Logistical Regression Case Study

Priyadharshni S M

Taher Chitalwala

PROBLEM STATEMENT AND APPROACH

Objective:

Evaluate the entire data sets of X Education to identify factors that lead to a higher lead conversion. Multiple data sources are available with the Company, however the lead conversion is very poor at 30%. The objective is to improve the lead conversion rate to ~80%

Overall Approach:

1. Exploratory data analysis was performed to understand the data
2. The data was reviewed for errors and inconsistencies. Missing values were treated in line with the methodology provided below
3. Three versions of the model were prepared.
4. The model was reviewed on the training and test set for accuracy

STEP 1: EXPLORATORY DATA ANALYSIS (1/2)

Activity Performed:

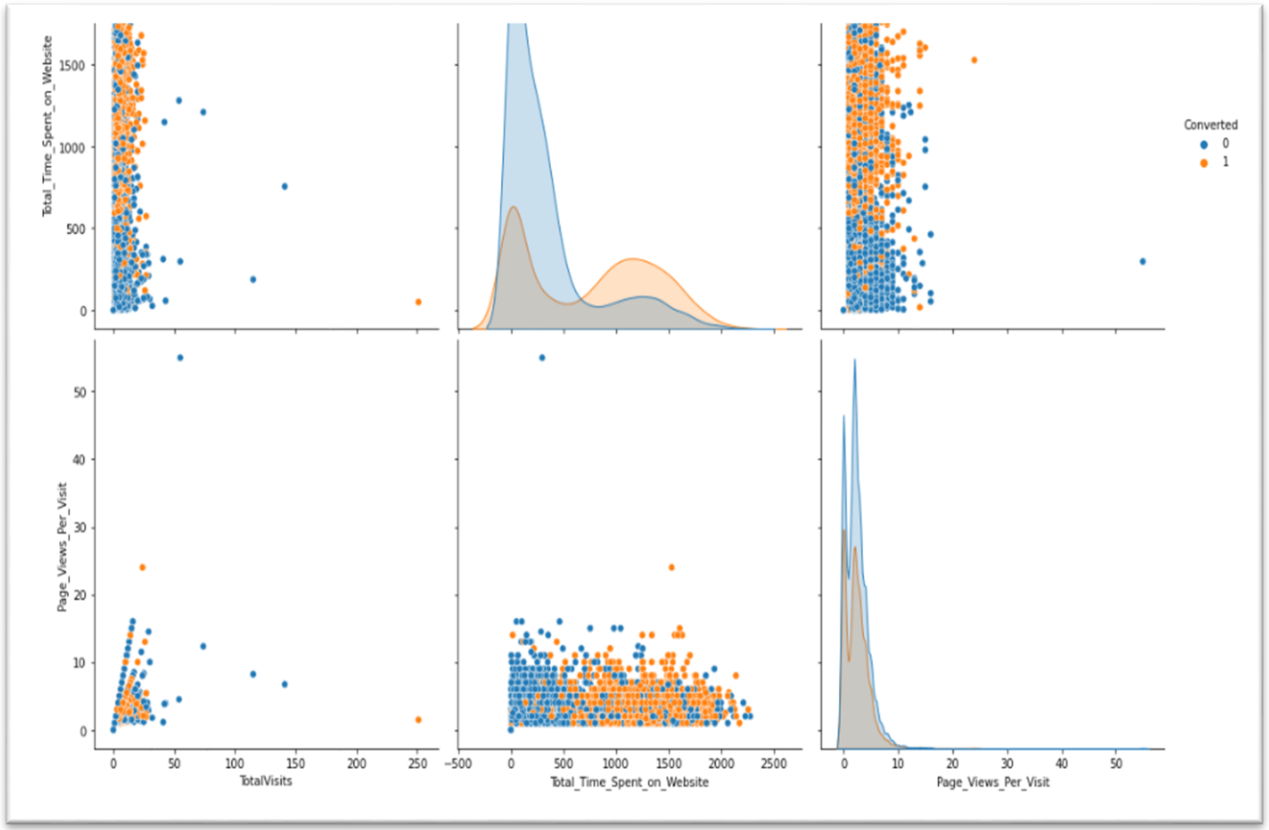
1. Data was reviewed for basic hygiene information *(nulls, outliers, shape etc.).*
2. Data was split into 3 categories for review through charts and tables *(Binary, Categorical, and Numerical Values)*

Key Findings:

