	Binomial	Df Model:	11
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2079.1
Date:	Sun, 15 Oct 2023	Deviance:	4158.1
Time:	19:09:00	Pearson chi2:	4.80e+03
No. Iterations:	7	Pseudo R-squ. (CS):	0.3642
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.2040	0.196	1.043	0.297	-0.179	0.587
TotalVisits	11.1489	2.665	4.184	0.000	5.926	16.371
Total Time Spent on Website	4.4223	0.185	23.899	0.000	4.060	4.785
Lead Origin_Lead Add Form	4.2051	0.258	16.275	0.000	3.699	4.712
Lead Source_Olark Chat	1.4526	0.122	11.934	0.000	1.214	1.691
Lead Source_Welingak Website	2.1526	1.037	2.076	0.038	0.121	4.185
Do Not Email_Yes	-1.5037	0.193	-7.774	0.000	-1.883	-1.125
Last Activity_Had a Phone Conversation	2.7552	0.802	3.438	0.001	1.184	4.326
Last Activity_SMS Sent	1.1856	0.082	14.421	0.000	1.024	1.347
What is your current occupation_Student	-2.3578	0.281	-8.392	0.000	-2.908	-1.807
What is your current occupation_Unemployed	-2.5445	0.186	-13.699	0.000	-2.908	-2.180
Last Notable Activity_Unreachable	2.7846	0.807	3.449	0.001	1.202	4.367

Following three categorical/dummy variables contribute most towards the probability of getting a lead to be converted and have maximum co-relation coefficient

1.) **Lead Origin_Lead Add Form = 4.2051**
2.) **Lead Source_Olark Chat= 1.4526**
3.) **Lead Source_Welingak Website = 2.1526**

3. Contd(2)

- **The model demonstrates effective prediction of the conversion rate, with key variables identified as significant contributors. These include metrics related to lead sources, lead origin, and last activity. Notable sources of influence on conversion include total visits, time spent on the website, and engagement through the lead add form.**
- **This analysis provides actionable insights for decision-making, allowing the company to make informed calls based on the identified influential factors. The model's robust performance instills confidence in its predictive capabilities, enhancing the company's ability to strategize effectively for conversion optimization.**