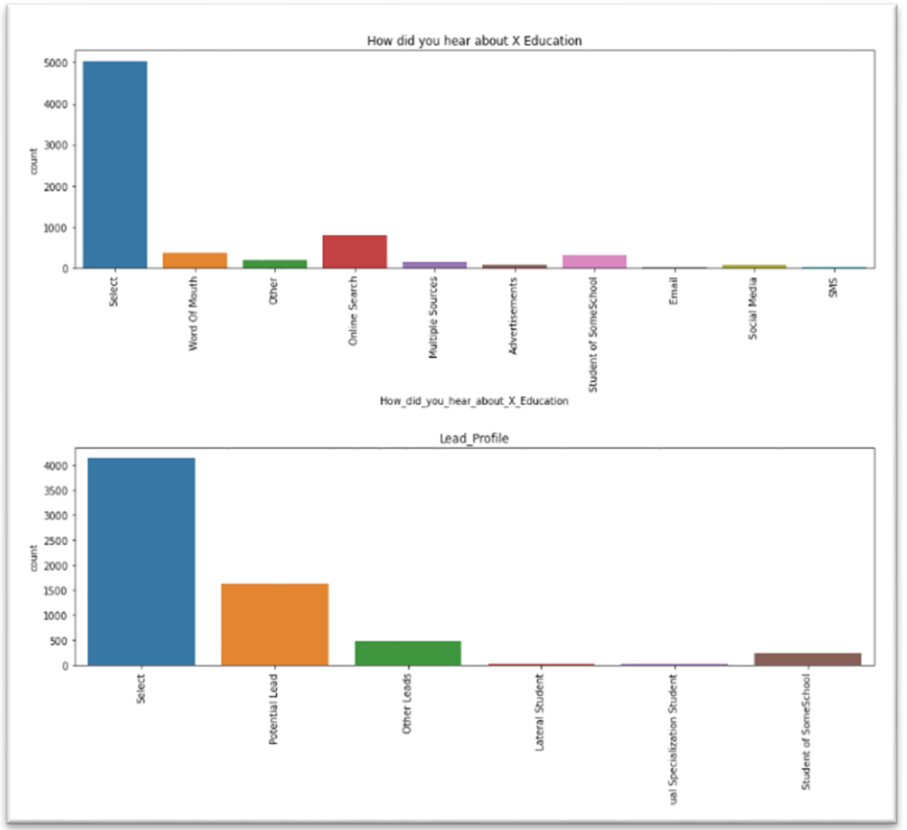
1. Data was concentrated in sections which would prove inefficient with regression.
   Country: 70% was India, 26% was blanks, and 4% was Other counties)
   What matters while choosing a course: 96% Better career prospects
   Employment Status: 85% Unemployed
2. Binary values were a complete. Fields were identified as a 99-100% No.
   *(**100% Values** – Magazine, Additional Updates, Supply Content Updates, Payment through Cheque, Get DM Content)*
   *(**99% Values** – Newspaper and Newspaper Article, Forums, Do Not call,, Digital Advertisement, Through Recommendations)*
3. Multiple columns had the single largest value as Null, NA
4. Time spent on the internet had a clear relationship with Leads. *(Refer pair-plots)*

STEP 1: EXPLORATORY DATA ANALYSIS (2/2)

Time Spent on Internet and Pages Per Visit Examples



Null Value Examples – Lead Profile and Others



STEP 2: CLEANING AND DATA HYGIENE

Activity Performed:

1. Select values were replaced with Null
2. Where categories of each feature were less than 2% of the count, values are aggregated into a residual category 'Others'.
3. In case 70% or more of the values were not available for a single row, the row was deleted. (No results deleted in this test)
4. For missing numerical values, Median values were inserted
5. For missing categorical values, Mode were inserted

Note: *(2% Threshold was selected due to large aggregations in case of a larger threshold)*

STEP 3: MODEL TESTING AND EVALUATION: (1/3)

Activity Performed:

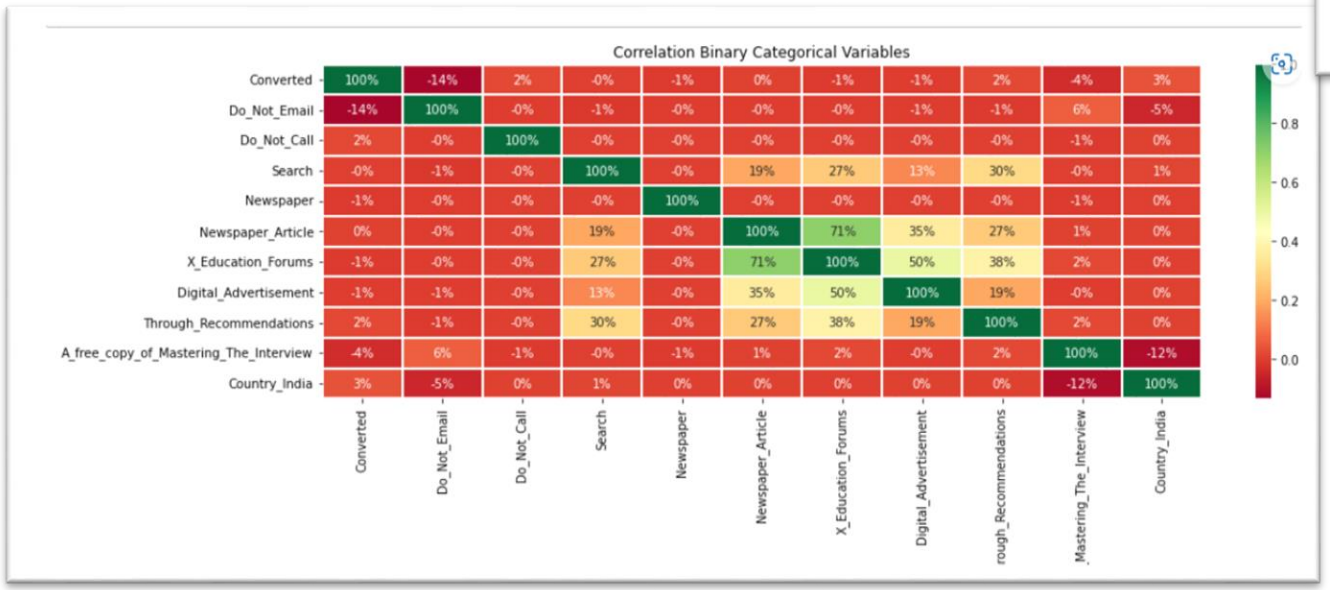
1. The correlation matrix was reviewed prior to modelling
2. RFE and variable resulted in the final 14 variables
3. The variables were reviewed for VIF and statistical significance.

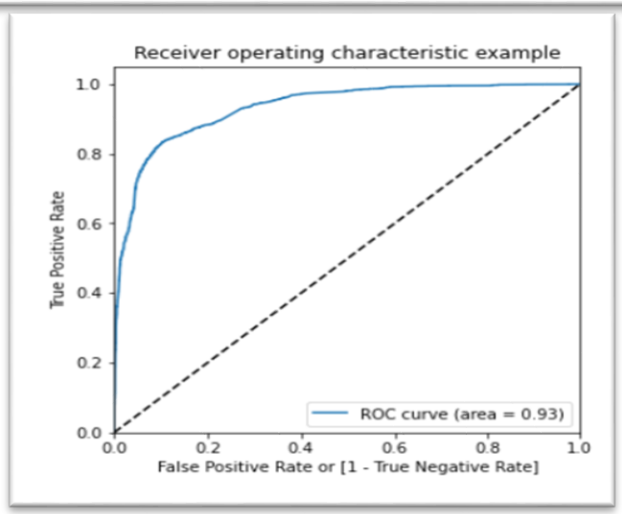
Final Model

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6468
Model:	GLM	**Df Residuals:**	6453
Model Family:	Binomial	**Df Model:**	14
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2035.4
Date:	Tue, 18 Oct 2022	**Deviance:**	4070.7
Time:	02:00:19	**Pearson chi2:**	9.89e+03
No. Iterations:	9		
Covariance Type:	nonrobust		

Correlation Matrix





STEP 3: MODEL TESTING AND EVALUATION: (2/3)

Model Evaluation:

1. The detailed model factors are listed on the right side.
2. The positive and negative model factors have been separately added.
3. All factors listed by RFE have logical concurrence *(Rounded Off to the nearest decimals)*
4. None of the factors listed include variables that hold significant concentration

Variable	Coeff	P Value	VIF
const	-5.4	0	9.7
Do_Not_Email	-1.3	0	1.1
Lead_Origin_Lead Add Form	3.37	0	1.3
Last_Activity_Email Opened	1.1	0	1.6
Last_Activity_Others	1.17	0	1.2
Last_Activity_SMS Sent	0.89	0	4.2
Tags_Busy	3.33	0	1.2
Tags_Closed by Horizzon	9.2	0	1.5
Tags_Others	3.22	0	1.6
Tags_Ringing	-0.8	0.02	1.8
Tags_Will revert after reading the email	4.1	0	2.3
Lead_Profile_Others	3.92	0.004	1
Asymmetrique_Activity_Index_Low	-1.6	0	1.1
Last_Notable_Activity_SMS Sent	2.24	0	3.7
Total_Time_Spent_on_Website	1.04	0	1.1

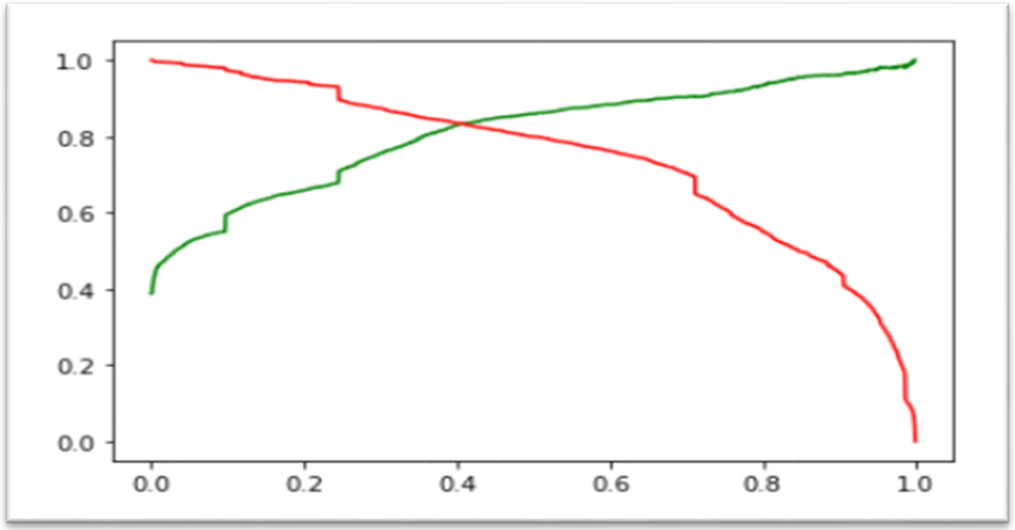
STEP 3: MODEL TESTING AND EVALUATION: (3/3)

Model Evaluation:

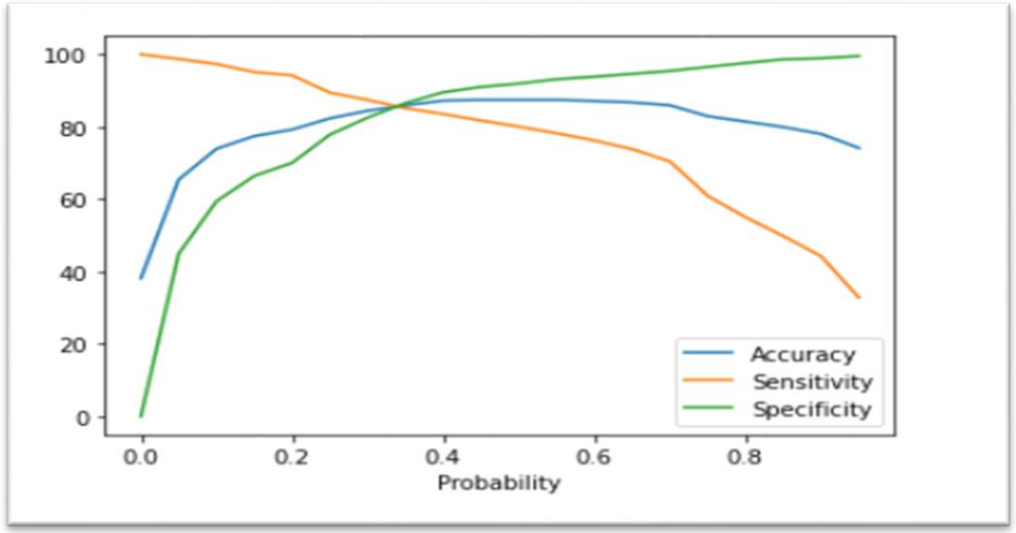
1. The correlation matrix was reviewed prior to modelling
2. RFE and variable resulted in the final 14 variables
3. The variables were reviewed for VIF and statistical significance.
4. The optimum cut off based on training and test sets lie between 0.35 and 0.4. Hence 0.37 was selected.

Matrix for Performance	Training	Test Set
Overall Accuracy	86.5	85.8
Sensitivity/ Recall	84.47	85.11
Specificity	87.76	86.29
False positive rate	12.24	13.71
Positive Predictive rate/ Precision	80.96	80.21
Negative Predictive rate	90.17	89.88

Precision Vs Recall of the Model



Accuracy, Sensitivity and Specificity



CONCLUSION

- Data cleaning has been performed using fairly moderate assumptions.
- Not all variables for dropped due to redundancy.
- RFE has factored to exclude redundant variables observed within the EDA section
- The ROC curve and model parameters appear sound and logical
- Accuracy, sensitivity, specificity, & positive predictive value are within acceptable levels.

THANK YOU